
LLVM Documentation

Release 5

LLVM project

2018-03-16

Contents

1	LLVM Design & Overview	3
1.1	LLVM Language Reference Manual	3
2	User Guides	241
2.1	Building LLVM with CMake	241
2.2	CMake Primer	251
2.3	Advanced Build Configurations	258
2.4	How To Build On ARM	260
2.5	How To Cross-Compile Clang/LLVM using Clang/LLVM	262
2.6	LLVM Command Guide	264
2.7	Getting Started with the LLVM System	318
2.8	Getting Started with the LLVM System using Microsoft Visual Studio	338
2.9	Frequently Asked Questions (FAQ)	341
2.10	The LLVM Lexicon	347
2.11	How To Add Your Build Configuration To LLVM Buildbot Infrastructure	350
2.12	yaml2obj	352
2.13	How to submit an LLVM bug report	355
2.14	Sphinx Quickstart Template	358
2.15	Code Reviews with Phabricator	360
2.16	LLVM Testing Infrastructure Guide	364
2.17	LLVM Tutorial: Table of Contents	378
2.18	LLVM 5.0.0 Release Notes	747
2.19	LLVM's Analysis and Transform Passes	751
2.20	YAML I/O	771
2.21	The Often Misunderstood GEP Instruction	787
2.22	Performance Tips for Frontend Authors	794
2.23	MCJIT Design and Implementation	799
2.24	LLVM Community Code of Conduct	805
2.25	Compiling CUDA with clang	807
2.26	Reporting Guide	815
2.27	Benchmarking tips	817
2.28	A guide to Dockerfiles for building LLVM	818
3	Programming Documentation	823
3.1	LLVM Atomic Instructions and Concurrency Guide	823
3.2	LLVM Coding Standards	831
3.3	CommandLine 2.0 Library Manual	855

3.4	Architecture & Platform Information for Compiler Writers	878
3.5	Extending LLVM: Adding instructions, intrinsics, types, etc.	882
3.6	How to set up LLVM-style RTTI for your class hierarchy	886
3.7	LLVM Programmer's Manual	892
3.8	LLVM Extensions	949
3.9	libFuzzer – a library for coverage-guided fuzz testing.	953
3.10	Scudo Hardened Allocator	966
3.11	Using -opt-bisect-limit to debug optimization errors	969
4	Subsystem Documentation	973
4.1	LLVM Alias Analysis Infrastructure	973
4.2	MemorySSA	984
4.3	LLVM Bitcode File Format	990
4.4	LLVM Block Frequency Terminology	1011
4.5	LLVM Branch Weight Metadata	1013
4.6	LLVM bugpoint tool: design and usage	1016
4.7	The LLVM Target-Independent Code Generator	1019
4.8	Exception Handling in LLVM	1053
4.9	LLVM Link Time Optimization: Design and Implementation	1065
4.10	Segmented Stacks in LLVM	1069
4.11	TableGen Fundamentals	1070
4.12	TableGen	1070
4.13	Debugging JIT-ed Code With GDB	1099
4.14	The LLVM gold plugin	1101
4.15	LLVM's Optional Rich Disassembly Output	1103
4.16	System Library	1105
4.17	Source Level Debugging with LLVM	1108
4.18	Auto-Vectorization in LLVM	1130
4.19	Writing an LLVM Backend	1140
4.20	Garbage Collection with LLVM	1171
4.21	Writing an LLVM Pass	1186
4.22	How To Use Attributes	1206
4.23	User Guide for NVPTX Back-end	1207
4.24	User Guide for AMDGPU Backend	1222
4.25	Stack maps and patch points in LLVM	1290
4.26	Design and Usage of the InAlloca Attribute	1297
4.27	Using ARM NEON instructions in big endian mode	1299
4.28	LLVM Code Coverage Mapping Format	1305
4.29	Garbage Collection Safepoints in LLVM	1314
4.30	MergeFunctions pass, how it works	1326
4.31	Type Metadata	1338
4.32	FaultMaps and implicit checks	1341
4.33	Machine IR (MIR) Format Reference Manual	1343
4.34	Coroutines in LLVM	1351
4.35	Global Instruction Selection	1371
4.36	XRay Instrumentation	1379
4.37	Debugging with XRay	1383
4.38	The PDB File Format	1389
5	Development Process Documentation	1405
5.1	LLVM Developer Policy	1405
5.2	Creating an LLVM Project	1415
5.3	LLVMBuild Guide	1419
5.4	How To Release LLVM To The Public	1423

5.5	Advice on Packaging LLVM	1429
5.6	How To Validate a New Release	1430
6	Community	1435
6.1	Mailing Lists	1435
6.2	IRC	1435
6.3	Community wide proposals	1436
7	Indices and tables	1451
	Bibliography	1453

The LLVM compiler infrastructure supports a wide range of projects, from industrial strength compilers to specialized JIT applications to small research projects.

Similarly, documentation is broken down into several high-level groupings targeted at different audiences:

Several introductory papers and presentations.

1.1 LLVM Language Reference Manual

- *Abstract*
- *Introduction*
 - *Well-Formedness*
- *Identifiers*
- *High Level Structure*
 - *Module Structure*
 - *Linkage Types*
 - *Calling Conventions*
 - *Visibility Styles*
 - *DLL Storage Classes*
 - *Thread Local Storage Models*
 - *Structure Types*
 - *Non-Integral Pointer Type*
 - *Global Variables*
 - *Functions*
 - *Aliases*

- *IFuncs*
- *Comdats*
- *Named Metadata*
- *Parameter Attributes*
- *Garbage Collector Strategy Names*
- *Prefix Data*
- *Prologue Data*
- *Personality Function*
- *Attribute Groups*
- *Function Attributes*
- *Global Attributes*
- *Operand Bundles*
 - * *Deoptimization Operand Bundles*
 - * *Funclet Operand Bundles*
 - * *GC Transition Operand Bundles*
- *Module-Level Inline Assembly*
- *Data Layout*
- *Target Triple*
- *Pointer Aliasing Rules*
- *Volatile Memory Accesses*
- *Memory Model for Concurrent Operations*
- *Atomic Memory Ordering Constraints*
- *Fast-Math Flags*
- *Use-list Order Directives*
- *Source Filename*
- *Type System*
 - *Void Type*
 - *Function Type*
 - *First Class Types*
 - * *Single Value Types*
 - *Integer Type*
 - *Floating Point Types*
 - *X86_mmx Type*
 - *Pointer Type*
 - *Vector Type*

- * *Label Type*
- * *Token Type*
- * *Metadata Type*
- * *Aggregate Types*
 - *Array Type*
 - *Structure Type*
 - *Opaque Structure Types*
- *Constants*
 - *Simple Constants*
 - *Complex Constants*
 - *Global Variable and Function Addresses*
 - *Undefined Values*
 - *Poison Values*
 - *Addresses of Basic Blocks*
 - *Constant Expressions*
- *Other Values*
 - *Inline Assembler Expressions*
 - * *Inline Asm Constraint String*
 - *Output constraints*
 - *Input constraints*
 - *Indirect inputs and outputs*
 - *Clobber constraints*
 - *Constraint Codes*
 - *Supported Constraint Code List*
 - * *Asm template argument modifiers*
 - * *Inline Asm Metadata*
- *Metadata*
 - *Metadata Nodes and Metadata Strings*
 - * *Specialized Metadata Nodes*
 - *DICompileUnit*
 - *DIFile*
 - *DIBasicType*
 - *DISubroutineType*
 - *IDerivedType*
 - *DICompositeType*

- *DISubrange*
- *DIEnumerator*
- *DITemplateTypeParameter*
- *DITemplateValueParameter*
- *DINamespace*
- *DIGlobalVariable*
- *DISubprogram*
- *DILexicalBlock*
- *DILexicalBlockFile*
- *DILocation*
- *DILocalVariable*
- *DIExpression*
- *DIObjCProperty*
- *DIImportedEntity*
- *DIMacro*
- *DIMacroFile*
- * *'tbaa' Metadata*
 - *Semantics*
 - *Representation*
- * *'tbaa.struct' Metadata*
- * *'noalias' and 'alias.scope' Metadata*
- * *'fpmath' Metadata*
- * *'range' Metadata*
- * *'absolute_symbol' Metadata*
- * *'unpredictable' Metadata*
- * *'llvm.loop'*
- * *'llvm.loop.vectorize' and 'llvm.loop.interleave'*
- * *'llvm.loop.interleave.count' Metadata*
- * *'llvm.loop.vectorize.enable' Metadata*
- * *'llvm.loop.vectorize.width' Metadata*
- * *'llvm.loop.unroll'*
- * *'llvm.loop.unroll.count' Metadata*
- * *'llvm.loop.unroll.disable' Metadata*
- * *'llvm.loop.unroll.runtime.disable' Metadata*
- * *'llvm.loop.unroll.enable' Metadata*

- * *'llvm.loop.unroll.full' Metadata*
- * *'llvm.loop.licm_versioning.disable' Metadata*
- * *'llvm.loop.distribute.enable' Metadata*
- * *'llvm.mem'*
- * *'llvm.mem.parallel_loop_access' Metadata*
- * *'invariant.group' Metadata*
- * *'type' Metadata*
- * *'associated' Metadata*
- * *'prof' Metadata*
 - *branch_weights*
 - *function_entry_count*
 - *VP*
- *Module Flags Metadata*
 - *Objective-C Garbage Collection Module Flags Metadata*
 - *C type width Module Flags Metadata*
- *Automatic Linker Flags Named Metadata*
- *Intrinsic Global Variables*
 - *The 'llvm.used' Global Variable*
 - *The 'llvm.compiler.used' Global Variable*
 - *The 'llvm.global_ctors' Global Variable*
 - *The 'llvm.global_dtors' Global Variable*
- *Instruction Reference*
 - *Terminator Instructions*
 - * *'ret' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
 - * *'br' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*

- * *'switch' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Implementation:*
 - *Example:*
- * *'indirectbr' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Implementation:*
 - *Example:*
- * *'invoke' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'resume' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'catchswitch' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'catchret' Instruction*
 - *Syntax:*
 - *Overview:*

- *Arguments:*
- *Semantics:*
- *Example:*
- * *'cleanupret' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'unreachable' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Semantics:*
- *Binary Operations*
 - * *'add' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
 - * *'fadd' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
 - * *'sub' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
 - * *'fsub' Instruction*
 - *Syntax:*
 - *Overview:*

- *Arguments:*
- *Semantics:*
- *Example:*
- * *'mul' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'fmul' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'udiv' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'sdiv' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'fdiv' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'urem' Instruction*
 - *Syntax:*

- *Overview:*
- *Arguments:*
- *Semantics:*
- *Example:*
- * *'srem' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'frem' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- *Bitwise Binary Operations*
 - * *'shl' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
 - * *'lshr' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
 - * *'ashr' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*

- * *'and' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'or' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'xor' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- *Vector Operations*
 - * *'extractelement' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
 - * *'insertelement' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
 - * *'shufflevector' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*

- *Semantics:*
 - *Example:*
- *Aggregate Operations*
 - * *‘extractvalue’ Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
 - * *‘insertvalue’ Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- *Memory Access and Addressing Operations*
 - * *‘alloca’ Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
 - * *‘load’ Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Examples:*
 - * *‘store’ Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
 - * *‘fence’ Instruction*

- *Syntax:*
- *Overview:*
- *Arguments:*
- *Semantics:*
- *Example:*
- * *'cmpxchg' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'atomicrmw' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'getelementptr' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
 - *Vector of pointers:*
- *Conversion Operations*
 - * *'trunc .. to' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
 - * *'zext .. to' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*

- *Semantics:*
- *Example:*
- * *'sext .. to' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'fptrunc .. to' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'fpext .. to' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'fptoui .. to' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'fptosi .. to' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'uitofp .. to' Instruction*
 - *Syntax:*
 - *Overview:*

- *Arguments:*
- *Semantics:*
- *Example:*
- * *'sitofp .. to' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'ptrtoint .. to' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'inttoptr .. to' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'bitcast .. to' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'addrspacecast .. to' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- *Other Operations*
 - * *'icmp' Instruction*

- *Syntax:*
- *Overview:*
- *Arguments:*
- *Semantics:*
- *Example:*
- * *'fcmp' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'phi' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'select' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'call' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'va_arg' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*

- * *'landingpad' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'catchpad' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- * *'cleanuppad' Instruction*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Example:*
- *Intrinsic Functions*
 - *Variable Argument Handling Intrinsics*
 - * *'llvm.va_start' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - * *'llvm.va_end' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - * *'llvm.va_copy' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*

- *Accurate Garbage Collection Intrinsics*
 - * *Experimental Statepoint Intrinsics*
 - * *'llvm.gcroot' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - * *'llvm.gcread' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - * *'llvm.gcwrite' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- *Code Generator Intrinsics*
 - * *'llvm.returnaddress' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - * *'llvm.addressofreturnaddress' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Semantics:*
 - * *'llvm.frameaddress' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - * *'llvm.localescape' and 'llvm.localrecover' Intrinsics*
 - *Syntax:*
 - *Overview:*

- *Arguments:*
- *Semantics:*
- * *'llvm.read_register' and 'llvm.write_register' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Semantics:*
- * *'llvm.stacksave' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Semantics:*
- * *'llvm.stackrestore' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Semantics:*
- * *'llvm.get.dynamic.area.offset' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Semantics:*
- * *'llvm.prefetch' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.pcmarker' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.readcyclecounter' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Semantics:*
- * *'llvm.clear_cache' Intrinsic*
 - *Syntax:*
 - *Overview:*

- *Semantics:*
- * *'llvm.instrprof_increment' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.instrprof_increment_step' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.instrprof_value_profile' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.thread.pointer' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Semantics:*
- *Standard C Library Intrinsics*
 - * *'llvm.memcpy' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - * *'llvm.memmove' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - * *'llvm.memset.*' Intrinsics*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*

- *Semantics:*
- * *'llvm.sqrt.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.powi.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.sin.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.cos.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.pow.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.exp.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.exp2.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*

- *Semantics:*
- * *'llvm.log.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.log10.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.log2.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.fma.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.fabs.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.minnum.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.maxnum.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*

- *Semantics:*
- * `'llvm.copysign.*'` *Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * `'llvm.floor.*'` *Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * `'llvm.ceil.*'` *Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * `'llvm.trunc.*'` *Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * `'llvm rint.*'` *Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * `'llvm.nearbyint.*'` *Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * `'llvm.round.*'` *Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*

- *Semantics:*
- *Bit Manipulation Intrinsics*
 - * *'llvm.bitreverse.*' Intrinsics*
 - *Syntax:*
 - *Overview:*
 - *Semantics:*
 - * *'llvm.bswap.*' Intrinsics*
 - *Syntax:*
 - *Overview:*
 - *Semantics:*
 - * *'llvm.ctpop.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - * *'llvmctlz.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - * *'llvm.cttz.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- *Arithmetic with Overflow Intrinsics*
 - * *'llvm.sadd.with.overflow.*' Intrinsics*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Examples:*
 - * *'llvm.uadd.with.overflow.*' Intrinsics*
 - *Syntax:*
 - *Overview:*

- *Arguments:*
- *Semantics:*
- *Examples:*
- * *'llvm.ssub.with.overflow.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Examples:*
- * *'llvm.usub.with.overflow.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Examples:*
- * *'llvm.smul.with.overflow.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Examples:*
- * *'llvm.umul.with.overflow.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Examples:*
- *Specialised Arithmetic Intrinsic*
 - * *'llvm.canonicalize.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - * *'llvm.fmuladd.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*

- *Semantics:*
- *Examples:*
- *Experimental Vector Reduction Intrinsic*
 - * `'llvm.experimental.vector.reduce.add.*'` *Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - * `'llvm.experimental.vector.reduce.fadd.*'` *Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Examples:*
 - * `'llvm.experimental.vector.reduce.mul.*'` *Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - * `'llvm.experimental.vector.reduce.fmul.*'` *Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Examples:*
 - * `'llvm.experimental.vector.reduce.and.*'` *Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - * `'llvm.experimental.vector.reduce.or.*'` *Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - * `'llvm.experimental.vector.reduce.xor.*'` *Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - * `'llvm.experimental.vector.reduce.smax.*'` *Intrinsic*
 - *Syntax:*

- *Overview:*
- *Arguments:*
- * `'llvm.experimental.vector.reduce.smin.*'` *Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
- * `'llvm.experimental.vector.reduce.umax.*'` *Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
- * `'llvm.experimental.vector.reduce.umin.*'` *Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
- * `'llvm.experimental.vector.reduce.fmax.*'` *Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
- * `'llvm.experimental.vector.reduce.fmin.*'` *Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
- *Half Precision Floating Point Intrinsics*
 - * `'llvm.convert.to.fp16'` *Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Examples:*
 - * `'llvm.convert.from.fp16'` *Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Examples:*

- *Debugger Intrinsics*
- *Exception Handling Intrinsics*
- *Trampoline Intrinsics*
 - * *`llvm.init.trampoline` Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - * *`llvm.adjust.trampoline` Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- *Masked Vector Load and Store Intrinsics*
 - * *`llvm.masked.load.*` Intrinsics*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - * *`llvm.masked.store.*` Intrinsics*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- *Masked Vector Gather and Scatter Intrinsics*
 - * *`llvm.masked.gather.*` Intrinsics*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - * *`llvm.masked.scatter.*` Intrinsics*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*

– *Memory Use Markers*

* *'llvm.lifetime.start' Intrinsic*

- *Syntax:*
- *Overview:*
- *Arguments:*
- *Semantics:*

* *'llvm.lifetime.end' Intrinsic*

- *Syntax:*
- *Overview:*
- *Arguments:*
- *Semantics:*

* *'llvm.invariant.start' Intrinsic*

- *Syntax:*
- *Overview:*
- *Arguments:*
- *Semantics:*

* *'llvm.invariant.end' Intrinsic*

- *Syntax:*
- *Overview:*
- *Arguments:*
- *Semantics:*

* *'llvm.invariant.group.barrier' Intrinsic*

- *Syntax:*
- *Overview:*
- *Arguments:*
- *Semantics:*

– *Constrained Floating Point Intrinsics*

* *'llvm.experimental.constrained.fadd' Intrinsic*

- *Syntax:*
- *Overview:*
- *Arguments:*
- *Semantics:*

* *'llvm.experimental.constrained.fsub' Intrinsic*

- *Syntax:*
- *Overview:*

- *Arguments:*
- *Semantics:*
- * *'llvm.experimental.constrained.fmul' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.experimental.constrained.fdiv' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.experimental.constrained.frem' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- *Constrained libm-equivalent Intrinsics*
 - * *'llvm.experimental.constrained.sqrt' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - * *'llvm.experimental.constrained.pow' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - * *'llvm.experimental.constrained.powi' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - * *'llvm.experimental.constrained.sin' Intrinsic*
 - *Syntax:*

- *Overview:*
- *Arguments:*
- *Semantics:*
- * *'llvm.experimental.constrained.cos' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.experimental.constrained.exp' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.experimental.constrained.exp2' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.experimental.constrained.log' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.experimental.constrained.log10' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.experimental.constrained.log2' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.experimental.constrained rint' Intrinsic*
 - *Syntax:*

- *Overview:*
- *Arguments:*
- *Semantics:*
- * *'llvm.experimental.constrained.nearbyint' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- *General Ininsics*
 - * *'llvm.var.annotation' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - * *'llvm.ptr.annotation.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - * *'llvm.annotation.*' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - * *'llvm.trap' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - * *'llvm.debugtrap' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - * *'llvm.stackprotector' Intrinsic*

- *Syntax:*
- *Overview:*
- *Arguments:*
- *Semantics:*
- * *'llvm.stackguard' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.objectsize' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.expect' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.assume' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.ssa_copy' Intrinsic*
 - *Syntax:*
 - *Arguments:*
 - *Overview:*
- * *'llvm.type.test' Intrinsic*
 - *Syntax:*
 - *Arguments:*
 - *Overview:*
- * *'llvm.type.checked.load' Intrinsic*
 - *Syntax:*
 - *Arguments:*

- *Overview:*
- * *'llvm.donothing' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
- * *'llvm.experimental.deoptimize' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Lowering:*
- * *'llvm.experimental.guard' Intrinsic*
 - *Syntax:*
 - *Overview:*
- * *'llvm.load.relative' Intrinsic*
 - *Syntax:*
 - *Overview:*
- *Stack Map Intrinsics*
- *Element Wise Atomic Memory Intrinsics*
 - * *'llvm.memcpy.element.unordered.atomic' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Lowering:*
 - * *'llvm.memmove.element.unordered.atomic' Intrinsic*
 - *Syntax:*
 - *Overview:*
 - *Arguments:*
 - *Semantics:*
 - *Lowering:*
 - * *'llvm.memset.element.unordered.atomic' Intrinsic*
 - *Syntax:*
 - *Overview:*

- *Arguments:*
- *Semantics:*
- *Lowering:*

1.1.1 Abstract

This document is a reference manual for the LLVM assembly language. LLVM is a Static Single Assignment (SSA) based representation that provides type safety, low-level operations, flexibility, and the capability of representing ‘all’ high-level languages cleanly. It is the common code representation used throughout all phases of the LLVM compilation strategy.

1.1.2 Introduction

The LLVM code representation is designed to be used in three different forms: as an in-memory compiler IR, as an on-disk bytecode representation (suitable for fast loading by a Just-In-Time compiler), and as a human readable assembly language representation. This allows LLVM to provide a powerful intermediate representation for efficient compiler transformations and analysis, while providing a natural means to debug and visualize the transformations. The three different forms of LLVM are all equivalent. This document describes the human readable representation and notation.

The LLVM representation aims to be light-weight and low-level while being expressive, typed, and extensible at the same time. It aims to be a “universal IR” of sorts, by being at a low enough level that high-level ideas may be cleanly mapped to it (similar to how microprocessors are “universal IR’s”, allowing many source languages to be mapped to them). By providing type information, LLVM can be used as the target of optimizations: for example, through pointer analysis, it can be proven that a C automatic variable is never accessed outside of the current function, allowing it to be promoted to a simple SSA value instead of a memory location.

Well-Formedness

It is important to note that this document describes ‘well formed’ LLVM assembly language. There is a difference between what the parser accepts and what is considered ‘well formed’. For example, the following instruction is syntactically okay, but not well formed:

```
%x = add i32 1, %x
```

because the definition of %x does not dominate all of its uses. The LLVM infrastructure provides a verification pass that may be used to verify that an LLVM module is well formed. This pass is automatically run by the parser after parsing input assembly and by the optimizer before it outputs bytecode. The violations pointed out by the verifier pass indicate bugs in transformation passes or input to the parser.

1.1.3 Identifiers

LLVM identifiers come in two basic types: global and local. Global identifiers (functions, global variables) begin with the '@' character. Local identifiers (register names, types) begin with the '%' character. Additionally, there are three different formats for identifiers, for different purposes:

1. Named values are represented as a string of characters with their prefix. For example, %foo, @DivisionByZero, %a.really.long.identifier. The actual regular expression used is ‘[%@] [-a-zA-Z\$. _] [-a-zA-Z\$. _0-9] *’. Identifiers that require other characters in their names can be surrounded with quotes. Special characters may be escaped using “\xx” where xx is the ASCII code for

the character in hexadecimal. In this way, any character can be used in a name value, even quotes themselves. The "\01" prefix can be used on global variables to suppress mangling.

2. Unnamed values are represented as an unsigned numeric value with their prefix. For example, %12, @2, %44.
3. Constants, which are described in the section *Constants* below.

LLVM requires that values start with a prefix for two reasons: Compilers don't need to worry about name clashes with reserved words, and the set of reserved words may be expanded in the future without penalty. Additionally, unnamed identifiers allow a compiler to quickly come up with a temporary variable without having to avoid symbol table conflicts.

Reserved words in LLVM are very similar to reserved words in other languages. There are keywords for different opcodes ('add', 'bitcast', 'ret', etc...), for primitive type names ('void', 'i32', etc...), and others. These reserved words cannot conflict with variable names, because none of them start with a prefix character ('%' or '@').

Here is an example of LLVM code to multiply the integer variable '%X' by 8:

The easy way:

```
%result = mul i32 %X, 8
```

After strength reduction:

```
%result = shl i32 %X, 3
```

And the hard way:

```
%0 = add i32 %X, %X          ; yields i32:%0
%1 = add i32 %0, %0          ; yields i32:%1
%result = add i32 %1, %1
```

This last way of multiplying %X by 8 illustrates several important lexical features of LLVM:

1. Comments are delimited with a ';' and go until the end of line.
2. Unnamed temporaries are created when the result of a computation is not assigned to a named value.
3. Unnamed temporaries are numbered sequentially (using a per-function incrementing counter, starting with 0). Note that basic blocks and unnamed function parameters are included in this numbering. For example, if the entry basic block is not given a label name and all function parameters are named, then it will get number 0.

It also shows a convention that we follow in this document. When demonstrating instructions, we will follow an instruction with a comment that defines the type and name of value produced.

1.1.4 High Level Structure

Module Structure

LLVM programs are composed of Module's, each of which is a translation unit of the input programs. Each module consists of functions, global variables, and symbol table entries. Modules may be combined together with the LLVM linker, which merges function (and global variable) definitions, resolves forward declarations, and merges symbol table entries. Here is an example of the "hello world" module:

```
; Declare the string constant as a global constant.
@.str = private unnamed_addr constant [13 x i8] c"hello world\0A\00"

; External declaration of the puts function
declare i32 @puts(i8* nocapture) nounwind
```

```
; Definition of main function
define i32 @main() {    ; i32()*
    ; Convert [13 x i8]* to i8*...
    %cast210 = getelementptr [13 x i8], [13 x i8]* @.str, i64 0, i64 0

    ; Call puts function to write out the string to stdout.
    call i32 @puts(i8* %cast210)
    ret i32 0
}

; Named metadata
!0 = !{i32 42, null, !"string"}
!foo = !{!0}
```

This example is made up of a *global variable* named “.str”, an external declaration of the “puts” function, a *function definition* for “main” and *named metadata* “foo”.

In general, a module is made up of a list of global values (where both functions and global variables are global values). Global values are represented by a pointer to a memory location (in this case, a pointer to an array of char, and a pointer to a function), and have one of the following *linkage types*.

Linkage Types

All Global Variables and Functions have one of the following types of linkage:

private Global values with “private” linkage are only directly accessible by objects in the current module.

In particular, linking code into a module with a private global value may cause the private to be renamed as necessary to avoid collisions. Because the symbol is private to the module, all references can be updated. This doesn’t show up in any symbol table in the object file.

internal Similar to private, but the value shows as a local symbol (STB_LOCAL in the case of ELF) in the object file. This corresponds to the notion of the ‘static’ keyword in C.

available_externally Globals with “available_externally” linkage are never emitted into the object file corresponding to the LLVM module. From the linker’s perspective, an `available_externally` global is equivalent to an external declaration. They exist to allow inlining and other optimizations to take place given knowledge of the definition of the global, which is known to be somewhere outside the module. Globals with `available_externally` linkage are allowed to be discarded at will, and allow inlining and other optimizations. This linkage type is only allowed on definitions, not declarations.

linkonce Globals with “linkonce” linkage are merged with other globals of the same name when linkage occurs. This can be used to implement some forms of inline functions, templates, or other code which must be generated in each translation unit that uses it, but where the body may be overridden with a more definitive definition later. Unreferenced `linkonce` globals are allowed to be discarded. Note that `linkonce` linkage does not actually allow the optimizer to inline the body of this function into callers because it doesn’t know if this definition of the function is the definitive definition within the program or whether it will be overridden by a stronger definition. To enable inlining and other optimizations, use “`linkonce_odr`” linkage.

weak “weak” linkage has the same merging semantics as `linkonce` linkage, except that unreferenced globals with weak linkage may not be discarded. This is used for globals that are declared “weak” in C source code.

common “common” linkage is most similar to “weak” linkage, but they are used for tentative definitions in C, such as “`int X;`” at global scope. Symbols with “common” linkage are merged in the same way as `weak` symbols, and they may not be deleted if unreferenced. `common` symbols may not have an explicit section, must have a zero initializer, and may not be marked ‘*constant*’. Functions and aliases may not have common linkage.

appending “appending” linkage may only be applied to global variables of pointer to array type. When two global variables with appending linkage are linked together, the two global arrays are appended together. This is the LLVM, typesafe, equivalent of having the system linker append together “sections” with identical names when .o files are linked.

Unfortunately this doesn’t correspond to any feature in .o files, so it can only be used for variables like `llvm.global_ctors` which `llvm` interprets specially.

extern_weak The semantics of this linkage follow the ELF object file model: the symbol is weak until linked, if not linked, the symbol becomes null instead of being an undefined reference.

linkonce_odr, weak_odr Some languages allow differing globals to be merged, such as two functions with different semantics. Other languages, such as C++, ensure that only equivalent globals are ever merged (the “one definition rule” — “ODR”). Such languages can use the `linkonce_odr` and `weak_odr` linkage types to indicate that the global will only be merged with equivalent globals. These linkage types are otherwise the same as their non-`odr` versions.

external If none of the above identifiers are used, the global is externally visible, meaning that it participates in linkage and can be used to resolve external symbol references.

It is illegal for a function *declaration* to have any linkage type other than `external` or `extern_weak`.

Calling Conventions

LLVM *functions*, *calls* and *invokes* can all have an optional calling convention specified for the call. The calling convention of any pair of dynamic caller/callee must match, or the behavior of the program is undefined. The following calling conventions are supported by LLVM, and more may be added in the future:

“ccc” - The C calling convention This calling convention (the default if no other calling convention is specified) matches the target C calling conventions. This calling convention supports varargs function calls and tolerates some mismatch in the declared prototype and implemented declaration of the function (as does normal C).

“fastcc” - The fast calling convention This calling convention attempts to make calls as fast as possible (e.g. by passing things in registers). This calling convention allows the target to use whatever tricks it wants to produce fast code for the target, without having to conform to an externally specified ABI (Application Binary Interface). *Tail calls can only be optimized when this, the GHC or the HiPE convention is used.* This calling convention does not support varargs and requires the prototype of all callees to exactly match the prototype of the function definition.

“coldcc” - The cold calling convention This calling convention attempts to make code in the caller as efficient as possible under the assumption that the call is not commonly executed. As such, these calls often preserve all registers so that the call does not break any live ranges in the caller side. This calling convention does not support varargs and requires the prototype of all callees to exactly match the prototype of the function definition. Furthermore the inliner doesn’t consider such function calls for inlining.

“cc 10” - GHC convention This calling convention has been implemented specifically for use by the [Glasgow Haskell Compiler \(GHC\)](#). It passes everything in registers, going to extremes to achieve this by disabling callee save registers. This calling convention should not be used lightly but only for specific situations such as an alternative to the *register pinning* performance technique often used when implementing functional programming languages. At the moment only X86 supports this convention and it has the following limitations:

- On X86-32 only supports up to 4 bit type parameters. No floating point types are supported.
- On X86-64 only supports up to 10 bit type parameters and 6 floating point parameters.

This calling convention supports [tail call optimization](#) but requires both the caller and callee are using it.

“cc 11” - The HiPE calling convention This calling convention has been implemented specifically for use by the [High-Performance Erlang \(HiPE\)](#) compiler, the native code compiler of the [Ericsson’s Open Source Erlang/OTP](#)

system. It uses more registers for argument passing than the ordinary C calling convention and defines no callee-saved registers. The calling convention properly supports *tail call optimization* but requires that both the caller and the callee use it. It uses a *register pinning* mechanism, similar to GHC's convention, for keeping frequently accessed runtime components pinned to specific hardware registers. At the moment only X86 supports this convention (both 32 and 64 bit).

“webkit_jscc” - WebKit's JavaScript calling convention This calling convention has been implemented for *WebKit FTL JIT*. It passes arguments on the stack right to left (as *cdecl* does), and returns a value in the platform's customary return register.

“anyregcc” - Dynamic calling convention for code patching This is a special convention that supports patching an arbitrary code sequence in place of a call site. This convention forces the call arguments into registers but allows them to be dynamically allocated. This can currently only be used with calls to `llvm.experimental.patchpoint` because only this intrinsic records the location of its arguments in a side table. See *Stack maps and patch points in LLVM*.

“preserve_mostcc” - The *PreserveMost* calling convention This calling convention attempts to make the code in the caller as unintrusive as possible. This convention behaves identically to the C calling convention on how arguments and return values are passed, but it uses a different set of caller/callee-saved registers. This alleviates the burden of saving and recovering a large register set before and after the call in the caller. If the arguments are passed in callee-saved registers, then they will be preserved by the callee across the call. This doesn't apply for values returned in callee-saved registers.

- On X86-64 the callee preserves all general purpose registers, except for R11. R11 can be used as a scratch register. Floating-point registers (XMMs/YMMs) are not preserved and need to be saved by the caller.

The idea behind this convention is to support calls to runtime functions that have a hot path and a cold path. The hot path is usually a small piece of code that doesn't use many registers. The cold path might need to call out to another function and therefore only needs to preserve the caller-saved registers, which haven't already been saved by the caller. The *PreserveMost* calling convention is very similar to the *coldcc* calling convention in terms of caller/callee-saved registers, but they are used for different types of function calls. *coldcc* is for function calls that are rarely executed, whereas *preserve_mostcc* function calls are intended to be on the hot path and definitely executed a lot. Furthermore *preserve_mostcc* doesn't prevent the inliner from inlining the function call.

This calling convention will be used by a future version of the ObjectiveC runtime and should therefore still be considered experimental at this time. Although this convention was created to optimize certain runtime calls to the ObjectiveC runtime, it is not limited to this runtime and might be used by other runtimes in the future too. The current implementation only supports X86-64, but the intention is to support more architectures in the future.

“preserve_allcc” - The *PreserveAll* calling convention This calling convention attempts to make the code in the caller even less intrusive than the *PreserveMost* calling convention. This calling convention also behaves identical to the C calling convention on how arguments and return values are passed, but it uses a different set of caller/callee-saved registers. This removes the burden of saving and recovering a large register set before and after the call in the caller. If the arguments are passed in callee-saved registers, then they will be preserved by the callee across the call. This doesn't apply for values returned in callee-saved registers.

- On X86-64 the callee preserves all general purpose registers, except for R11. R11 can be used as a scratch register. Furthermore it also preserves all floating-point registers (XMMs/YMMs).

The idea behind this convention is to support calls to runtime functions that don't need to call out to any other functions.

This calling convention, like the *PreserveMost* calling convention, will be used by a future version of the ObjectiveC runtime and should be considered experimental at this time.

“cxx_fast_tlscc” - The *CXX_FAST_TLS* calling convention for access functions Clang generates an access function to access C++-style TLS. The access function generally has an entry block, an exit block and an

initialization block that is run at the first time. The entry and exit blocks can access a few TLS IR variables, each access will be lowered to a platform-specific sequence.

This calling convention aims to minimize overhead in the caller by preserving as many registers as possible (all the registers that are preserved on the fast path, composed of the entry and exit blocks).

This calling convention behaves identical to the *C* calling convention on how arguments and return values are passed, but it uses a different set of caller/callee-saved registers.

Given that each platform has its own lowering sequence, hence its own set of preserved registers, we can't use the existing *PreserveMost*.

- On X86-64 the callee preserves all general purpose registers, except for RDI and RAX.

“swiftcc” - This calling convention is used for Swift language.

- On X86-64 RCX and R8 are available for additional integer returns, and XMM2 and XMM3 are available for additional FP/vector returns.
- On iOS platforms, we use AAPCS-VFP calling convention.

“cc <n>” - Numbered convention Any calling convention may be specified by number, allowing target-specific calling conventions to be used. Target specific calling conventions start at 64.

More calling conventions can be added/defined on an as-needed basis, to support Pascal conventions or any other well-known target-independent convention.

Visibility Styles

All Global Variables and Functions have one of the following visibility styles:

“default” - Default style On targets that use the ELF object file format, default visibility means that the declaration is visible to other modules and, in shared libraries, means that the declared entity may be overridden. On Darwin, default visibility means that the declaration is visible to other modules. Default visibility corresponds to “external linkage” in the language.

“hidden” - Hidden style Two declarations of an object with hidden visibility refer to the same object if they are in the same shared object. Usually, hidden visibility indicates that the symbol will not be placed into the dynamic symbol table, so no other module (executable or shared library) can reference it directly.

“protected” - Protected style On ELF, protected visibility indicates that the symbol will be placed in the dynamic symbol table, but that references within the defining module will bind to the local symbol. That is, the symbol cannot be overridden by another module.

A symbol with `internal` or `private` linkage must have default visibility.

DLL Storage Classes

All Global Variables, Functions and Aliases can have one of the following DLL storage class:

`dllimport` “`dllimport`” causes the compiler to reference a function or variable via a global pointer to a pointer that is set up by the DLL exporting the symbol. On Microsoft Windows targets, the pointer name is formed by combining `__imp_` and the function or variable name.

`dlexport` “`dlexport`” causes the compiler to provide a global pointer to a pointer in a DLL, so that it can be referenced with the `dllimport` attribute. On Microsoft Windows targets, the pointer name is formed by combining `__imp_` and the function or variable name. Since this storage class exists for defining a dll interface, the compiler, assembler and linker know it is externally referenced and must refrain from deleting the symbol.

Thread Local Storage Models

A variable may be defined as `thread_local`, which means that it will not be shared by threads (each thread will have a separated copy of the variable). Not all targets support thread-local variables. Optionally, a TLS model may be specified:

localdynamic For variables that are only used within the current shared library.

initialexec For variables in modules that will not be loaded dynamically.

localexec For variables defined in the executable and only used within it.

If no explicit model is given, the “general dynamic” model is used.

The models correspond to the ELF TLS models; see [ELF Handling For Thread-Local Storage](#) for more information on under which circumstances the different models may be used. The target may choose a different TLS model if the specified model is not supported, or if a better choice of model can be made.

A model can also be specified in an alias, but then it only governs how the alias is accessed. It will not have any effect in the aliasee.

For platforms without linker support of ELF TLS model, the `-femulated-tls` flag can be used to generate GCC compatible emulated TLS code.

Structure Types

LLVM IR allows you to specify both “identified” and “literal” *structure types*. Literal types are unique structurally, but identified types are never unique. An *opaque structural type* can also be used to forward declare a type that is not yet available.

An example of an identified structure specification is:

```
%mytype = type { %mytype*, i32 }
```

Prior to the LLVM 3.0 release, identified types were structurally unique. Only literal types are unique in recent versions of LLVM.

Non-Integral Pointer Type

Note: non-integral pointer types are a work in progress, and they should be considered experimental at this time.

LLVM IR optionally allows the frontend to denote pointers in certain address spaces as “non-integral” via the *data-layout string*. Non-integral pointer types represent pointers that have an *unspecified* bitwise representation; that is, the integral representation may be target dependent or unstable (not backed by a fixed integer).

`inttoptr` instructions converting integers to non-integral pointer types are ill-typed, and so are `ptrtoint` instructions converting values of non-integral pointer types to integers. Vector versions of said instructions are ill-typed as well.

Global Variables

Global variables define regions of memory allocated at compilation time instead of run-time.

Global variable definitions must be initialized.

Global variables in other translation units can also be declared, in which case they don’t have an initializer.

Either global variable definitions or declarations may have an explicit section to be placed in and may have an optional explicit alignment specified.

A variable may be defined as a global `constant`, which indicates that the contents of the variable will **never** be modified (enabling better optimization, allowing the global data to be placed in the read-only section of an executable, etc). Note that variables that need runtime initialization cannot be marked `constant` as there is a store to the variable.

LLVM explicitly allows *declarations* of global variables to be marked constant, even if the final definition of the global is not. This capability can be used to enable slightly better optimization of the program, but requires the language definition to guarantee that optimizations based on the ‘constantness’ are valid for the translation units that do not include the definition.

As SSA values, global variables define pointer values that are in scope (i.e. they dominate) all basic blocks in the program. Global variables always define a pointer to their “content” type because they describe a region of memory, and all memory objects in LLVM are accessed through pointers.

Global variables can be marked with `unnamed_addr` which indicates that the address is not significant, only the content. Constants marked like this can be merged with other constants if they have the same initializer. Note that a constant with significant address *can* be merged with a `unnamed_addr` constant, the result being a constant whose address is significant.

If the `local_unnamed_addr` attribute is given, the address is known to not be significant within the module.

A global variable may be declared to reside in a target-specific numbered address space. For targets that support them, address spaces may affect how optimizations are performed and/or what target instructions are used to access the variable. The default address space is zero. The address space qualifier must precede any other attributes.

LLVM allows an explicit section to be specified for globals. If the target supports it, it will emit globals to the section specified. Additionally, the global can be placed in a `comdat` if the target has the necessary support.

By default, global initializers are optimized by assuming that global variables defined within the module are not modified from their initial values before the start of the global initializer. This is true even for variables potentially accessible from outside the module, including those with external linkage or appearing in `@llvm.used` or `dllexport` variables. This assumption may be suppressed by marking the variable with `externally_initialized`.

An explicit alignment may be specified for a global, which must be a power of 2. If not present, or if the alignment is set to zero, the alignment of the global is set by the target to whatever it feels convenient. If an explicit alignment is specified, the global is forced to have exactly that alignment. Targets and optimizers are not allowed to over-align the global if the global has an assigned section. In this case, the extra alignment could be observable: for example, code could assume that the globals are densely packed in their section and try to iterate over them as an array, alignment padding would break this iteration. The maximum alignment is $1 \ll 29$.

Globals can also have a *DLL storage class*, an optional *global attributes* and an optional list of attached *metadata*.

Variables and aliases can have a *Thread Local Storage Model*.

Syntax:

```
@<GlobalVarName> = [Linkage] [Visibility] [DLLStorageClass] [ThreadLocal]
                   [(unnamed_addr|local_unnamed_addr)] [AddrSpace]
                   [ExternallyInitialized]
                   <global | constant> <Type> [<InitializerConstant>]
                   [, section "name"] [, comdat [($name)]]
                   [, align <Alignment>] (, !name !N)*
```

For example, the following defines a global in a numbered address space with an initializer, section, and alignment:

```
@G = addrspace(5) constant float 1.0, section "foo", align 4
```

The following example just declares a global variable

```
@G = external global i32
```

The following example defines a thread-local global with the `initialexec` TLS model:

```
@G = thread_local(initialexec) global i32 0, align 4
```

Functions

LLVM function definitions consist of the “define” keyword, an optional *linkage type*, an optional *visibility style*, an optional *DLL storage class*, an optional *calling convention*, an optional `unnamed_addr` attribute, a return type, an optional *parameter attribute* for the return type, a function name, a (possibly empty) argument list (each with optional *parameter attributes*), optional *function attributes*, an optional section, an optional alignment, an optional *comdat*, an optional *garbage collector name*, an optional *prefix*, an optional *prologue*, an optional *personality*, an optional list of attached *metadata*, an opening curly brace, a list of basic blocks, and a closing curly brace.

LLVM function declarations consist of the “declare” keyword, an optional *linkage type*, an optional *visibility style*, an optional *DLL storage class*, an optional *calling convention*, an optional `unnamed_addr` or `local_unnamed_addr` attribute, a return type, an optional *parameter attribute* for the return type, a function name, a possibly empty list of arguments, an optional alignment, an optional *garbage collector name*, an optional *prefix*, and an optional *prologue*.

A function definition contains a list of basic blocks, forming the CFG (Control Flow Graph) for the function. Each basic block may optionally start with a label (giving the basic block a symbol table entry), contains a list of instructions, and ends with a *terminator* instruction (such as a branch or function return). If an explicit label is not provided, a block is assigned an implicit numbered label, using the next value from the same counter as used for unnamed temporaries (*see above*). For example, if a function entry block does not have an explicit label, it will be assigned label “%0”, then the first unnamed temporary in that block will be “%1”, etc.

The first basic block in a function is special in two ways: it is immediately executed on entrance to the function, and it is not allowed to have predecessor basic blocks (i.e. there can not be any branches to the entry block of a function). Because the block can have no predecessors, it also cannot have any *PHI nodes*.

LLVM allows an explicit section to be specified for functions. If the target supports it, it will emit functions to the section specified. Additionally, the function can be placed in a COMDAT.

An explicit alignment may be specified for a function. If not present, or if the alignment is set to zero, the alignment of the function is set by the target to whatever it feels convenient. If an explicit alignment is specified, the function is forced to have at least that much alignment. All alignments must be a power of 2.

If the `unnamed_addr` attribute is given, the address is known to not be significant and two identical functions can be merged.

If the `local_unnamed_addr` attribute is given, the address is known to not be significant within the module.

Syntax:

```
define [linkage] [visibility] [DLLStorageClass]
  [cconv] [ret attrs]
  <ResultType> @<FunctionName> ([argument list])
  [(unnamed_addr|local_unnamed_addr)] [fn Attrs] [section "name"]
  [comdat [($name)]] [align N] [gc] [prefix Constant]
  [prologue Constant] [personality Constant] (!name !N)* { ... }
```

The argument list is a comma separated sequence of arguments where each argument is of the following form:

Syntax:

```
<type> [parameter Attrs] [name]
```

Aliases

Aliases, unlike function or variables, don't create any new data. They are just a new symbol and metadata for an existing position.

Aliases have a name and an aliasee that is either a global value or a constant expression.

Aliases may have an optional *linkage type*, an optional *visibility style*, an optional *DLL storage class* and an optional *tls model*.

Syntax:

```
@<Name> = [Linkage] [Visibility] [DLLStorageClass] [ThreadLocal] [(unnamed_addr|local_
↳unnamed_addr)] alias <AliaseeTy>, <AliaseeTy>* @<Aliasee>
```

The linkage must be one of `private`, `internal`, `linkonce`, `weak`, `linkonce_odr`, `weak_odr`, `external`. Note that some system linkers might not correctly handle dropping a weak symbol that is aliased.

Aliases that are not `unnamed_addr` are guaranteed to have the same address as the aliasee expression. `unnamed_addr` ones are only guaranteed to point to the same content.

If the `local_unnamed_addr` attribute is given, the address is known to not be significant within the module.

Since aliases are only a second name, some restrictions apply, of which some can only be checked when producing an object file:

- The expression defining the aliasee must be computable at assembly time. Since it is just a name, no relocations can be used.
- No alias in the expression can be weak as the possibility of the intermediate alias being overridden cannot be represented in an object file.
- No global value in the expression can be a declaration, since that would require a relocation, which is not possible.

IFuncs

IFuncs, like as aliases, don't create any new data or func. They are just a new symbol that dynamic linker resolves at runtime by calling a resolver function.

IFuncs have a name and a resolver that is a function called by dynamic linker that returns address of another function associated with the name.

IFunc may have an optional *linkage type* and an optional *visibility style*.

Syntax:

```
@<Name> = [Linkage] [Visibility] ifunc <IFuncTy>, <ResolverTy>* @<Resolver>
```

Comdats

Comdat IR provides access to COFF and ELF object file COMDAT functionality.

Comdats have a name which represents the COMDAT key. All global objects that specify this key will only end up in the final object file if the linker chooses that key over some other key. Aliases are placed in the same COMDAT that their aliasee computes to, if any.

Comdats have a selection kind to provide input on how the linker should choose between keys in two different object files.

Syntax:

```
$<Name> = comdat SelectionKind
```

The selection kind must be one of the following:

any The linker may choose any COMDAT key, the choice is arbitrary.

exactmatch The linker may choose any COMDAT key but the sections must contain the same data.

largest The linker will choose the section containing the largest COMDAT key.

noduplicates The linker requires that only section with this COMDAT key exist.

samesize The linker may choose any COMDAT key but the sections must contain the same amount of data.

Note that the Mach-O platform doesn't support COMDATs and ELF only supports **any** as a selection kind.

Here is an example of a COMDAT group where a function will only be selected if the COMDAT key's section is the largest:

```
$foo = comdat largest
@foo = global i32 2, comdat($foo)

define void @bar() comdat($foo) {
    ret void
}
```

As a syntactic sugar the `$name` can be omitted if the name is the same as the global name:

```
$foo = comdat any
@foo = global i32 2, comdat
```

In a COFF object file, this will create a COMDAT section with selection kind `IMAGE_COMDAT_SELECT_LARGEST` containing the contents of the `@foo` symbol and another COMDAT section with selection kind `IMAGE_COMDAT_SELECT_ASSOCIATIVE` which is associated with the first COMDAT section and contains the contents of the `@bar` symbol.

There are some restrictions on the properties of the global object. It, or an alias to it, must have the same name as the COMDAT group when targeting COFF. The contents and size of this object may be used during link-time to determine which COMDAT groups get selected depending on the selection kind. Because the name of the object must match the name of the COMDAT group, the linkage of the global object must not be local; local symbols can get renamed if a collision occurs in the symbol table.

The combined use of COMDATS and section attributes may yield surprising results. For example:

```
$foo = comdat any
$bar = comdat any
@g1 = global i32 42, section "sec", comdat($foo)
@g2 = global i32 42, section "sec", comdat($bar)
```

From the object file perspective, this requires the creation of two sections with the same name. This is necessary because both globals belong to different COMDAT groups and COMDATs, at the object file level, are represented by sections.

Note that certain IR constructs like global variables and functions may create COMDATs in the object file in addition to any which are specified using COMDAT IR. This arises when the code generator is configured to emit globals in individual sections (e.g. when `-data-sections` or `-function-sections` is supplied to `llc`).

Named Metadata

Named metadata is a collection of metadata. *Metadata nodes* (but not metadata strings) are the only valid operands for a named metadata.

1. Named metadata are represented as a string of characters with the metadata prefix. The rules for metadata names are the same as for identifiers, but quoted names are not allowed. "`\xx`" type escapes are still valid, which allows any character to be part of a name.

Syntax:

```
; Some unnamed metadata nodes, which are referenced by the named metadata.
!0 = !{"zero"}
!1 = !{"one"}
!2 = !{"two"}
; A named metadata.
!name = !{!0, !1, !2}
```

Parameter Attributes

The return type and each parameter of a function type may have a set of *parameter attributes* associated with them. Parameter attributes are used to communicate additional information about the result or parameters of a function. Parameter attributes are considered to be part of the function, not of the function type, so functions with different parameter attributes can have the same function type.

Parameter attributes are simple keywords that follow the type specified. If multiple parameter attributes are needed, they are space separated. For example:

```
declare i32 @printf(i8* noalias nocapture, ...)
declare i32 @atoi(i8 zeroext)
declare signext i8 @returns_signed_char()
```

Note that any attributes for the function result (`nounwind`, `readonly`) come immediately after the argument list.

Currently, only the following parameter attributes are defined:

zeroext This indicates to the code generator that the parameter or return value should be zero-extended to the extent required by the target's ABI by the caller (for a parameter) or the callee (for a return value).

signext This indicates to the code generator that the parameter or return value should be sign-extended to the extent required by the target's ABI (which is usually 32-bits) by the caller (for a parameter) or the callee (for a return value).

inreg This indicates that this parameter or return value should be treated in a special target-dependent fashion while emitting code for a function call or return (usually, by putting it in a register as opposed to memory, though some targets use it to distinguish between two different kinds of registers). Use of this attribute is target-specific.

byval This indicates that the pointer parameter should really be passed by value to the function. The attribute implies that a hidden copy of the pointee is made between the caller and the callee, so the callee is unable to modify the value in the caller. This attribute is only valid on LLVM pointer arguments. It is generally used to pass structs and arrays by value, but is also valid on pointers to scalars. The copy is considered to belong to the caller not the callee (for example, `readonly` functions should not write to `byval` parameters). This is not a valid attribute for return values.

The `byval` attribute also supports specifying an alignment with the `align` attribute. It indicates the alignment of the stack slot to form and the known alignment of the pointer specified to the call site. If the alignment is not specified, then the code generator makes a target-specific assumption.

`inalloca`

The `inalloca` argument attribute allows the caller to take the address of outgoing stack arguments. An `inalloca` argument must be a pointer to stack memory produced by an `alloca` instruction. The `alloca`, or argument allocation, must also be tagged with the `inalloca` keyword. Only the last argument may have the `inalloca` attribute, and that argument is guaranteed to be passed in memory.

An argument allocation may be used by a call at most once because the call may deallocate it. The `inalloca` attribute cannot be used in conjunction with other attributes that affect argument storage, like `inreg`, `nest`, `sret`, or `byval`. The `inalloca` attribute also disables LLVM's implicit lowering of large aggregate return values, which means that frontend authors must lower them with `sret` pointers.

When the call site is reached, the argument allocation must have been the most recent stack allocation that is still live, or the results are undefined. It is possible to allocate additional stack space after an argument allocation and before its call site, but it must be cleared off with *llvm.stackrestore*.

See *Design and Usage of the InAlloca Attribute* for more information on how to use this attribute.

sret This indicates that the pointer parameter specifies the address of a structure that is the return value of the function in the source program. This pointer must be guaranteed by the caller to be valid: loads and stores to the structure may be assumed by the callee not to trap and to be properly aligned. This is not a valid attribute for return values.

align <n> This indicates that the pointer value may be assumed by the optimizer to have the specified alignment.

Note that this attribute has additional semantics when combined with the `byval` attribute.

noalias This indicates that objects accessed via pointer values *based* on the argument or return value are not also accessed, during the execution of the function, via pointer values not *based* on the argument or return value. The attribute on a return value also has additional semantics described below. The caller shares the responsibility with the callee for ensuring that these requirements are met. For further details, please see the discussion of the NoAlias response in *alias analysis*.

Note that this definition of `noalias` is intentionally similar to the definition of `restrict` in C99 for function arguments.

For function return values, C99's `restrict` is not meaningful, while LLVM's `noalias` is. Furthermore, the semantics of the `noalias` attribute on return values are stronger than the semantics of the attribute when used on function arguments. On function return values, the `noalias` attribute indicates that the function acts like a system memory allocation function, returning a pointer to allocated storage disjoint from the storage for any other object accessible to the caller.

nocapture This indicates that the callee does not make any copies of the pointer that outlive the callee itself. This is not a valid attribute for return values. Addresses used in volatile operations are considered to be captured.

nest This indicates that the pointer parameter can be excised using the *trampoline intrinsics*. This is not a valid attribute for return values and can only be applied to one parameter.

returned This indicates that the function always returns the argument as its return value. This is a hint to the optimizer and code generator used when generating the caller, allowing value propagation, tail call optimization, and omission of register saves and restores in some cases; it is not checked or enforced when generating the callee. The parameter and the function return type must be valid operands for the *bitcast instruction*. This is not a valid attribute for return values and can only be applied to one parameter.

nonnull This indicates that the parameter or return pointer is not null. This attribute may only be applied to pointer typed parameters. This is not checked or enforced by LLVM, the caller must ensure that the pointer passed in is non-null, or the callee must ensure that the returned pointer is non-null.

dereferenceable (<n>) This indicates that the parameter or return pointer is dereferenceable. This attribute may only be applied to pointer typed parameters. A pointer that is dereferenceable can be loaded from speculatively without a risk of trapping. The number of bytes known to be dereferenceable must be provided in parentheses. It is legal for the number of bytes to be less than the size of the pointee type. The `nonnull`

attribute does not imply dereferenceability (consider a pointer to one element past the end of an array), however `dereferenceable(<n>)` does imply `nonnull` in `addrspace(0)` (which is the default address space).

`dereferenceable_or_null(<n>)` This indicates that the parameter or return value isn't both non-null and non-dereferenceable (up to `<n>` bytes) at the same time. All non-null pointers tagged with `dereferenceable_or_null(<n>)` are `dereferenceable(<n>)`. For address space 0 `dereferenceable_or_null(<n>)` implies that a pointer is exactly one of `dereferenceable(<n>)` or `null`, and in other address spaces `dereferenceable_or_null(<n>)` implies that a pointer is at least one of `dereferenceable(<n>)` or `null` (i.e. it may be both `null` and `dereferenceable(<n>)`). This attribute may only be applied to pointer typed parameters.

`swiftself` This indicates that the parameter is the self/context parameter. This is not a valid attribute for return values and can only be applied to one parameter.

`swifterror` This attribute is motivated to model and optimize Swift error handling. It can be applied to a parameter with pointer to pointer type or a pointer-sized alloca. At the call site, the actual argument that corresponds to a `swifterror` parameter has to come from a `swifterror` alloca or the `swifterror` parameter of the caller. A `swifterror` value (either the parameter or the alloca) can only be loaded and stored from, or used as a `swifterror` argument. This is not a valid attribute for return values and can only be applied to one parameter.

These constraints allow the calling convention to optimize access to `swifterror` variables by associating them with a specific register at call boundaries rather than placing them in memory. Since this does change the calling convention, a function which uses the `swifterror` attribute on a parameter is not ABI-compatible with one which does not.

These constraints also allow LLVM to assume that a `swifterror` argument does not alias any other memory visible within a function and that a `swifterror` alloca passed as an argument does not escape.

Garbage Collector Strategy Names

Each function may specify a garbage collector strategy name, which is simply a string:

```
define void @f() gc "name" { ... }
```

The supported values of `name` includes those *built in to LLVM* and any provided by loaded plugins. Specifying a GC strategy will cause the compiler to alter its output in order to support the named garbage collection algorithm. Note that LLVM itself does not contain a garbage collector, this functionality is restricted to generating machine code which can interoperate with a collector provided externally.

Prefix Data

Prefix data is data associated with a function which the code generator will emit immediately before the function's entrypoint. The purpose of this feature is to allow frontends to associate language-specific runtime metadata with specific functions and make it available through the function pointer while still allowing the function pointer to be called.

To access the data for a given function, a program may bitcast the function pointer to a pointer to the constant's type and dereference index -1. This implies that the IR symbol points just past the end of the prefix data. For instance, take the example of a function annotated with a single `i32`,

```
define void @f() prefix i32 123 { ... }
```

The prefix data can be referenced as,

```
%0 = bitcast void* () @f to i32*
@a = getelementptr inbounds i32, i32* %0, i32 -1
%b = load i32, i32* @a
```

Prefix data is laid out as if it were an initializer for a global variable of the prefix data's type. The function will be placed such that the beginning of the prefix data is aligned. This means that if the size of the prefix data is not a multiple of the alignment size, the function's entrypoint will not be aligned. If alignment of the function's entrypoint is desired, padding must be added to the prefix data.

A function may have prefix data but no body. This has similar semantics to the `available_externally` linkage in that the data may be used by the optimizers but will not be emitted in the object file.

Prologue Data

The `prologue` attribute allows arbitrary code (encoded as bytes) to be inserted prior to the function body. This can be used for enabling function hot-patching and instrumentation.

To maintain the semantics of ordinary function calls, the prologue data must have a particular format. Specifically, it must begin with a sequence of bytes which decode to a sequence of machine instructions, valid for the module's target, which transfer control to the point immediately succeeding the prologue data, without performing any other visible action. This allows the inliner and other passes to reason about the semantics of the function definition without needing to reason about the prologue data. Obviously this makes the format of the prologue data highly target dependent.

A trivial example of valid prologue data for the x86 architecture is `i8 144`, which encodes the `nop` instruction:

```
define void @f() prologue i8 144 { ... }
```

Generally prologue data can be formed by encoding a relative branch instruction which skips the metadata, as in this example of valid prologue data for the `x86_64` architecture, where the first two bytes encode `jmp .+10`:

```
%0 = type <{ i8, i8, i8* }>

define void @f() prologue %0 <{ i8 235, i8 8, i8* @md}> { ... }
```

A function may have prologue data but no body. This has similar semantics to the `available_externally` linkage in that the data may be used by the optimizers but will not be emitted in the object file.

Personality Function

The `personality` attribute permits functions to specify what function to use for exception handling.

Attribute Groups

Attribute groups are groups of attributes that are referenced by objects within the IR. They are important for keeping `.ll` files readable, because a lot of functions will use the same set of attributes. In the degenerative case of a `.ll` file that corresponds to a single `.c` file, the single attribute group will capture the important command line flags used to build that file.

An attribute group is a module-level object. To use an attribute group, an object references the attribute group's ID (e.g. `#37`). An object may refer to more than one attribute group. In that situation, the attributes from the different groups are merged.

Here is an example of attribute groups for a function that should always be inlined, has a stack alignment of 4, and which shouldn't use SSE instructions:


```

; Target-independent attributes:
attributes #0 = { alwaysinline alignstack=4 }

; Target-dependent attributes:
attributes #1 = { "no-sse" }

; Function @f has attributes: alwaysinline, alignstack=4, and "no-sse".
define void @f() #0 #1 { ... }

```

Function Attributes

Function attributes are set to communicate additional information about a function. Function attributes are considered to be part of the function, not of the function type, so functions with different function attributes can have the same function type.

Function attributes are simple keywords that follow the type specified. If multiple attributes are needed, they are space separated. For example:

```

define void @f() noinline { ... }
define void @f() alwaysinline { ... }
define void @f() alwaysinline optsize { ... }
define void @f() optsize { ... }

```

alignstack(<n>) This attribute indicates that, when emitting the prologue and epilogue, the backend should forcibly align the stack pointer. Specify the desired alignment, which must be a power of two, in parentheses.

allocsize(<EltSizeParam>[, <NumEltsParam>]) This attribute indicates that the annotated function will always return at least a given number of bytes (or null). Its arguments are zero-indexed parameter numbers; if one argument is provided, then it's assumed that at least `CallSite.Args[EltSizeParam]` bytes will be available at the returned pointer. If two are provided, then it's assumed that `CallSite.Args[EltSizeParam] * CallSite.Args[NumEltsParam]` bytes are available. The referenced parameters must be integer types. No assumptions are made about the contents of the returned block of memory.

alwaysinline This attribute indicates that the inliner should attempt to inline this function into callers whenever possible, ignoring any active inlining size threshold for this caller.

builtin This indicates that the callee function at a call site should be recognized as a built-in function, even though the function's declaration uses the `nobuiltin` attribute. This is only valid at call sites for direct calls to functions that are declared with the `nobuiltin` attribute.

cold This attribute indicates that this function is rarely called. When computing edge weights, basic blocks post-dominated by a cold function call are also considered to be cold; and, thus, given low weight.

convergent In some parallel execution models, there exist operations that cannot be made control-dependent on any additional values. We call such operations *convergent*, and mark them with this attribute.

The `convergent` attribute may appear on functions or call/invoke instructions. When it appears on a function, it indicates that calls to this function should not be made control-dependent on additional values. For example, the intrinsic `llvm.nvvm.barrier0` is convergent, so calls to this intrinsic cannot be made control-dependent on additional values.

When it appears on a call/invoke, the `convergent` attribute indicates that we should treat the call as though we're calling a convergent function. This is particularly useful on indirect calls; without this we may treat such calls as though the target is non-convergent.

The optimizer may remove the `convergent` attribute on functions when it can prove that the function does not execute any convergent operations. Similarly, the optimizer may remove `convergent` on calls/invokes when it can prove that the call/invoke cannot call a convergent function.

inaccessiblememory This attribute indicates that the function may only access memory that is not accessible by the module being compiled. This is a weaker form of `readnone`.

inaccessiblemem_or_argmemory This attribute indicates that the function may only access memory that is either not accessible by the module being compiled, or is pointed to by its pointer arguments. This is a weaker form of `argmemory`

inlinehint This attribute indicates that the source code contained a hint that inlining this function is desirable (such as the “inline” keyword in C/C++). It is just a hint; it imposes no requirements on the inliner.

jumpable This attribute indicates that the function should be added to a jump-instruction table at code-generation time, and that all address-taken references to this function should be replaced with a reference to the appropriate jump-instruction-table function pointer. Note that this creates a new pointer for the original function, which means that code that depends on function-pointer identity can break. So, any function annotated with `jumpable` must also be `unnamed_addr`.

minsize This attribute suggests that optimization passes and code generator passes make choices that keep the code size of this function as small as possible and perform optimizations that may sacrifice runtime performance in order to minimize the size of the generated code.

naked This attribute disables prologue / epilogue emission for the function. This can have very system-specific consequences.

nobuiltin This indicates that the callee function at a call site is not recognized as a built-in function. LLVM will retain the original call and not replace it with equivalent code based on the semantics of the built-in function, unless the call site uses the `builtin` attribute. This is valid at call sites and on function declarations and definitions.

noduplicate This attribute indicates that calls to the function cannot be duplicated. A call to a `noduplicate` function may be moved within its parent function, but may not be duplicated within its parent function.

A function containing a `noduplicate` call may still be an inlining candidate, provided that the call is not duplicated by inlining. That implies that the function has internal linkage and only has one call site, so the original call is dead after inlining.

noimplicitfloat This attribute disables implicit floating point instructions.

noinline This attribute indicates that the inliner should never inline this function in any situation. This attribute may not be used together with the `alwaysinline` attribute.

nonlazybind This attribute suppresses lazy symbol binding for the function. This may make calls to the function faster, at the cost of extra program startup time if the function is not called during program startup.

noredzone This attribute indicates that the code generator should not use a red zone, even if the target-specific ABI normally permits it.

noreturn This function attribute indicates that the function never returns normally. This produces undefined behavior at runtime if the function ever does dynamically return.

norecurse This function attribute indicates that the function does not call itself either directly or indirectly down any possible call path. This produces undefined behavior at runtime if the function ever does recurse.

nounwind This function attribute indicates that the function never raises an exception. If the function does raise an exception, its runtime behavior is undefined. However, functions marked `nounwind` may still trap or generate asynchronous exceptions. Exception handling schemes that are recognized by LLVM to handle asynchronous exceptions, such as SEH, will still provide their implementation defined semantics.

optnone This function attribute indicates that most optimization passes will skip this function, with the exception of interprocedural optimization passes. Code generation defaults to the “fast” instruction selector. This attribute cannot be used together with the `alwaysinline` attribute; this attribute is also incompatible with the `minsize` attribute and the `optsize` attribute.

This attribute requires the `noinline` attribute to be specified on the function as well, so the function is never inlined into any caller. Only functions with the `alwaysinline` attribute are valid candidates for inlining into the body of this function.

optsize This attribute suggests that optimization passes and code generator passes make choices that keep the code size of this function low, and otherwise do optimizations specifically to reduce code size as long as they do not significantly impact runtime performance.

"patchable-function" This attribute tells the code generator that the code generated for this function needs to follow certain conventions that make it possible for a runtime function to patch over it later. The exact effect of this attribute depends on its string value, for which there currently is one legal possibility:

- **"prologue-short-redirect"** - This style of patchable function is intended to support patching a function prologue to redirect control away from the function in a thread safe manner. It guarantees that the first instruction of the function will be large enough to accommodate a short jump instruction, and will be sufficiently aligned to allow being fully changed via an atomic compare-and-swap instruction. While the first requirement can be satisfied by inserting large enough NOP, LLVM can and will try to re-purpose an existing instruction (i.e. one that would have to be emitted anyway) as the patchable instruction larger than a short jump.

"prologue-short-redirect" is currently only supported on x86-64.

This attribute by itself does not imply restrictions on inter-procedural optimizations. All of the semantic effects the patching may have to be separately conveyed via the linkage type.

"probe-stack" This attribute indicates that the function will trigger a guard region in the end of the stack. It ensures that accesses to the stack must be no further apart than the size of the guard region to a previous access of the stack. It takes one required string value, the name of the stack probing function that will be called.

If a function that has a "probe-stack" attribute is inlined into a function with another "probe-stack" attribute, the resulting function has the "probe-stack" attribute of the caller. If a function that has a "probe-stack" attribute is inlined into a function that has no "probe-stack" attribute at all, the resulting function has the "probe-stack" attribute of the callee.

readnone On a function, this attribute indicates that the function computes its result (or decides to unwind an exception) based strictly on its arguments, without dereferencing any pointer arguments or otherwise accessing any mutable state (e.g. memory, control registers, etc) visible to caller functions. It does not write through any pointer arguments (including `byval` arguments) and never changes any state visible to callers. This means while it cannot unwind exceptions by calling the C++ exception throwing methods (since they write to memory), there may be non-C++ mechanisms that throw exceptions without writing to LLVM visible memory.

On an argument, this attribute indicates that the function does not dereference that pointer argument, even though it may read or write the memory that the pointer points to if accessed through other pointers.

readonly On a function, this attribute indicates that the function does not write through any pointer arguments (including `byval` arguments) or otherwise modify any state (e.g. memory, control registers, etc) visible to caller functions. It may dereference pointer arguments and read state that may be set in the caller. A readonly function always returns the same value (or unwinds an exception identically) when called with the same set of arguments and global state. This means while it cannot unwind exceptions by calling the C++ exception throwing methods (since they write to memory), there may be non-C++ mechanisms that throw exceptions without writing to LLVM visible memory.

On an argument, this attribute indicates that the function does not write through this pointer argument, even though it may write to the memory that the pointer points to.

"stack-probe-size" This attribute controls the behavior of stack probes: either the "probe-stack" attribute, or ABI-required stack probes, if any. It defines the size of the guard region. It ensures that if the function may use more stack space than the size of the guard region, stack probing sequence will be emitted. It takes one required integer value, which is 4096 by default.

If a function that has a `"stack-probe-size"` attribute is inlined into a function with another `"stack-probe-size"` attribute, the resulting function has the `"stack-probe-size"` attribute that has the lower numeric value. If a function that has a `"stack-probe-size"` attribute is inlined into a function that has no `"stack-probe-size"` attribute at all, the resulting function has the `"stack-probe-size"` attribute of the callee.

writeonly On a function, this attribute indicates that the function may write to but does not read from memory.

On an argument, this attribute indicates that the function may write to but does not read through this pointer argument (even though it may read from the memory that the pointer points to).

argmemonly This attribute indicates that the only memory accesses inside function are loads and stores from objects pointed to by its pointer-typed arguments, with arbitrary offsets. Or in other words, all memory operations in the function can refer to memory only using pointers based on its function arguments. Note that `argmemonly` can be used together with `readonly` attribute in order to specify that function reads only from its arguments.

returns_twice This attribute indicates that this function can return twice. The C `setjmp` is an example of such a function. The compiler disables some optimizations (like tail calls) in the caller of these functions.

safestack This attribute indicates that [SafeStack](#) protection is enabled for this function.

If a function that has a `safestack` attribute is inlined into a function that doesn't have a `safestack` attribute or which has an `ssp`, `sspstrong` or `sspreq` attribute, then the resulting function will have a `safestack` attribute.

sanitize_address This attribute indicates that AddressSanitizer checks (dynamic address safety analysis) are enabled for this function.

sanitize_memory This attribute indicates that MemorySanitizer checks (dynamic detection of accesses to uninitialized memory) are enabled for this function.

sanitize_thread This attribute indicates that ThreadSanitizer checks (dynamic thread safety analysis) are enabled for this function.

speculatable This function attribute indicates that the function does not have any effects besides calculating its result and does not have undefined behavior. Note that `speculatable` is not enough to conclude that along any particular execution path the number of calls to this function will not be externally observable. This attribute is only valid on functions and declarations, not on individual call sites. If a function is incorrectly marked as `speculatable` and really does exhibit undefined behavior, the undefined behavior may be observed even if the call site is dead code.

ssp This attribute indicates that the function should emit a stack smashing protector. It is in the form of a “canary” — a random value placed on the stack before the local variables that's checked upon return from the function to see if it has been overwritten. A heuristic is used to determine if a function needs stack protectors or not. The heuristic used will enable protectors for functions with:

- Character arrays larger than `ssp-buffer-size` (default 8).
- Aggregates containing character arrays larger than `ssp-buffer-size`.
- Calls to `alloca()` with variable sizes or constant sizes greater than `ssp-buffer-size`.

Variables that are identified as requiring a protector will be arranged on the stack such that they are adjacent to the stack protector guard.

If a function that has an `ssp` attribute is inlined into a function that doesn't have an `ssp` attribute, then the resulting function will have an `ssp` attribute.

sspreq This attribute indicates that the function should *always* emit a stack smashing protector. This overrides the `ssp` function attribute.

Variables that are identified as requiring a protector will be arranged on the stack such that they are adjacent to the stack protector guard. The specific layout rules are:

1. Large arrays and structures containing large arrays (\geq `ssp-buffer-size`) are closest to the stack protector.
2. Small arrays and structures containing small arrays ($<$ `ssp-buffer-size`) are 2nd closest to the protector.
3. Variables that have had their address taken are 3rd closest to the protector.

If a function that has an `sspreq` attribute is inlined into a function that doesn't have an `sspreq` attribute or which has an `ssp` or `sspstrong` attribute, then the resulting function will have an `sspreq` attribute.

sspstrong This attribute indicates that the function should emit a stack smashing protector. This attribute causes a strong heuristic to be used when determining if a function needs stack protectors. The strong heuristic will enable protectors for functions with:

- Arrays of any size and type
- Aggregates containing an array of any size and type.
- Calls to `alloca()`.
- Local variables that have had their address taken.

Variables that are identified as requiring a protector will be arranged on the stack such that they are adjacent to the stack protector guard. The specific layout rules are:

1. Large arrays and structures containing large arrays (\geq `ssp-buffer-size`) are closest to the stack protector.
2. Small arrays and structures containing small arrays ($<$ `ssp-buffer-size`) are 2nd closest to the protector.
3. Variables that have had their address taken are 3rd closest to the protector.

This overrides the `ssp` function attribute.

If a function that has an `sspstrong` attribute is inlined into a function that doesn't have an `sspstrong` attribute, then the resulting function will have an `sspstrong` attribute.

"thunk" This attribute indicates that the function will delegate to some other function with a tail call. The prototype of a thunk should not be used for optimization purposes. The caller is expected to cast the thunk prototype to match the thunk target prototype.

unwindtable This attribute indicates that the ABI being targeted requires that an unwind table entry be produced for this function even if we can show that no exceptions pass by it. This is normally the case for the ELF x86-64 abi, but it can be disabled for some compilation units.

Global Attributes

Attributes may be set to communicate additional information about a global variable. Unlike *function attributes*, attributes on a global variable are grouped into a single *attribute group*.

Operand Bundles

Operand bundles are tagged sets of SSA values that can be associated with certain LLVM instructions (currently only `calls` and `invoke s`). In a way they are like metadata, but dropping them is incorrect and will change program semantics.

Syntax:

```
operand bundle set ::= '[' operand bundle (, operand bundle ) * ']'
operand bundle ::= tag '(' [ bundle operand ] (, bundle operand ) * ')'
bundle operand ::= SSA value
tag ::= string constant
```

Operand bundles are **not** part of a function’s signature, and a given function may be called from multiple places with different kinds of operand bundles. This reflects the fact that the operand bundles are conceptually a part of the `call` (or `invoke`), not the callee being dispatched to.

Operand bundles are a generic mechanism intended to support runtime-introspection-like functionality for managed languages. While the exact semantics of an operand bundle depend on the bundle tag, there are certain limitations to how much the presence of an operand bundle can influence the semantics of a program. These restrictions are described as the semantics of an “unknown” operand bundle. As long as the behavior of an operand bundle is describable within these restrictions, LLVM does not need to have special knowledge of the operand bundle to not miscompile programs containing it.

- The bundle operands for an unknown operand bundle escape in unknown ways before control is transferred to the callee or invokee.
- Calls and invokes with operand bundles have unknown read / write effect on the heap on entry and exit (even if the call target is `readnone` or `readonly`), unless they’re overridden with callsite specific attributes.
- An operand bundle at a call site cannot change the implementation of the called function. Inter-procedural optimizations work as usual as long as they take into account the first two properties.

More specific types of operand bundles are described below.

Deoptimization Operand Bundles

Deoptimization operand bundles are characterized by the “`deopt`” operand bundle tag. These operand bundles represent an alternate “safe” continuation for the call site they’re attached to, and can be used by a suitable runtime to deoptimize the compiled frame at the specified call site. There can be at most one “`deopt`” operand bundle attached to a call site. Exact details of deoptimization is out of scope for the language reference, but it usually involves rewriting a compiled frame into a set of interpreted frames.

From the compiler’s perspective, deoptimization operand bundles make the call sites they’re attached to at least `readonly`. They read through all of their pointer typed operands (even if they’re not otherwise escaped) and the entire visible heap. Deoptimization operand bundles do not capture their operands except during deoptimization, in which case control will not be returned to the compiled frame.

The inliner knows how to inline through calls that have deoptimization operand bundles. Just like inlining through a normal call site involves composing the normal and exceptional continuations, inlining through a call site with a deoptimization operand bundle needs to appropriately compose the “safe” deoptimization continuation. The inliner does this by prepending the parent’s deoptimization continuation to every deoptimization continuation in the inlined body. E.g. inlining `@f` into `@g` in the following example

```
define void @f() {
  call void @x()    ;; no deopt state
  call void @y() [ "deopt" (i32 10) ]
  call void @y() [ "deopt" (i32 10), "unknown" (i8* null) ]
  ret void
}

define void @g() {
  call void @f() [ "deopt" (i32 20) ]
  ret void
}
```

will result in

```
define void @g() {
  call void @x()  ;; still no deopt state
  call void @y() [ "deopt" (i32 20, i32 10) ]
  call void @y() [ "deopt" (i32 20, i32 10), "unknown" (i8* null) ]
  ret void
}
```

It is the frontend’s responsibility to structure or encode the deoptimization state in a way that syntactically prepending the caller’s deoptimization state to the callee’s deoptimization state is semantically equivalent to composing the caller’s deoptimization continuation after the callee’s deoptimization continuation.

Funclet Operand Bundles

Funclet operand bundles are characterized by the "funclet" operand bundle tag. These operand bundles indicate that a call site is within a particular funclet. There can be at most one "funclet" operand bundle attached to a call site and it must have exactly one bundle operand.

If any funclet EH pads have been “entered” but not “exited” (per the [description in the EH doc](#)), it is undefined behavior to execute a `call` or `invoke` which:

- does not have a "funclet" bundle and is not a `call` to a nounwind intrinsic, or
- has a "funclet" bundle whose operand is not the most-recently-entered not-yet-exited funclet EH pad.

Similarly, if no funclet EH pads have been entered-but-not-yet-exited, executing a `call` or `invoke` with a "funclet" bundle is undefined behavior.

GC Transition Operand Bundles

GC transition operand bundles are characterized by the "gc-transition" operand bundle tag. These operand bundles mark a call as a transition between a function with one GC strategy to a function with a different GC strategy. If coordinating the transition between GC strategies requires additional code generation at the call site, these bundles may contain any values that are needed by the generated code. For more details, see [GC Transitions](#).

Module-Level Inline Assembly

Modules may contain “module-level inline asm” blocks, which corresponds to the GCC “file scope inline asm” blocks. These blocks are internally concatenated by LLVM and treated as a single unit, but may be separated in the `.ll` file if desired. The syntax is very simple:

```
module asm "inline asm code goes here"
module asm "more can go here"
```

The strings can contain any character by escaping non-printable characters. The escape sequence used is simply “\xx” where “xx” is the two digit hex code for the number.

Note that the assembly string *must* be parseable by LLVM’s integrated assembler (unless it is disabled), even when emitting a `.s` file.

Data Layout

A module may specify a target specific data layout string that specifies how data is to be laid out in memory. The syntax for the data layout is simply:

```
target datalayout = "layout specification"
```

The *layout specification* consists of a list of specifications separated by the minus sign character ('-'). Each specification starts with a letter and may include other information after the letter to define some aspect of the data layout. The specifications accepted are as follows:

E Specifies that the target lays out data in big-endian form. That is, the bits with the most significance have the lowest address location.

e Specifies that the target lays out data in little-endian form. That is, the bits with the least significance have the lowest address location.

S<size> Specifies the natural alignment of the stack in bits. Alignment promotion of stack variables is limited to the natural stack alignment to avoid dynamic stack realignment. The stack alignment must be a multiple of 8-bits. If omitted, the natural stack alignment defaults to “unspecified”, which does not prevent any alignment promotions.

A<address space> Specifies the address space of objects created by ‘alloca’. Defaults to the default address space of 0.

p[n] : <size> : <abi> : <pref> This specifies the *size* of a pointer and its <abi> and <pref>ered alignments for address space n. All sizes are in bits. The address space, n, is optional, and if not specified, denotes the default address space 0. The value of n must be in the range [1,2²³).

i<size> : <abi> : <pref> This specifies the alignment for an integer type of a given bit <size>. The value of <size> must be in the range [1,2²³).

v<size> : <abi> : <pref> This specifies the alignment for a vector type of a given bit <size>.

f<size> : <abi> : <pref> This specifies the alignment for a floating point type of a given bit <size>. Only values of <size> that are supported by the target will work. 32 (float) and 64 (double) are supported on all targets; 80 or 128 (different flavors of long double) are also supported on some targets.

a : <abi> : <pref> This specifies the alignment for an object of aggregate type.

m : <mangling> If present, specifies that llvm names are mangled in the output. The options are

- e: ELF mangling: Private symbols get a .L prefix.
- m: Mips mangling: Private symbols get a \$ prefix.
- o: Mach-O mangling: Private symbols get L prefix. Other symbols get a _ prefix.
- w: Windows COFF prefix: Similar to Mach-O, but stdcall and fastcall functions also get a suffix based on the frame size.
- x: Windows x86 COFF prefix: Similar to Windows COFF, but use a _ prefix for __cdecl functions.

n<size1> : <size2> : <size3> . . . This specifies a set of native integer widths for the target CPU in bits. For example, it might contain n32 for 32-bit PowerPC, n32 : 64 for PowerPC 64, or n8 : 16 : 32 : 64 for X86-64. Elements of this set are considered to support most general arithmetic operations efficiently.

ni : <address space0> : <address space1> : <address space2> . . . This specifies pointer types with the specified address spaces as *Non-Integral Pointer Type*s. The 0 address space cannot be specified as non-integral.

On every specification that takes a <abi> : <pref>, specifying the <pref> alignment is optional. If omitted, the preceding : should be omitted too and <pref> will be equal to <abi>.

When constructing the data layout for a given target, LLVM starts with a default set of specifications which are then (possibly) overridden by the specifications in the `datalayout` keyword. The default specifications are given in this list:

- `E` - big endian
- `p:64:64:64` - 64-bit pointers with 64-bit alignment.
- `p[n]:64:64:64` - Other address spaces are assumed to be the same as the default address space.
- `S0` - natural stack alignment is unspecified
- `i1:8:8` - `i1` is 8-bit (byte) aligned
- `i8:8:8` - `i8` is 8-bit (byte) aligned
- `i16:16:16` - `i16` is 16-bit aligned
- `i32:32:32` - `i32` is 32-bit aligned
- `i64:32:64` - `i64` has ABI alignment of 32-bits but preferred alignment of 64-bits
- `f16:16:16` - half is 16-bit aligned
- `f32:32:32` - float is 32-bit aligned
- `f64:64:64` - double is 64-bit aligned
- `f128:128:128` - quad is 128-bit aligned
- `v64:64:64` - 64-bit vector is 64-bit aligned
- `v128:128:128` - 128-bit vector is 128-bit aligned
- `a:0:64` - aggregates are 64-bit aligned

When LLVM is determining the alignment for a given type, it uses the following rules:

1. If the type sought is an exact match for one of the specifications, that specification is used.
2. If no match is found, and the type sought is an integer type, then the smallest integer type that is larger than the bitwidth of the sought type is used. If none of the specifications are larger than the bitwidth then the largest integer type is used. For example, given the default specifications above, the `i7` type will use the alignment of `i8` (next largest) while both `i65` and `i256` will use the alignment of `i64` (largest specified).
3. If no match is found, and the type sought is a vector type, then the largest vector type that is smaller than the sought vector type will be used as a fall back. This happens because `<128 x double>` can be implemented in terms of `64 <2 x double>`, for example.

The function of the data layout string may not be what you expect. Notably, this is not a specification from the frontend of what alignment the code generator should use.

Instead, if specified, the target data layout is required to match what the ultimate *code generator* expects. This string is used by the mid-level optimizers to improve code, and this only works if it matches what the ultimate code generator uses. There is no way to generate IR that does not embed this target-specific detail into the IR. If you don't specify the string, the default specifications will be used to generate a Data Layout and the optimization phases will operate accordingly and introduce target specificity into the IR with respect to these default specifications.

Target Triple

A module may specify a target triple string that describes the target host. The syntax for the target triple is simply:

```
target triple = "x86_64-apple-macosx10.7.0"
```

The *target triple* string consists of a series of identifiers delimited by the minus sign character ('-'). The canonical forms are:

```
ARCHITECTURE-VENDOR-OPERATING_SYSTEM  
ARCHITECTURE-VENDOR-OPERATING_SYSTEM-ENVIRONMENT
```

This information is passed along to the backend so that it generates code for the proper architecture. It's possible to override this on the command line with the `-mtriple` command line option.

Pointer Aliasing Rules

Any memory access must be done through a pointer value associated with an address range of the memory access, otherwise the behavior is undefined. Pointer values are associated with address ranges according to the following rules:

- A pointer value is associated with the addresses associated with any value it is *based* on.
- An address of a global variable is associated with the address range of the variable's storage.
- The result value of an allocation instruction is associated with the address range of the allocated storage.
- A null pointer in the default address-space is associated with no address.
- An integer constant other than zero or a pointer value returned from a function not defined within LLVM may be associated with address ranges allocated through mechanisms other than those provided by LLVM. Such ranges shall not overlap with any ranges of addresses allocated by mechanisms provided by LLVM.

A pointer value is *based* on another pointer value according to the following rules:

- A pointer value formed from a `getelementptr` operation is *based* on the second value operand of the `getelementptr`.
- The result value of a `bitcast` is *based* on the operand of the `bitcast`.
- A pointer value formed by an `inttoptr` is *based* on all pointer values that contribute (directly or indirectly) to the computation of the pointer's value.
- The “*based on*” relationship is transitive.

Note that this definition of “*based*” is intentionally similar to the definition of “*based*” in C99, though it is slightly weaker.

LLVM IR does not associate types with memory. The result type of a `load` merely indicates the size and alignment of the memory from which to load, as well as the interpretation of the value. The first operand type of a `store` similarly only indicates the size and alignment of the store.

Consequently, type-based alias analysis, aka TBAA, aka `-fstrict-aliasing`, is not applicable to general unadorned LLVM IR. *Metadata* may be used to encode additional information which specialized optimization passes may use to implement type-based alias analysis.

Volatile Memory Accesses

Certain memory accesses, such as *load*'s, *store*'s, and *llvm.memcpy*'s may be marked `volatile`. The optimizers must not change the number of volatile operations or change their order of execution relative to other volatile operations. The optimizers *may* change the order of volatile operations relative to non-volatile operations. This is not Java's “volatile” and has no cross-thread synchronization behavior.

IR-level volatile loads and stores cannot safely be optimized into `llvm.memcpy` or `llvm.memmove` intrinsics even when those intrinsics are flagged volatile. Likewise, the backend should never split or merge target-legal volatile load/store instructions.

Rationale

Platforms may rely on volatile loads and stores of natively supported data width to be executed as single instruction. For example, in C this holds for an l-value of volatile primitive type with native hardware support, but not necessarily for aggregate types. The frontend upholds these expectations, which are intentionally unspecified in the IR. The rules above ensure that IR transformations do not violate the frontend's contract with the language.

Memory Model for Concurrent Operations

The LLVM IR does not define any way to start parallel threads of execution or to register signal handlers. Nonetheless, there are platform-specific ways to create them, and we define LLVM IR's behavior in their presence. This model is inspired by the C++0x memory model.

For a more informal introduction to this model, see the [LLVM Atomic Instructions and Concurrency Guide](#).

We define a *happens-before* partial order as the least partial order that

- Is a superset of single-thread program order, and
- When a *synchronizes-with* b, includes an edge from a to b. *Synchronizes-with* pairs are introduced by platform-specific techniques, like pthread locks, thread creation, thread joining, etc., and by atomic instructions. (See also [Atomic Memory Ordering Constraints](#)).

Note that program order does not introduce *happens-before* edges between a thread and signals executing inside that thread.

Every (defined) read operation (load instructions, memcpy, atomic loads/read-modify-writes, etc.) R reads a series of bytes written by (defined) write operations (store instructions, atomic stores/read-modify-writes, memcpy, etc.). For the purposes of this section, initialized globals are considered to have a write of the initializer which is atomic and happens before any other read or write of the memory in question. For each byte of a read R, R_{byte} may see any write to the same byte, except:

- If write_1 happens before write_2 , and write_2 happens before R_{byte} , then R_{byte} does not see write_1 .
- If R_{byte} happens before write_3 , then R_{byte} does not see write_3 .

Given that definition, R_{byte} is defined as follows:

- If R is volatile, the result is target-dependent. (Volatile is supposed to give guarantees which can support `sig_atomic_t` in C/C++, and may be used for accesses to addresses that do not behave like normal memory. It does not generally provide cross-thread synchronization.)
- Otherwise, if there is no write to the same byte that happens before R_{byte} , R_{byte} returns `undef` for that byte.
- Otherwise, if R_{byte} may see exactly one write, R_{byte} returns the value written by that write.
- Otherwise, if R is atomic, and all the writes R_{byte} may see are atomic, it chooses one of the values written. See the [Atomic Memory Ordering Constraints](#) section for additional constraints on how the choice is made.
- Otherwise R_{byte} returns `undef`.

R returns the value composed of the series of bytes it read. This implies that some bytes within the value may be `undef` **without** the entire value being `undef`. Note that this only defines the semantics of the operation; it doesn't mean that targets will emit more than one instruction to read the series of bytes.

Note that in cases where none of the atomic intrinsics are used, this model places only one restriction on IR transformations on top of what is required for single-threaded execution: introducing a store to a byte which might not otherwise be stored is not allowed in general. (Specifically, in the case where another thread might write to and read from an address, introducing a store can change a load that may see exactly one write into a load that may see multiple writes.)

Atomic Memory Ordering Constraints

Atomic instructions (*cmpxchg*, *atomicrmw*, *fence*, *atomic load*, and *atomic store*) take ordering parameters that determine which other atomic instructions on the same address they *synchronize with*. These semantics are borrowed from Java and C++0x, but are somewhat more colloquial. If these descriptions aren't precise enough, check those specs (see spec references in the *atomics guide*). *fence* instructions treat these orderings somewhat differently since they don't take an address. See that instruction's documentation for details.

For a simpler introduction to the ordering constraints, see the *LLVM Atomic Instructions and Concurrency Guide*.

unordered The set of values that can be read is governed by the happens-before partial order. A value cannot be read unless some operation wrote it. This is intended to provide a guarantee strong enough to model Java's non-volatile shared variables. This ordering cannot be specified for read-modify-write operations; it is not strong enough to make them atomic in any interesting way.

monotonic In addition to the guarantees of *unordered*, there is a single total order for modifications by *monotonic* operations on each address. All modification orders must be compatible with the happens-before order. There is no guarantee that the modification orders can be combined to a global total order for the whole program (and this often will not be possible). The read in an atomic read-modify-write operation (*cmpxchg* and *atomicrmw*) reads the value in the modification order immediately before the value it writes. If one atomic read happens before another atomic read of the same address, the later read must see the same value or a later value in the address's modification order. This disallows reordering of *monotonic* (or stronger) operations on the same address. If an address is written *monotonic*-ally by one thread, and other threads *monotonic*-ally read that address repeatedly, the other threads must eventually see the write. This corresponds to the C++0x/C1x `memory_order_relaxed`.

acquire In addition to the guarantees of *monotonic*, a *synchronizes-with* edge may be formed with a release operation. This is intended to model C++'s `memory_order_acquire`.

release In addition to the guarantees of *monotonic*, if this operation writes a value which is subsequently read by an *acquire* operation, it *synchronizes-with* that operation. (This isn't a complete description; see the C++0x definition of a release sequence.) This corresponds to the C++0x/C1x `memory_order_release`.

acq_rel (acquire+release) Acts as both an *acquire* and *release* operation on its address. This corresponds to the C++0x/C1x `memory_order_acq_rel`.

seq_cst (sequentially consistent) In addition to the guarantees of *acq_rel* (*acquire* for an operation that only reads, *release* for an operation that only writes), there is a global total order on all sequentially-consistent operations on all addresses, which is consistent with the *happens-before* partial order and with the modification orders of all the affected addresses. Each sequentially-consistent read sees the last preceding write to the same address in this global order. This corresponds to the C++0x/C1x `memory_order_seq_cst` and Java `volatile`.

If an atomic operation is marked `syncscope("singlethread")`, it only *synchronizes with* and only participates in the `seq_cst` total orderings of other operations running in the same thread (for example, in signal handlers).

If an atomic operation is marked `syncscope("<target-scope>")`, where `<target-scope>` is a target specific synchronization scope, then it is target dependent if it *synchronizes with* and participates in the `seq_cst` total orderings of other operations.

Otherwise, an atomic operation that is not marked `syncscope("singlethread")` or `syncscope("<target-scope>")` *synchronizes with* and participates in the `seq_cst` total orderings of other operations that are not marked `syncscope("singlethread")` or `syncscope("<target-scope>")`.

Fast-Math Flags

LLVM IR floating-point binary ops (*fadd*, *fsub*, *fmul*, *fdiv*, *frem*, *fcmp*) and *call* instructions have the following flags that can be set to enable otherwise unsafe floating point transformations.

nnan No NaNs - Allow optimizations to assume the arguments and result are not NaN. Such optimizations are required to retain defined behavior over NaNs, but the value of the result is undefined.

ninf No Infs - Allow optimizations to assume the arguments and result are not +/-Inf. Such optimizations are required to retain defined behavior over +/-Inf, but the value of the result is undefined.

nsz No Signed Zeros - Allow optimizations to treat the sign of a zero argument or result as insignificant.

arcp Allow Reciprocal - Allow optimizations to use the reciprocal of an argument rather than perform division.

contract Allow floating-point contraction (e.g. fusing a multiply followed by an addition into a fused multiply-and-add).

fast Fast - Allow algebraically equivalent transformations that may dramatically change results in floating point (e.g. reassociate). This flag implies all the others.

Use-list Order Directives

Use-list directives encode the in-memory order of each use-list, allowing the order to be recreated. `<order-indexes>` is a comma-separated list of indexes that are assigned to the referenced value's uses. The referenced value's use-list is immediately sorted by these indexes.

Use-list directives may appear at function scope or global scope. They are not instructions, and have no effect on the semantics of the IR. When they're at function scope, they must appear after the terminator of the final basic block.

If basic blocks have their address taken via `blockaddress()` expressions, `uselistorder_bb` can be used to reorder their use-lists from outside their function's scope.

Syntax

```
uselistorder <ty> <value>, { <order-indexes> }
uselistorder_bb @function, %block { <order-indexes> }
```

Examples

```
define void @foo(i32 %arg1, i32 %arg2) {
entry:
; ... instructions ...
bb:
; ... instructions ...

; At function scope.
uselistorder i32 %arg1, { 1, 0, 2 }
uselistorder label %bb, { 1, 0 }
}

; At global scope.
uselistorder i32* @global, { 1, 2, 0 }
uselistorder i32 7, { 1, 0 }
uselistorder i32 (i32) @bar, { 1, 0 }
uselistorder_bb @foo, %bb, { 5, 1, 3, 2, 0, 4 }
```

Source Filename

The *source filename* string is set to the original module identifier, which will be the name of the compiled source file when compiling from source through the clang front end, for example. It is then preserved through the IR and bitcode.

This is currently necessary to generate a consistent unique global identifier for local functions used in profile data, which prepends the source file name to the local function name.

The syntax for the source file name is simply:

```
source_filename = "/path/to/source.c"
```

1.1.5 Type System

The LLVM type system is one of the most important features of the intermediate representation. Being typed enables a number of optimizations to be performed on the intermediate representation directly, without having to do extra analyses on the side before the transformation. A strong type system makes it easier to read the generated code and enables novel analyses and transformations that are not feasible to perform on normal three address code representations.

Void Type

Overview

The void type does not represent any value and has no size.

Syntax

```
void
```

Function Type

Overview

The function type can be thought of as a function signature. It consists of a return type and a list of formal parameter types. The return type of a function type is a void type or first class type — except for *label* and *metadata* types.

Syntax

```
<returntype> (<parameter list>)
```

...where ‘<parameter list>’ is a comma-separated list of type specifiers. Optionally, the parameter list may include a type `...`, which indicates that the function takes a variable number of arguments. Variable argument functions can access their arguments with the *variable argument handling intrinsic* functions. ‘<returntype>’ is any type except *label* and *metadata*.

Examples

<code>i32 (i32)</code>	function taking an <code>i32</code> , returning an <code>i32</code>
<code>float (i16, i32 *) *</code>	<i>Pointer</i> to a function that takes an <code>i16</code> and a <i>pointer</i> to <code>i32</code> , returning <code>float</code> .
<code>i32 (i8*, ...)</code>	A vararg function that takes at least one <i>pointer</i> to <code>i8</code> (char in C), which returns an integer. This is the signature for <code>printf</code> in LLVM.
<code>{i32, i32} (i32)</code>	A function taking an <code>i32</code> , returning a <i>structure</i> containing two <code>i32</code> values

First Class Types

The *first class* types are perhaps the most important. Values of these types are the only ones which can be produced by instructions.

Single Value Types

These are the types that are valid in registers from CodeGen's perspective.

Integer Type

Overview

The integer type is a very simple type that simply specifies an arbitrary bit width for the integer type desired. Any bit width from 1 bit to $2^{23}-1$ (about 8 million) can be specified.

Syntax

```
iN
```

The number of bits the integer will occupy is specified by the `N` value.

Examples:

<code>i1</code>	a single-bit integer.
<code>i32</code>	a 32-bit integer.
<code>i1942652</code>	a really big integer of over 1 million bits.

Floating Point Types

Type	Description
<code>half</code>	16-bit floating point value
<code>float</code>	32-bit floating point value
<code>double</code>	64-bit floating point value
<code>fp128</code>	128-bit floating point value (112-bit mantissa)
<code>x86_fp80</code>	80-bit floating point value (X87)
<code>ppc_fp128</code>	128-bit floating point value (two 64-bits)

X86_mmx Type

Overview

The `x86_mmx` type represents a value held in an MMX register on an x86 machine. The operations allowed on it are quite limited: parameters and return values, load and store, and bitcast. User-specified MMX instructions are represented as intrinsic or asm calls with arguments and/or results of this type. There are no arrays, vectors or constants of this type.

Syntax

```
x86_mmx
```

Pointer Type

Overview

The pointer type is used to specify memory locations. Pointers are commonly used to reference objects in memory.

Pointer types may have an optional address space attribute defining the numbered address space where the pointed-to object resides. The default address space is number zero. The semantics of non-zero address spaces are target-specific.

Note that LLVM does not permit pointers to void (`void*`) nor does it permit pointers to labels (`label*`). Use `i8*` instead.

Syntax

```
<type> *
```

Examples

<code>[4 x i32]*</code>	A <i>pointer to array</i> of four <code>i32</code> values.
<code>i32 (i32*) *</code>	A <i>pointer to a function</i> that takes an <code>i32*</code> , returning an <code>i32</code> .
<code>i32 addrspace(5) *</code>	A <i>pointer</i> to an <code>i32</code> value that resides in address space #5.

Vector Type

Overview

A vector type is a simple derived type that represents a vector of elements. Vector types are used when multiple primitive data are operated in parallel using a single instruction (SIMD). A vector type requires a size (number of elements) and an underlying primitive data type. Vector types are considered *first class*.

Syntax

```
< <# elements> x <elementtype> >
```

The number of elements is a constant integer value larger than 0; `elementtype` may be any integer, floating point or pointer type. Vectors of size zero are not allowed.

Examples

<code><4 x i32></code>	Vector of 4 32-bit integer values.
<code><8 x float></code>	Vector of 8 32-bit floating-point values.
<code><2 x i64></code>	Vector of 2 64-bit integer values.
<code><4 x i64*></code>	Vector of 4 pointers to 64-bit integer values.

Label Type

Overview

The label type represents code labels.

Syntax

```
label
```


Token Type

Overview

The token type is used when a value is associated with an instruction but all uses of the value must not attempt to introspect or obscure it. As such, it is not appropriate to have a *phi* or *select* of type token.

Syntax

```
token
```

Metadata Type

Overview

The metadata type represents embedded metadata. No derived types may be created from metadata except for *function* arguments.

Syntax

```
metadata
```

Aggregate Types

Aggregate Types are a subset of derived types that can contain multiple member types. *Arrays* and *structs* are aggregate types. *Vectors* are not considered to be aggregate types.

Array Type

Overview

The array type is a very simple derived type that arranges elements sequentially in memory. The array type requires a size (number of elements) and an underlying data type.

Syntax

```
[<# elements> x <elementtype>]
```

The number of elements is a constant integer value; *elementtype* may be any type with a size.

Examples

[40 x i32]	Array of 40 32-bit integer values.
[41 x i32]	Array of 41 32-bit integer values.
[4 x i8]	Array of 4 8-bit integer values.

Here are some examples of multidimensional arrays:

[3 x [4 x i32]]	3x4 array of 32-bit integer values.
[12 x [10 x float]]	12x10 array of single precision floating point values.
[2 x [3 x [4 x i16]]]	2x3x4 array of 16-bit integer values.

There is no restriction on indexing beyond the end of the array implied by a static type (though there are restrictions on indexing beyond the bounds of an allocated object in some cases). This means that single-dimension ‘variable sized array’ addressing can be implemented in LLVM with a zero length array type. An implementation of ‘pascal style arrays’ in LLVM could use the type “{ i32, [0 x float]}”, for example.

Structure Type

Overview

The structure type is used to represent a collection of data members together in memory. The elements of a structure may be any type that has a size.

Structures in memory are accessed using ‘load’ and ‘store’ by getting a pointer to a field with the ‘getelementptr’ instruction. Structures in registers are accessed using the ‘extractvalue’ and ‘insertvalue’ instructions.

Structures may optionally be “packed” structures, which indicate that the alignment of the struct is one byte, and that there is no padding between the elements. In non-packed structs, padding between field types is inserted as defined by the DataLayout string in the module, which is required to match what the underlying code generator expects.

Structures can either be “literal” or “identified”. A literal structure is defined inline with other types (e.g. {i32, i32}*) whereas identified types are always defined at the top level with a name. Literal types are uniqued by their contents and can never be recursive or opaque since there is no way to write one. Identified types can be recursive, can be opaqued, and are never uniqued.

Syntax

```
%T1 = type { <type list> }      ; Identified normal struct type
%T2 = type <{ <type list> }>    ; Identified packed struct type
```

Examples

{ i32, i32, i32 }	A triple of three i32 values
{ float, i32 (i32) * }	A pair, where the first element is a float and the second element is a <i>pointer</i> to a <i>function</i> that takes an i32, returning an i32.
<{ i8, i32 }>	A packed struct known to be 5 bytes in size.

Opaque Structure Types

Overview

Opaque structure types are used to represent named structure types that do not have a body specified. This corresponds (for example) to the C notion of a forward declared structure.

Syntax

```
%X = type opaque
%52 = type opaque
```

Examples

opaque	An opaque type.
--------	-----------------

1.1.6 Constants

LLVM has several different basic types of constants. This section describes them all and their syntax.

Simple Constants

Boolean constants The two strings ‘true’ and ‘false’ are both valid constants of the `i1` type.

Integer constants Standard integers (such as ‘4’) are constants of the *integer* type. Negative numbers may be used with integer types.

Floating point constants Floating point constants use standard decimal notation (e.g. 123.421), exponential notation (e.g. 1.23421e+2), or a more precise hexadecimal notation (see below). The assembler requires the exact decimal value of a floating-point constant. For example, the assembler accepts 1.25 but rejects 1.3 because 1.3 is a repeating decimal in binary. Floating point constants must have a *floating point* type.

Null pointer constants The identifier ‘null’ is recognized as a null pointer constant and must be of *pointer type*.

Token constants The identifier ‘none’ is recognized as an empty token constant and must be of *token type*.

The one non-intuitive notation for constants is the hexadecimal form of floating point constants. For example, the form ‘double 0x432ff973cafa8000’ is equivalent to (but harder to read than) ‘double 4.5e+15’. The only time hexadecimal floating point constants are required (and the only time that they are generated by the disassembler) is when a floating point constant must be emitted but it cannot be represented as a decimal floating point number in a reasonable number of digits. For example, NaN’s, infinities, and other special values are represented in their IEEE hexadecimal format so that assembly and disassembly do not cause any bits to change in the constants.

When using the hexadecimal form, constants of types half, float, and double are represented using the 16-digit form shown above (which matches the IEEE754 representation for double); half and float values must, however, be exactly representable as IEEE 754 half and single precision, respectively. Hexadecimal format is always used for long double, and there are three forms of long double. The 80-bit format used by x86 is represented as 0xK followed by 20 hexadecimal digits. The 128-bit format used by PowerPC (two adjacent doubles) is represented by 0xM followed by 32 hexadecimal digits. The IEEE 128-bit format is represented by 0xL followed by 32 hexadecimal digits. Long doubles will only work if they match the long double format on your target. The IEEE 16-bit format (half precision) is represented by 0xH followed by 4 hexadecimal digits. All hexadecimal formats are big-endian (sign bit at the left).

There are no constants of type `x86_mmx`.

Complex Constants

Complex constants are a (potentially recursive) combination of simple constants and smaller complex constants.

Structure constants Structure constants are represented with notation similar to structure type definitions (a comma separated list of elements, surrounded by braces ({})). For example: “{ i32 4, float 17.0, i32* @G }”, where “@G” is declared as “@G = external global i32”. Structure constants must have *structure type*, and the number and types of elements must match those specified by the type.

Array constants Array constants are represented with notation similar to array type definitions (a comma separated list of elements, surrounded by square brackets ([])). For example: “[i32 42, i32 11, i32 74]”. Array constants must have *array type*, and the number and types of elements must match those specified by the type. As a special case, character array constants may also be represented as a double-quoted string using the `c` prefix. For example: “c"Hello World\0A\00”.

Vector constants Vector constants are represented with notation similar to vector type definitions (a comma separated list of elements, surrounded by less-than/greater-than’s (<>)). For example: “< i32 42, i32 11, i32 74, i32 100 >”. Vector constants must have *vector type*, and the number and types of elements must match those specified by the type.

Zero initialization The string ‘zeroinitializer’ can be used to zero initialize a value to zero of *any* type, including scalar and *aggregate* types. This is often used to avoid having to print large zero initializers (e.g. for large arrays) and is always exactly equivalent to using explicit zero initializers.

Metadata node A metadata node is a constant tuple without types. For example: “!{!0, !{!2, !0}, !"test"}”. Metadata can reference constant values, for example: “!{!0, i32 0, i8* @global, i64 (i64)* @function, !"str"}”. Unlike other typed constants that are meant to be interpreted as part of the instruction stream, metadata is a place to attach additional information such as debug info.

Global Variable and Function Addresses

The addresses of *global variables* and *functions* are always implicitly valid (link-time) constants. These constants are explicitly referenced when the *identifier for the global* is used and always have *pointer* type. For example, the following is a legal LLVM file:

```
@X = global i32 17
@Y = global i32 42
@Z = global [2 x i32*] [ i32* @X, i32* @Y ]
```

Undefined Values

The string ‘undef’ can be used anywhere a constant is expected, and indicates that the user of the value may receive an unspecified bit-pattern. Undefined values may be of any type (other than ‘label’ or ‘void’) and be used anywhere a constant is permitted.

Undefined values are useful because they indicate to the compiler that the program is well defined no matter what value is used. This gives the compiler more freedom to optimize. Here are some examples of (potentially surprising) transformations that are valid (in pseudo IR):

```
%A = add %X, undef
%B = sub %X, undef
%C = xor %X, undef
Safe:
%A = undef
%B = undef
%C = undef
```

This is safe because all of the output bits are affected by the undef bits. Any output bit can have a zero or one depending on the input bits.

```
%A = or %X, undef
%B = and %X, undef
Safe:
%A = -1
%B = 0
Safe:
%A = %X ;; By choosing undef as 0
%B = %X ;; By choosing undef as -1
Unsafe:
%A = undef
%B = undef
```

These logical operations have bits that are not always affected by the input. For example, if %X has a zero bit, then the output of the ‘and’ operation will always be a zero for that bit, no matter what the corresponding bit from the ‘undef’ is. As such, it is unsafe to optimize or assume that the result of the ‘and’ is ‘undef’. However, it is safe to

assume that all bits of the ‘undef’ could be 0, and optimize the ‘and’ to 0. Likewise, it is safe to assume that all the bits of the ‘undef’ operand to the ‘or’ could be set, allowing the ‘or’ to be folded to -1.

```
%A = select undef, %X, %Y
%B = select undef, 42, %Y
%C = select %X, %Y, undef
Safe:
  %A = %X      (or %Y)
  %B = 42      (or %Y)
  %C = %Y
Unsafe:
  %A = undef
  %B = undef
  %C = undef
```

This set of examples shows that undefined ‘select’ (and conditional branch) conditions can go *either way*, but they have to come from one of the two operands. In the %A example, if %X and %Y were both known to have a clear low bit, then %A would have to have a cleared low bit. However, in the %C example, the optimizer is allowed to assume that the ‘undef’ operand could be the same as %Y, allowing the whole ‘select’ to be eliminated.

```
%A = xor undef, undef

%B = undef
%C = xor %B, %B

%D = undef
%E = icmp slt %D, 4
%F = icmp gte %D, 4

Safe:
  %A = undef
  %B = undef
  %C = undef
  %D = undef
  %E = undef
  %F = undef
```

This example points out that two ‘undef’ operands are not necessarily the same. This can be surprising to people (and also matches C semantics) where they assume that “X^X” is always zero, even if X is undefined. This isn’t true for a number of reasons, but the short answer is that an ‘undef’ “variable” can arbitrarily change its value over its “live range”. This is true because the variable doesn’t actually *have a live range*. Instead, the value is logically read from arbitrary registers that happen to be around when needed, so the value is not necessarily consistent over time. In fact, %A and %C need to have the same semantics or the core LLVM “replace all uses with” concept would not hold.

```
%A = fdiv undef, %X
%B = fdiv %X, undef
Safe:
  %A = undef
b: unreachable
```

These examples show the crucial difference between an *undefined value* and *undefined behavior*. An undefined value (like ‘undef’) is allowed to have an arbitrary bit-pattern. This means that the %A operation can be constant folded to ‘undef’, because the ‘undef’ could be an SNaN, and fdiv is not (currently) defined on SNaN’s. However, in the second example, we can make a more aggressive assumption: because the undef is allowed to be an arbitrary value, we are allowed to assume that it could be zero. Since a divide by zero has *undefined behavior*, we are allowed to assume that the operation does not execute at all. This allows us to delete the divide and all code after it. Because the undefined operation “can’t happen”, the optimizer can assume that it occurs in dead code.

```

a: store undef -> %X
b: store %X -> undef
Safe:
a: <deleted>
b: unreachable

```

These examples reiterate the `fdiv` example: a store *of* an undefined value can be assumed to not have any effect; we can assume that the value is overwritten with bits that happen to match what was already there. However, a store *to* an undefined location could clobber arbitrary memory, therefore, it has undefined behavior.

Poison Values

Poison values are similar to *undef values*, however they also represent the fact that an instruction or constant expression that cannot evoke side effects has nevertheless detected a condition that results in undefined behavior.

There is currently no way of representing a poison value in the IR; they only exist when produced by operations such as *add* with the `nsw` flag.

Poison value behavior is defined in terms of value *dependence*:

- Values other than *phi* nodes depend on their operands.
- *Phi* nodes depend on the operand corresponding to their dynamic predecessor basic block.
- Function arguments depend on the corresponding actual argument values in the dynamic callers of their functions.
- *Call* instructions depend on the *ret* instructions that dynamically transfer control back to them.
- *Invoke* instructions depend on the *ret*, *resume*, or exception-throwing call instructions that dynamically transfer control back to them.
- Non-volatile loads and stores depend on the most recent stores to all of the referenced memory addresses, following the order in the IR (including loads and stores implied by intrinsics such as *@llvm.memcpy*.)
- An instruction with externally visible side effects depends on the most recent preceding instruction with externally visible side effects, following the order in the IR. (This includes *volatile operations*.)
- An instruction *control-depends* on a *terminator instruction* if the terminator instruction has multiple successors and the instruction is always executed when control transfers to one of the successors, and may not be executed when control is transferred to another.
- Additionally, an instruction also *control-depends* on a terminator instruction if the set of instructions it otherwise depends on would be different if the terminator had transferred control to a different successor.
- Dependence is transitive.

Poison values have the same behavior as *undef values*, with the additional effect that any instruction that has a *dependence* on a poison value has undefined behavior.

Here are some examples:

```

entry:
  %poison = sub nuw i32 0, 1           ; Results in a poison value.
  %still_poison = and i32 %poison, 0   ; 0, but also poison.
  %poison_yet_again = getelementptr i32, i32* @h, i32 %still_poison
  store i32 0, i32* %poison_yet_again ; memory at @h[0] is poisoned

  store i32 %poison, i32* @g           ; Poison value stored to memory.
  %poison2 = load i32, i32* @g         ; Poison value loaded back from memory.

```

```

store volatile i32 %poison, i32* @g ; External observation; undefined behavior.

%narrowaddr = bitcast i32* @g to i16*
%wideaddr = bitcast i32* @g to i64*
%poison3 = load i16, i16* %narrowaddr ; Returns a poison value.
%poison4 = load i64, i64* %wideaddr ; Returns a poison value.

%cmp = icmp slt i32 %poison, 0 ; Returns a poison value.
br i1 %cmp, label %true, label %end ; Branch to either destination.

true:
store volatile i32 0, i32* @g ; This is control-dependent on %cmp, so
                               ; it has undefined behavior.

br label %end

end:
%p = phi i32 [ 0, %entry ], [ 1, %true ]
                               ; Both edges into this PHI are
                               ; control-dependent on %cmp, so this
                               ; always results in a poison value.

store volatile i32 0, i32* @g ; This would depend on the store in %true
                               ; if %cmp is true, or the store in %entry
                               ; otherwise, so this is undefined behavior.

br i1 %cmp, label %second_true, label %second_end
                               ; The same branch again, but this time the
                               ; true block doesn't have side effects.

second_true:
; No side effects!
ret void

second_end:
store volatile i32 0, i32* @g ; This time, the instruction always depends
                               ; on the store in %end. Also, it is
                               ; control-equivalent to %end, so this is
                               ; well-defined (ignoring earlier undefined
                               ; behavior in this example).

```

Addresses of Basic Blocks

blockaddress(@function, %block)

The ‘blockaddress’ constant computes the address of the specified basic block in the specified function, and always has an `i8*` type. Taking the address of the entry block is illegal.

This value only has defined behavior when used as an operand to the ‘*indirectbr*’ instruction, or for comparisons against null. Pointer equality tests between labels addresses results in undefined behavior — though, again, comparison against null is ok, and no label is equal to the null pointer. This may be passed around as an opaque pointer sized value as long as the bits are not inspected. This allows `ptrtoint` and arithmetic to be performed on these values so long as the original value is reconstituted before the *indirectbr* instruction.

Finally, some targets may provide defined semantics when using the value as the operand to an inline assembly, but that is target specific.

Constant Expressions

Constant expressions are used to allow expressions involving other constants to be used as constants. Constant expressions may be of any *first class* type and may involve any LLVM operation that does not have side effects (e.g. load and call are not supported). The following is the syntax for constant expressions:

trunc (CST to TYPE) Truncate a constant to another type. The bit size of CST must be larger than the bit size of TYPE. Both types must be integers.

zext (CST to TYPE) Zero extend a constant to another type. The bit size of CST must be smaller than the bit size of TYPE. Both types must be integers.

sext (CST to TYPE) Sign extend a constant to another type. The bit size of CST must be smaller than the bit size of TYPE. Both types must be integers.

fptrunc (CST to TYPE) Truncate a floating point constant to another floating point type. The size of CST must be larger than the size of TYPE. Both types must be floating point.

fpext (CST to TYPE) Floating point extend a constant to another type. The size of CST must be smaller or equal to the size of TYPE. Both types must be floating point.

fptoui (CST to TYPE) Convert a floating point constant to the corresponding unsigned integer constant. TYPE must be a scalar or vector integer type. CST must be of scalar or vector floating point type. Both CST and TYPE must be scalars, or vectors of the same number of elements. If the value won't fit in the integer type, the results are undefined.

fptosi (CST to TYPE) Convert a floating point constant to the corresponding signed integer constant. TYPE must be a scalar or vector integer type. CST must be of scalar or vector floating point type. Both CST and TYPE must be scalars, or vectors of the same number of elements. If the value won't fit in the integer type, the results are undefined.

uitofp (CST to TYPE) Convert an unsigned integer constant to the corresponding floating point constant. TYPE must be a scalar or vector floating point type. CST must be of scalar or vector integer type. Both CST and TYPE must be scalars, or vectors of the same number of elements. If the value won't fit in the floating point type, the results are undefined.

sitofp (CST to TYPE) Convert a signed integer constant to the corresponding floating point constant. TYPE must be a scalar or vector floating point type. CST must be of scalar or vector integer type. Both CST and TYPE must be scalars, or vectors of the same number of elements. If the value won't fit in the floating point type, the results are undefined.

ptrtoint (CST to TYPE) Convert a pointer typed constant to the corresponding integer constant. TYPE must be an integer type. CST must be of pointer type. The CST value is zero extended, truncated, or unchanged to make it fit in TYPE.

inttoptr (CST to TYPE) Convert an integer constant to a pointer constant. TYPE must be a pointer type. CST must be of integer type. The CST value is zero extended, truncated, or unchanged to make it fit in a pointer size. This one is *really* dangerous!

bitcast (CST to TYPE) Convert a constant, CST, to another TYPE. The constraints of the operands are the same as those for the *bitcast instruction*.

addrspacecast (CST to TYPE) Convert a constant pointer or constant vector of pointer, CST, to another TYPE in a different address space. The constraints of the operands are the same as those for the *addrspacecast instruction*.

getelementptr (TY, CSTPTR, IDX0, IDX1, ...), getelementptr inbounds (TY, CSTPTR, IDX0, IDX1, ...) Perform the *getelementptr operation* on constants. As with the *getelementptr* instruction, the index list may have one or more indexes, which are required to make sense for the type of "pointer to TY".

select (COND, VAL1, VAL2) Perform the *select operation* on constants.

icmp **COND** (**VAL1**, **VAL2**) Performs the *icmp operation* on constants.

fcmp **COND** (**VAL1**, **VAL2**) Performs the *fcmp operation* on constants.

extractelement (**VAL**, **IDX**) Perform the *extractelement operation* on constants.

insertelement (**VAL**, **ELT**, **IDX**) Perform the *insertelement operation* on constants.

shufflevector (**VEC1**, **VEC2**, **IDXMASK**) Perform the *shufflevector operation* on constants.

extractvalue (**VAL**, **IDX0**, **IDX1**, ...) Perform the *extractvalue operation* on constants. The index list is interpreted in a similar manner as indices in a ‘*getelementptr*’ operation. At least one index value must be specified.

insertvalue (**VAL**, **ELT**, **IDX0**, **IDX1**, ...) Perform the *insertvalue operation* on constants. The index list is interpreted in a similar manner as indices in a ‘*getelementptr*’ operation. At least one index value must be specified.

OPCODE (**LHS**, **RHS**) Perform the specified operation of the LHS and RHS constants. OPCODE may be any of the *binary* or *bitwise binary* operations. The constraints on operands are the same as those for the corresponding instruction (e.g. no bitwise operations on floating point values are allowed).

1.1.7 Other Values

Inline Assembler Expressions

LLVM supports inline assembler expressions (as opposed to *Module-Level Inline Assembly*) through the use of a special value. This value represents the inline assembler as a template string (containing the instructions to emit), a list of operand constraints (stored as a string), a flag that indicates whether or not the inline asm expression has side effects, and a flag indicating whether the function containing the asm needs to align its stack conservatively.

The template string supports argument substitution of the operands using “\$” followed by a number, to indicate substitution of the given register/memory location, as specified by the constraint string. “\${NUM:MODIFIER}” may also be used, where MODIFIER is a target-specific annotation for how to print the operand (See *Asm template argument modifiers*).

A literal “\$” may be included by using “\$\$” in the template. To include other special characters into the output, the usual “\XX” escapes may be used, just as in other strings. Note that after template substitution, the resulting assembly string is parsed by LLVM’s integrated assembler unless it is disabled – even when emitting a .s file – and thus must contain assembly syntax known to LLVM.

LLVM also supports a few more substitutions useful for writing inline assembly:

- `${:uid}`: Expands to a decimal integer unique to this inline assembly blob. This substitution is useful when declaring a local label. Many standard compiler optimizations, such as inlining, may duplicate an inline asm blob. Adding a blob-unique identifier ensures that the two labels will not conflict during assembly. This is used to implement GCC’s *%= special format string*.
- `${:comment}`: Expands to the comment character of the current target’s assembly dialect. This is usually #, but many targets use other strings, such as ;, //, or !.
- `${:private}`: Expands to the assembler private label prefix. Labels with this prefix will not appear in the symbol table of the assembled object. Typically the prefix is L, but targets may use other strings. .L is relatively popular.

LLVM’s support for inline asm is modeled closely on the requirements of Clang’s GCC-compatible inline-asm support. Thus, the feature-set and the constraint and modifier codes listed here are similar or identical to those in GCC’s inline asm support. However, to be clear, the syntax of the template and constraint strings described here is *not* the same as the syntax accepted by GCC and Clang, and, while most constraint letters are passed through as-is by Clang, some get translated to other codes when converting from the C source to the LLVM assembly.

An example inline assembler expression is:

```
i32 (i32) asm "bswap $0", "=r,r"
```

Inline assembler expressions may **only** be used as the callee operand of a *call* or an *invoke* instruction. Thus, typically we have:

```
%X = call i32 @asm "bswap $0", "=r,r" (i32 %Y)
```

Inline asms with side effects not visible in the constraint list must be marked as having side effects. This is done through the use of the ‘sideeffect’ keyword, like so:

```
call void @asm sideeffect "eieio", ""()
```

In some cases inline asms will contain code that will not work unless the stack is aligned in some way, such as calls or SSE instructions on x86, yet will not contain code that does that alignment within the asm. The compiler should make conservative assumptions about what the asm might contain and should generate its usual stack alignment code in the prologue if the ‘alignstack’ keyword is present:

```
call void @asm alignstack "eieio", ""()
```

Inline asms also support using non-standard assembly dialects. The assumed dialect is ATT. When the ‘inteldialect’ keyword is present, the inline asm is using the Intel dialect. Currently, ATT and Intel are the only supported dialects. An example is:

```
call void @asm inteldialect "eieio", ""()
```

If multiple keywords appear the ‘sideeffect’ keyword must come first, the ‘alignstack’ keyword second and the ‘inteldialect’ keyword last.

Inline Asm Constraint String

The constraint list is a comma-separated string, each element containing one or more constraint codes.

For each element in the constraint list an appropriate register or memory operand will be chosen, and it will be made available to assembly template string expansion as \$0 for the first constraint in the list, \$1 for the second, etc.

There are three different types of constraints, which are distinguished by a prefix symbol in front of the constraint code: Output, Input, and Clobber. The constraints must always be given in that order: outputs first, then inputs, then clobbers. They cannot be intermingled.

There are also three different categories of constraint codes:

- Register constraint. This is either a register class, or a fixed physical register. This kind of constraint will allocate a register, and if necessary, bitcast the argument or result to the appropriate type.
- Memory constraint. This kind of constraint is for use with an instruction taking a memory operand. Different constraints allow for different addressing modes used by the target.
- Immediate value constraint. This kind of constraint is for an integer or other immediate value which can be rendered directly into an instruction. The various target-specific constraints allow the selection of a value in the proper range for the instruction you wish to use it with.

Output constraints

Output constraints are specified by an “=” prefix (e.g. “=r”). This indicates that the assembly will write to this operand, and the operand will then be made available as a return value of the asm expression. Output constraints do

not consume an argument from the call instruction. (Except, see below about indirect outputs).

Normally, it is expected that no output locations are written to by the assembly expression until *all* of the inputs have been read. As such, LLVM may assign the same register to an output and an input. If this is not safe (e.g. if the assembly contains two instructions, where the first writes to one output, and the second reads an input and writes to a second output), then the “&” modifier must be used (e.g. “=&r”) to specify that the output is an “early-clobber” output. Marking an output as “early-clobber” ensures that LLVM will not use the same register for any inputs (other than an input tied to this output).

Input constraints

Input constraints do not have a prefix – just the constraint codes. Each input constraint will consume one argument from the call instruction. It is not permitted for the asm to write to any input register or memory location (unless that input is tied to an output). Note also that multiple inputs may all be assigned to the same register, if LLVM can determine that they necessarily all contain the same value.

Instead of providing a Constraint Code, input constraints may also “tie” themselves to an output constraint, by providing an integer as the constraint string. Tied inputs still consume an argument from the call instruction, and take up a position in the asm template numbering as is usual – they will simply be constrained to always use the same register as the output they’ve been tied to. For example, a constraint string of “=r, 0” says to assign a register for output, and use that register as an input as well (it being the 0’t constraint).

It is permitted to tie an input to an “early-clobber” output. In that case, no *other* input may share the same register as the input tied to the early-clobber (even when the other input has the same value).

You may only tie an input to an output which has a register constraint, not a memory constraint. Only a single input may be tied to an output.

There is also an “interesting” feature which deserves a bit of explanation: if a register class constraint allocates a register which is too small for the value type operand provided as input, the input value will be split into multiple registers, and all of them passed to the inline asm.

However, this feature is often not as useful as you might think.

Firstly, the registers are *not* guaranteed to be consecutive. So, on those architectures that have instructions which operate on multiple consecutive instructions, this is not an appropriate way to support them. (e.g. the 32-bit SparcV8 has a 64-bit load, which instruction takes a single 32-bit register. The hardware then loads into both the named register, and the next register. This feature of inline asm would not be useful to support that.)

A few of the targets provide a template string modifier allowing explicit access to the second register of a two-register operand (e.g. MIPS L, M, and D). On such an architecture, you can actually access the second allocated register (yet, still, not any subsequent ones). But, in that case, you’re still probably better off simply splitting the value into two separate operands, for clarity. (e.g. see the description of the A constraint on X86, which, despite existing only for use with this feature, is not really a good idea to use)

Indirect inputs and outputs

Indirect output or input constraints can be specified by the “*” modifier (which goes after the “=” in case of an output). This indicates that the asm will write to or read from the contents of an *address* provided as an input argument. (Note that in this way, indirect outputs act more like an *input* than an output: just like an input, they consume an argument of the call expression, rather than producing a return value. An indirect output constraint is an “output” only in that the asm is expected to write to the contents of the input memory location, instead of just read from it).

This is most typically used for memory constraint, e.g. “=*m”, to pass the address of a variable as a value.

It is also possible to use an indirect *register* constraint, but only on output (e.g. “=*r”). This will cause LLVM to allocate a register for an output value normally, and then, separately emit a store to the address provided as input, after

the provided inline asm. (It's not clear what value this functionality provides, compared to writing the store explicitly after the asm statement, and it can only produce worse code, since it bypasses many optimization passes. I would recommend not using it.)

Clobber constraints

A clobber constraint is indicated by a “~” prefix. A clobber does not consume an input operand, nor generate an output. Clobbers cannot use any of the general constraint code letters – they may use only explicit register constraints, e.g. “~{eax}”. The one exception is that a clobber string of “~{memory}” indicates that the assembly writes to arbitrary undeclared memory locations – not only the memory pointed to by a declared indirect output.

Note that clobbering named registers that are also present in output constraints is not legal.

Constraint Codes

After a potential prefix comes constraint code, or codes.

A Constraint Code is either a single letter (e.g. “r”), a “^” character followed by two letters (e.g. “^wC”), or “{” register-name “}” (e.g. “{eax}”).

The one and two letter constraint codes are typically chosen to be the same as GCC's constraint codes.

A single constraint may include one or more than constraint code in it, leaving it up to LLVM to choose which one to use. This is included mainly for compatibility with the translation of GCC inline asm coming from clang.

There are two ways to specify alternatives, and either or both may be used in an inline asm constraint list:

1. Append the codes to each other, making a constraint code set. E.g. “im” or “{eax}m”. This means “choose any of the options in the set”. The choice of constraint is made independently for each constraint in the constraint list.
2. Use “|” between constraint code sets, creating alternatives. Every constraint in the constraint list must have the same number of alternative sets. With this syntax, the same alternative in *all* of the items in the constraint list will be chosen together.

Putting those together, you might have a two operand constraint string like “rm|r,ri|rm”. This indicates that if operand 0 is r or m, then operand 1 may be one of r or i. If operand 0 is r, then operand 1 may be one of r or m. But, operand 0 and 1 cannot both be of type m.

However, the use of either of the alternatives features is *NOT* recommended, as LLVM is not able to make an intelligent choice about which one to use. (At the point it currently needs to choose, not enough information is available to do so in a smart way.) Thus, it simply tries to make a choice that's most likely to compile, not one that will be optimal performance. (e.g., given “rm”, it'll always choose to use memory, not registers). And, if given multiple registers, or multiple register classes, it will simply choose the first one. (In fact, it doesn't currently even ensure explicitly specified physical registers are unique, so specifying multiple physical registers as alternatives, like {r11}{r12}, {r11}{r12}, will assign r11 to both operands, not at all what was intended.)

Supported Constraint Code List

The constraint codes are, in general, expected to behave the same way they do in GCC. LLVM's support is often implemented on an ‘as-needed’ basis, to support C inline asm code which was supported by GCC. A mismatch in behavior between LLVM and GCC likely indicates a bug in LLVM.

Some constraint codes are typically supported by all targets:

- r: A register in the target's general purpose register class.

- *m*: A memory address operand. It is target-specific what addressing modes are supported, typical examples are register, or register + register offset, or register + immediate offset (of some target-specific size).
- *i*: An integer constant (of target-specific width). Allows either a simple immediate, or a relocatable value.
- *n*: An integer constant – *not* including relocatable values.
- *s*: An integer constant, but allowing *only* relocatable values.
- *X*: Allows an operand of any kind, no constraint whatsoever. Typically useful to pass a label for an asm branch or call.
- {*register-name*}: Requires exactly the named physical register.

Other constraints are target-specific:

AArch64:

- *z*: An immediate integer 0. Outputs WZR or XZR, as appropriate.
- *I*: An immediate integer valid for an ADD or SUB instruction, i.e. 0 to 4095 with optional shift by 12.
- *J*: An immediate integer that, when negated, is valid for an ADD or SUB instruction, i.e. -1 to -4095 with optional left shift by 12.
- *K*: An immediate integer that is valid for the ‘bitmask immediate 32’ of a logical instruction like AND, EOR, or ORR with a 32-bit register.
- *L*: An immediate integer that is valid for the ‘bitmask immediate 64’ of a logical instruction like AND, EOR, or ORR with a 64-bit register.
- *M*: An immediate integer for use with the MOV assembly alias on a 32-bit register. This is a superset of *K*: in addition to the bitmask immediate, also allows immediate integers which can be loaded with a single MOVZ or MOVL instruction.
- *N*: An immediate integer for use with the MOV assembly alias on a 64-bit register. This is a superset of *L*.
- *Q*: Memory address operand must be in a single register (no offsets). (However, LLVM currently does this for the *m* constraint as well.)
- *r*: A 32 or 64-bit integer register (W* or X*).
- *w*: A 32, 64, or 128-bit floating-point/SIMD register.
- *x*: A lower 128-bit floating-point/SIMD register (V0 to V15).

AMDGPU:

- *r*: A 32 or 64-bit integer register.
- [0–9] *v*: The 32-bit VGPR register, number 0-9.
- [0–9] *s*: The 32-bit SGPR register, number 0-9.

All ARM modes:

- *Q*, *Um*, *Un*, *Uq*, *Us*, *Ut*, *Uv*, *Uy*: Memory address operand. Treated the same as operand *m*, at the moment.

ARM and ARM’s Thumb2 mode:

- *j*: An immediate integer between 0 and 65535 (valid for MOVW)
- *I*: An immediate integer valid for a data-processing instruction.
- *J*: An immediate integer between -4095 and 4095.
- *K*: An immediate integer whose bitwise inverse is valid for a data-processing instruction. (Can be used with template modifier “B” to print the inverted value).

- **L**: An immediate integer whose negation is valid for a data-processing instruction. (Can be used with template modifier “n” to print the negated value).
- **M**: A power of two or a integer between 0 and 32.
- **N**: Invalid immediate constraint.
- **O**: Invalid immediate constraint.
- **r**: A general-purpose 32-bit integer register (`r0-r15`).
- **l**: In Thumb2 mode, low 32-bit GPR registers (`r0-r7`). In ARM mode, same as **r**.
- **h**: In Thumb2 mode, a high 32-bit GPR register (`r8-r15`). In ARM mode, invalid.
- **w**: A 32, 64, or 128-bit floating-point/SIMD register: `s0-s31`, `d0-d31`, or `q0-q15`.
- **x**: A 32, 64, or 128-bit floating-point/SIMD register: `s0-s15`, `d0-d7`, or `q0-q3`.
- **t**: A floating-point/SIMD register, only supports 32-bit values: `s0-s31`.

ARM’s Thumb1 mode:

- **I**: An immediate integer between 0 and 255.
- **J**: An immediate integer between -255 and -1.
- **K**: An immediate integer between 0 and 255, with optional left-shift by some amount.
- **L**: An immediate integer between -7 and 7.
- **M**: An immediate integer which is a multiple of 4 between 0 and 1020.
- **N**: An immediate integer between 0 and 31.
- **O**: An immediate integer which is a multiple of 4 between -508 and 508.
- **r**: A low 32-bit GPR register (`r0-r7`).
- **l**: A low 32-bit GPR register (`r0-r7`).
- **h**: A high GPR register (`r0-r7`).
- **w**: A 32, 64, or 128-bit floating-point/SIMD register: `s0-s31`, `d0-d31`, or `q0-q15`.
- **x**: A 32, 64, or 128-bit floating-point/SIMD register: `s0-s15`, `d0-d7`, or `q0-q3`.
- **t**: A floating-point/SIMD register, only supports 32-bit values: `s0-s31`.

Hexagon:

- **o, v**: A memory address operand, treated the same as constraint **m**, at the moment.
- **r**: A 32 or 64-bit register.

MSP430:

- **r**: An 8 or 16-bit register.

MIPS:

- **I**: An immediate signed 16-bit integer.
- **J**: An immediate integer zero.
- **K**: An immediate unsigned 16-bit integer.
- **L**: An immediate 32-bit integer, where the lower 16 bits are 0.
- **N**: An immediate integer between -65535 and -1.

- `O`: An immediate signed 15-bit integer.
- `P`: An immediate integer between 1 and 65535.
- `m`: A memory address operand. In MIPS-SE mode, allows a base address register plus 16-bit immediate offset. In MIPS mode, just a base register.
- `R`: A memory address operand. In MIPS-SE mode, allows a base address register plus a 9-bit signed offset. In MIPS mode, the same as constraint `m`.
- `ZC`: A memory address operand, suitable for use in a `pref`, `ll`, or `sc` instruction on the given subtarget (details vary).
- `r`, `d`, `y`: A 32 or 64-bit GPR register.
- `f`: A 32 or 64-bit FPU register (`F0–F31`), or a 128-bit MSA register (`W0–W31`). In the case of MSA registers, it is recommended to use the `w` argument modifier for compatibility with GCC.
- `c`: A 32-bit or 64-bit GPR register suitable for indirect jump (always 25).
- `l`: The `lo` register, 32 or 64-bit.
- `x`: Invalid.

NVPTX:

- `b`: A 1-bit integer register.
- `c` or `h`: A 16-bit integer register.
- `r`: A 32-bit integer register.
- `l` or `N`: A 64-bit integer register.
- `f`: A 32-bit float register.
- `d`: A 64-bit float register.

PowerPC:

- `I`: An immediate signed 16-bit integer.
- `J`: An immediate unsigned 16-bit integer, shifted left 16 bits.
- `K`: An immediate unsigned 16-bit integer.
- `L`: An immediate signed 16-bit integer, shifted left 16 bits.
- `M`: An immediate integer greater than 31.
- `N`: An immediate integer that is an exact power of 2.
- `O`: The immediate integer constant 0.
- `P`: An immediate integer constant whose negation is a signed 16-bit constant.
- `es`, `o`, `Q`, `Z`, `Zy`: A memory address operand, currently treated the same as `m`.
- `r`: A 32 or 64-bit integer register.
- `b`: A 32 or 64-bit integer register, excluding `R0` (that is: `R1–R31`).
- `f`: A 32 or 64-bit float register (`F0–F31`), or when QPX is enabled, a 128 or 256-bit QPX register (`Q0–Q31`; aliases the `F` registers).
- `v`: For $4 \times f32$ or $4 \times f64$ types, when QPX is enabled, a 128 or 256-bit QPX register (`Q0–Q31`), otherwise a 128-bit altivec vector register (`V0–V31`).
- `y`: Condition register (`CR0–CR7`).

- `wc`: An individual CR bit in a CR register.
- `wa`, `wd`, `wf`: Any 128-bit VSX vector register, from the full VSX register set (overlapping both the floating-point and vector register files).
- `ws`: A 32 or 64-bit floating point register, from the full VSX register set.

Sparc:

- `I`: An immediate 13-bit signed integer.
- `r`: A 32-bit integer register.
- `f`: Any floating-point register on SparcV8, or a floating point register in the “low” half of the registers on SparcV9.
- `e`: Any floating point register. (Same as `f` on SparcV8.)

SystemZ:

- `I`: An immediate unsigned 8-bit integer.
- `J`: An immediate unsigned 12-bit integer.
- `K`: An immediate signed 16-bit integer.
- `L`: An immediate signed 20-bit integer.
- `M`: An immediate integer 0x7ffffff.
- `Q`: A memory address operand with a base address and a 12-bit immediate unsigned displacement.
- `R`: A memory address operand with a base address, a 12-bit immediate unsigned displacement, and an index register.
- `S`: A memory address operand with a base address and a 20-bit immediate signed displacement.
- `T`: A memory address operand with a base address, a 20-bit immediate signed displacement, and an index register.
- `r` or `d`: A 32, 64, or 128-bit integer register.
- `a`: A 32, 64, or 128-bit integer address register (excludes R0, which in an address context evaluates as zero).
- `h`: A 32-bit value in the high part of a 64bit data register (LLVM-specific)
- `f`: A 32, 64, or 128-bit floating point register.

X86:

- `I`: An immediate integer between 0 and 31.
- `J`: An immediate integer between 0 and 64.
- `K`: An immediate signed 8-bit integer.
- `L`: An immediate integer, 0xff or 0xffff or (in 64-bit mode only) 0xffffffff.
- `M`: An immediate integer between 0 and 3.
- `N`: An immediate unsigned 8-bit integer.
- `O`: An immediate integer between 0 and 127.
- `e`: An immediate 32-bit signed integer.
- `Z`: An immediate 32-bit unsigned integer.
- `o`, `v`: Treated the same as `m`, at the moment.

- `q`: An 8, 16, 32, or 64-bit register which can be accessed as an 8-bit `l` integer register. On X86-32, this is the `a`, `b`, `c`, and `d` registers, and on X86-64, it is all of the integer registers.
- `Q`: An 8, 16, 32, or 64-bit register which can be accessed as an 8-bit `h` integer register. This is the `a`, `b`, `c`, and `d` registers.
- `r` or `l`: An 8, 16, 32, or 64-bit integer register.
- `R`: An 8, 16, 32, or 64-bit “legacy” integer register – one which has existed since i386, and can be accessed without the REX prefix.
- `f`: A 32, 64, or 80-bit ‘387 FPU stack pseudo-register.
- `y`: A 64-bit MMX register, if MMX is enabled.
- `x`: If SSE is enabled: a 32 or 64-bit scalar operand, or 128-bit vector operand in a SSE register. If AVX is also enabled, can also be a 256-bit vector operand in an AVX register. If AVX-512 is also enabled, can also be a 512-bit vector operand in an AVX512 register. Otherwise, an error.
- `Y`: The same as `x`, if SSE2 is enabled, otherwise an error.
- `A`: Special case: allocates EAX first, then EDX, for a single operand (in 32-bit mode, a 64-bit integer operand will get split into two registers). It is not recommended to use this constraint, as in 64-bit mode, the 64-bit operand will get allocated only to RAX – if two 32-bit operands are needed, you’re better off splitting it yourself, before passing it to the asm statement.

XCore:

- `r`: A 32-bit integer register.

Asm template argument modifiers

In the asm template string, modifiers can be used on the operand reference, like “`{0:n}`”.

The modifiers are, in general, expected to behave the same way they do in GCC. LLVM’s support is often implemented on an ‘as-needed’ basis, to support C inline asm code which was supported by GCC. A mismatch in behavior between LLVM and GCC likely indicates a bug in LLVM.

Target-independent:

- `c`: Print an immediate integer constant unadorned, without the target-specific immediate punctuation (e.g. no \$ prefix).
- `n`: Negate and print immediate integer constant unadorned, without the target-specific immediate punctuation (e.g. no \$ prefix).
- `l`: Print as an unadorned label, without the target-specific label punctuation (e.g. no \$ prefix).

AArch64:

- `w`: Print a GPR register with a `w*` name instead of `x*` name. E.g., instead of `x30`, print `w30`.
- `x`: Print a GPR register with a `x*` name. (this is the default, anyhow).
- `b`, `h`, `s`, `d`, `q`: Print a floating-point/SIMD register with a `b*`, `h*`, `s*`, `d*`, or `q*` name, rather than the default of `v*`.

AMDGPU:

- `r`: No effect.

ARM:

- `a`: Print an operand as an address (with `[` and `]` surrounding a register).

- `P`: No effect.
- `q`: No effect.
- `y`: Print a VFP single-precision register as an indexed double (e.g. print as `d4 [1]` instead of `s9`)
- `B`: Bitwise invert and print an immediate integer constant without `#` prefix.
- `L`: Print the low 16-bits of an immediate integer constant.
- `M`: Print as a register set suitable for `ldm/stm`. Also prints *all* register operands subsequent to the specified one (!), so use carefully.
- `Q`: Print the low-order register of a register-pair, or the low-order register of a two-register operand.
- `R`: Print the high-order register of a register-pair, or the high-order register of a two-register operand.
- `H`: Print the second register of a register-pair. (On a big-endian system, `H` is equivalent to `Q`, and on little-endian system, `H` is equivalent to `R`.)
- `e`: Print the low doubleword register of a NEON quad register.
- `f`: Print the high doubleword register of a NEON quad register.
- `m`: Print the base register of a memory operand without the `[` and `]` adornment.

Hexagon:

- `L`: Print the second register of a two-register operand. Requires that it has been allocated consecutively to the first.
- `I`: Print the letter 'i' if the operand is an integer constant, otherwise nothing. Used to print 'addi' vs 'add' instructions.

MSP430:

No additional modifiers.

MIPS:

- `X`: Print an immediate integer as hexadecimal
- `x`: Print the low 16 bits of an immediate integer as hexadecimal.
- `d`: Print an immediate integer as decimal.
- `m`: Subtract one and print an immediate integer as decimal.
- `z`: Print \$0 if an immediate zero, otherwise print normally.
- `L`: Print the low-order register of a two-register operand, or prints the address of the low-order word of a double-word memory operand.
- `M`: Print the high-order register of a two-register operand, or prints the address of the high-order word of a double-word memory operand.
- `D`: Print the second register of a two-register operand, or prints the second word of a double-word memory operand. (On a big-endian system, `D` is equivalent to `L`, and on little-endian system, `D` is equivalent to `M`.)
- `w`: No effect. Provided for compatibility with GCC which requires this modifier in order to print MSA registers (`W0-W31`) with the `f` constraint.

NVPTX:

- `r`: No effect.

PowerPC:

- **L**: Print the second register of a two-register operand. Requires that it has been allocated consecutively to the first.
- **I**: Print the letter ‘i’ if the operand is an integer constant, otherwise nothing. Used to print ‘addi’ vs ‘add’ instructions.
- **y**: For a memory operand, prints formatter for a two-register X-form instruction. (Currently always prints `r0, OPERAND`).
- **U**: Prints ‘u’ if the memory operand is an update form, and nothing otherwise. (NOTE: LLVM does not support update form, so this will currently always print nothing)
- **X**: Prints ‘x’ if the memory operand is an indexed form. (NOTE: LLVM does not support indexed form, so this will currently always print nothing)

Sparc:

- **r**: No effect.

SystemZ:

SystemZ implements only **n**, and does *not* support any of the other target-independent modifiers.

X86:

- **c**: Print an unadorned integer or symbol name. (The latter is target-specific behavior for this typically target-independent modifier).
- **A**: Print a register name with a ‘*’ before it.
- **b**: Print an 8-bit register name (e.g. `al`); do nothing on a memory operand.
- **h**: Print the upper 8-bit register name (e.g. `ah`); do nothing on a memory operand.
- **w**: Print the 16-bit register name (e.g. `ax`); do nothing on a memory operand.
- **k**: Print the 32-bit register name (e.g. `eax`); do nothing on a memory operand.
- **q**: Print the 64-bit register name (e.g. `rax`), if 64-bit registers are available, otherwise the 32-bit register name; do nothing on a memory operand.
- **n**: Negate and print an unadorned integer, or, for operands other than an immediate integer (e.g. a relocatable symbol expression), print a ‘-’ before the operand. (The behavior for relocatable symbol expressions is a target-specific behavior for this typically target-independent modifier)
- **H**: Print a memory reference with additional offset +8.
- **P**: Print a memory reference or operand for use as the argument of a call instruction. (E.g. omit `(rip)`, even though it’s PC-relative.)

XCore:

No additional modifiers.

Inline Asm Metadata

The call instructions that wrap inline asm nodes may have a “!srcloc” MDNode attached to it that contains a list of constant integers. If present, the code generator will use the integer as the location cookie value when report errors through the `LLVMContext` error reporting mechanisms. This allows a front-end to correlate backend errors that occur with inline asm back to the source code that produced it. For example:

```
call void asm sideeffect "something bad", "()", !srcloc !42
...
!42 = !{ i32 1234567 }
```

It is up to the front-end to make sense of the magic numbers it places in the IR. If the MDNode contains multiple constants, the code generator will use the one that corresponds to the line of the asm that the error occurs on.

1.1.8 Metadata

LLVM IR allows metadata to be attached to instructions in the program that can convey extra information about the code to the optimizers and code generator. One example application of metadata is source-level debug information. There are two metadata primitives: strings and nodes.

Metadata does not have a type, and is not a value. If referenced from a `call` instruction, it uses the `metadata` type.

All metadata are identified in syntax by an exclamation point (`!`).

Metadata Nodes and Metadata Strings

A metadata string is a string surrounded by double quotes. It can contain any character by escaping non-printable characters with `"\xx"` where `"xx"` is the two digit hex code. For example: `!"test\00"`.

Metadata nodes are represented with notation similar to structure constants (a comma separated list of elements, surrounded by braces and preceded by an exclamation point). Metadata nodes can have any values as their operand. For example:

```
!{ !"test\00", i32 10 }
```

Metadata nodes that aren't unique use the `distinct` keyword. For example:

```
!0 = distinct !{ !"test\00", i32 10 }
```

`distinct` nodes are useful when nodes shouldn't be merged based on their content. They can also occur when transformations cause uniquing collisions when metadata operands change.

A *named metadata* is a collection of metadata nodes, which can be looked up in the module symbol table. For example:

```
!foo = !{!4, !3}
```

Metadata can be used as function arguments. Here `llvm.dbg.value` function is using two metadata arguments:

```
call void @llvm.dbg.value(metadata !24, i64 0, metadata !25)
```

Metadata can be attached to an instruction. Here metadata `!21` is attached to the `add` instruction using the `!dbg` identifier:

```
%indvar.next = add i64 %indvar, 1, !dbg !21
```

Metadata can also be attached to a function or a global variable. Here metadata `!22` is attached to the `f1` and `f2` functions, and the globals `g1` and `g2` using the `!dbg` identifier:

```
declare !dbg !22 void @f1()
define void @f2() !dbg !22 {
    ret void
}

@g1 = global i32 0, !dbg !22
@g2 = external global i32, !dbg !22
```

A transformation is required to drop any metadata attachment that it does not know or know it can't preserve. Currently there is an exception for metadata attachment to globals for `!type` and `!absolute_symbol` which can't be unconditionally dropped unless the global is itself deleted.

Metadata attached to a module using named metadata may not be dropped, with the exception of debug metadata (named metadata with the name `!llvm.dbg.*`).

More information about specific metadata nodes recognized by the optimizers and code generator is found below.

Specialized Metadata Nodes

Specialized metadata nodes are custom data structures in metadata (as opposed to generic tuples). Their fields are labelled, and can be specified in any order.

These aren't inherently debug info centric, but currently all the specialized metadata nodes are related to debug info.

DICompileUnit

`DICompileUnit` nodes represent a compile unit. The `enums:`, `retainedTypes:`, `globals:`, `imports:` and `macros:` fields are tuples containing the debug info to be emitted along with the compile unit, regardless of code optimizations (some nodes are only emitted if there are references to them from instructions). The `debugInfoForProfiling:` field is a boolean indicating whether or not line-table discriminators are updated to provide more-accurate debug info for profiling results.

```
!0 = !DICompileUnit(language: DW_LANG_C99, file: !1, producer: "clang",
                    isOptimized: true, flags: "-O2", runtimeVersion: 2,
                    splitDebugFilename: "abc.debug", emissionKind: FullDebug,
                    enums: !2, retainedTypes: !3, globals: !4, imports: !5,
                    macros: !6, dwoId: 0x0abcd)
```

Compile unit descriptors provide the root scope for objects declared in a specific compilation unit. File descriptors are defined using this scope. These descriptors are collected by a named metadata node `!llvm.dbg.cu`. They keep track of global variables, type information, and imported entities (declarations and namespaces).

DIFile

`DIFile` nodes represent files. The `filename:` can include slashes.

```
!0 = !DIFile(filename: "path/to/file", directory: "/path/to/dir",
             checksumkind: CSK_MD5,
             checksum: "000102030405060708090a0b0c0d0e0f")
```

Files are sometimes used in `scope:` fields, and are the only valid target for `file:` fields. Valid values for `checksumkind:` field are: `{CSK_None, CSK_MD5, CSK_SHA1}`

DIBasicType

`DIBasicType` nodes represent primitive types, such as `int`, `bool` and `float`. `tag:` defaults to `DW_TAG_base_type`.

```
!0 = !DIBasicType(name: "unsigned char", size: 8, align: 8,
                  encoding: DW_ATE_unsigned_char)
!1 = !DIBasicType(tag: DW_TAG_unspecified_type, name: "decltype(nullptr)")
```

The `encoding`: describes the details of the type. Usually it's one of the following:

```
DW_ATE_address      = 1
DW_ATE_boolean      = 2
DW_ATE_float        = 4
DW_ATE_signed       = 5
DW_ATE_signed_char  = 6
DW_ATE_unsigned     = 7
DW_ATE_unsigned_char = 8
```

DISubroutineType

`DISubroutineType` nodes represent subroutine types. Their `types`: field refers to a tuple; the first operand is the return type, while the rest are the types of the formal arguments in order. If the first operand is `null`, that represents a function with no return value (such as `void foo() {}` in C++).

```
!0 = !BasicType(name: "int", size: 32, align: 32, DW_ATE_signed)
!1 = !BasicType(name: "char", size: 8, align: 8, DW_ATE_signed_char)
!2 = !DISubroutineType(types: [{null, !0, !1}] ; void (int, char)
```

DIDerivedType

`DIDerivedType` nodes represent types derived from other types, such as qualified types.

```
!0 = !DIBasicType(name: "unsigned char", size: 8, align: 8,
                  encoding: DW_ATE_unsigned_char)
!1 = !DIDerivedType(tag: DW_TAG_pointer_type, baseType: !0, size: 32,
                  align: 32)
```

The following tag: values are valid:

```
DW_TAG_member      = 13
DW_TAG_pointer_type = 15
DW_TAG_reference_type = 16
DW_TAG_typedef      = 22
DW_TAG_inheritance  = 28
DW_TAG_ptr_to_member_type = 31
DW_TAG_const_type   = 38
DW_TAG_friend       = 42
DW_TAG_volatile_type = 53
DW_TAG_restrict_type = 55
DW_TAG_atomic_type  = 71
```

`DW_TAG_member` is used to define a member of a *composite type*. The type of the member is the `baseType`:. The `offset`: is the member's bit offset. If the composite type has an `ODR identifier`: and does not set `flags: DIFwdDecl`, then the member is uniqued based only on its `name`: and `scope`:.

`DW_TAG_inheritance` and `DW_TAG_friend` are used in the `elements`: field of *composite types* to describe parents and friends.

`DW_TAG_typedef` is used to provide a name for the `baseType`:.

`DW_TAG_pointer_type`, `DW_TAG_reference_type`, `DW_TAG_const_type`, `DW_TAG_volatile_type`, `DW_TAG_restrict_type` and `DW_TAG_atomic_type` are used to qualify the `baseType`:.

Note that the `void *` type is expressed as a type derived from `NULL`.

DICompositeType

`DICompositeType` nodes represent types composed of other types, like structures and unions. `elements`: points to a tuple of the composed types.

If the source language supports ODR, the `identifier`: field gives the unique identifier used for type merging between modules. When specified, *subprogram declarations* and *member derived types* that reference the ODR-type in their `scope`: change unquing rules.

For a given `identifier`:, there should only be a single composite type that does not have `flags`: `DIFlagFwdDecl` set. LLVM tools that link modules together will unique such definitions at parse time via the `identifier`: field, even if the nodes are distinct.

```
!0 = !DIEnumerator(name: "SixKind", value: 7)
!1 = !DIEnumerator(name: "SevenKind", value: 7)
!2 = !DIEnumerator(name: "NegEightKind", value: -8)
!3 = !DICompositeType(tag: DW_TAG_enumeration_type, name: "Enum", file: !12,
                      line: 2, size: 32, align: 32, identifier: "_M4Enum",
                      elements: !(!0, !1, !2))
```

The following `tag`: values are valid:

```
DW_TAG_array_type      = 1
DW_TAG_class_type      = 2
DW_TAG_enumeration_type = 4
DW_TAG_structure_type  = 19
DW_TAG_union_type      = 23
```

For `DW_TAG_array_type`, the `elements`: should be *subrange descriptors*, each representing the range of subscripts at that level of indexing. The `DIFlagVector` flag to `flags`: indicates that an array type is a native packed vector.

For `DW_TAG_enumeration_type`, the `elements`: should be *enumerator descriptors*, each representing the definition of an enumeration value for the set. All enumeration type descriptors are collected in the `enums`: field of the *compile unit*.

For `DW_TAG_structure_type`, `DW_TAG_class_type`, and `DW_TAG_union_type`, the `elements`: should be *derived types* with `tag`: `DW_TAG_member`, `tag`: `DW_TAG_inheritance`, or `tag`: `DW_TAG_friend`; or *subprograms* with `isDefinition`: `false`.

DISubrange

`DISubrange` nodes are the elements for `DW_TAG_array_type` variants of *DICompositeType*. `count`: `-1` indicates an empty array.

```
!0 = !DISubrange(count: 5, lowerBound: 0) ; array counting from 0
!1 = !DISubrange(count: 5, lowerBound: 1) ; array counting from 1
!2 = !DISubrange(count: -1) ; empty array.
```

DIEnumerator

`DIEnumerator` nodes are the elements for `DW_TAG_enumeration_type` variants of *DICompositeType*.

```
!0 = !DIEnumerator(name: "SixKind", value: 7)
!1 = !DIEnumerator(name: "SevenKind", value: 7)
!2 = !DIEnumerator(name: "NegEightKind", value: -8)
```

DITemplateTypeParameter

DITemplateTypeParameter nodes represent type parameters to generic source language constructs. They are used (optionally) in *DICompositeType* and *DISubprogram* `templateParams`: fields.

```
!0 = !DITemplateTypeParameter(name: "Ty", type: !1)
```

DITemplateValueParameter

DITemplateValueParameter nodes represent value parameters to generic source language constructs. `tag`: defaults to `DW_TAG_template_value_parameter`, but if specified can also be set to `DW_TAG_GNU_template_template_param` or `DW_TAG_GNU_template_param_pack`. They are used (optionally) in *DICompositeType* and *DISubprogram* `templateParams`: fields.

```
!0 = !DITemplateValueParameter(name: "Ty", type: !1, value: i32 7)
```

DINamespace

DINamespace nodes represent namespaces in the source language.

```
!0 = !DINamespace(name: "myawesomeproject", scope: !1, file: !2, line: 7)
```

DIGlobalVariable

DIGlobalVariable nodes represent global variables in the source language.

```
!0 = !DIGlobalVariable(name: "foo", linkageName: "foo", scope: !1,
                        file: !2, line: 7, type: !3, isLocal: true,
                        isDefinition: false, variable: i32* @foo,
                        declaration: !4)
```

All global variables should be referenced by the *globals*: field of a *compile unit*.

DISubprogram

DISubprogram nodes represent functions from the source language. A DISubprogram may be attached to a function definition using `!dbg` metadata. The `variables`: field points at *variables* that must be retained, even if their IR counterparts are optimized out of the IR. The `type`: field must point at a *DISubroutineType*. When `isDefinition: false`, subprograms describe a declaration in the type tree as opposed to a definition of a function. If the scope is a composite type with an `ODR identifier`: and that does not set `flags: DIFwdDecl`, then the subprogram declaration is uniqued based only on its `linkageName`: and `scope`:.


```
define void @_Z3foov() !dbg !0 {
    ...
}

!0 = distinct !DISubprogram(name: "foo", linkageName: "_Zfoov", scope: !1,
                             file: !2, line: 7, type: !3, isLocal: true,
                             isDefinition: true, scopeLine: 8,
                             containingType: !4,
                             virtuality: DW_VIRTUALITY_pure_virtual,
                             virtualIndex: 10, flags: DIFlagPrototyped,
                             isOptimized: true, unit: !5, templateParams: !6,
                             declaration: !7, variables: !8, thrownTypes: !9)
```

DILexicalBlock

DILexicalBlock nodes describe nested blocks within a *subprogram*. The line number and column numbers are used to distinguish two lexical blocks at same depth. They are valid targets for `scope:` fields.

```
!0 = distinct !DILexicalBlock(scope: !1, file: !2, line: 7, column: 35)
```

Usually lexical blocks are `distinct` to prevent node merging based on operands.

DILexicalBlockFile

DILexicalBlockFile nodes are used to discriminate between sections of a *lexical block*. The `file:` field can be changed to indicate textual inclusion, or the `discriminator:` field can be used to discriminate between control flow within a single block in the source language.

```
!0 = !DILexicalBlock(scope: !3, file: !4, line: 7, column: 35)
!1 = !DILexicalBlockFile(scope: !0, file: !4, discriminator: 0)
!2 = !DILexicalBlockFile(scope: !0, file: !4, discriminator: 1)
```

DILocation

DILocation nodes represent source debug locations. The `scope:` field is mandatory, and points at an *DILexicalBlockFile*, an *DILexicalBlock*, or an *DISubprogram*.

```
!0 = !DILocation(line: 2900, column: 42, scope: !1, inlinedAt: !2)
```

DILocalVariable

DILocalVariable nodes represent local variables in the source language. If the `arg:` field is set to non-zero, then this variable is a subprogram parameter, and it will be included in the `variables:` field of its *DISubprogram*.

```
!0 = !DILocalVariable(name: "this", arg: 1, scope: !3, file: !2, line: 7,
                      type: !3, flags: DIFlagArtificial)
!1 = !DILocalVariable(name: "x", arg: 2, scope: !4, file: !2, line: 7,
                      type: !3)
!2 = !DILocalVariable(name: "y", scope: !5, file: !2, line: 7, type: !3)
```

DIExpression

DIExpression nodes represent expressions that are inspired by the DWARF expression language. They are used in *debug intrinsics* (such as `llvm.dbg.declare` and `llvm.dbg.value`) to describe how the referenced LLVM variable relates to the source language variable.

The current supported vocabulary is limited:

- `DW_OP_deref` dereferences the top of the expression stack.
- `DW_OP_plus` pops the last two entries from the expression stack, adds them together and appends the result to the expression stack.
- `DW_OP_minus` pops the last two entries from the expression stack, subtracts the last entry from the second last entry and appends the result to the expression stack.
- `DW_OP_plus_uconst, 93` adds 93 to the working expression.
- `DW_OP_LLVM_fragment, 16, 8` specifies the offset and size (16 and 8 here, respectively) of the variable fragment from the working expression. Note that contrary to `DW_OP_bit_piece`, the offset is describing the location within the described source variable.
- `DW_OP_swap` swaps top two stack entries.
- `DW_OP_xderef` provides extended dereference mechanism. The entry at the top of the stack is treated as an address. The second stack entry is treated as an address space identifier.
- `DW_OP_stack_value` marks a constant value.

DWARF specifies three kinds of simple location descriptions: Register, memory, and implicit location descriptions. Register and memory location descriptions describe the *location* of a source variable (in the sense that a debugger might modify its value), whereas implicit locations describe merely the *value* of a source variable. DIExpressions also follow this model: A DIExpression that doesn't have a trailing `DW_OP_stack_value` will describe an *address* when combined with a concrete location.

```
!0 = !DIExpression(DW_OP_deref)
!1 = !DIExpression(DW_OP_plus_uconst, 3)
!1 = !DIExpression(DW_OP_constu, 3, DW_OP_plus)
!2 = !DIExpression(DW_OP_bit_piece, 3, 7)
!3 = !DIExpression(DW_OP_deref, DW_OP_constu, 3, DW_OP_plus, DW_OP_LLVM_fragment, 3, 7)
!4 = !DIExpression(DW_OP_constu, 2, DW_OP_swap, DW_OP_xderef)
!5 = !DIExpression(DW_OP_constu, 42, DW_OP_stack_value)
```

DIObjCProperty

DIObjCProperty nodes represent Objective-C property nodes.

```
!3 = !DIObjCProperty(name: "foo", file: !1, line: 7, setter: "setFoo",
                     getter: "getFoo", attributes: 7, type: !2)
```

DIIImportedEntity

DIIImportedEntity nodes represent entities (such as modules) imported into a compile unit.

```
!2 = !DIIImportedEntity(tag: DW_TAG_imported_module, name: "foo", scope: !0,
                       entity: !1, line: 7)
```

DIMacro

`DIMacro` nodes represent definition or undefinition of a macro identifiers. The `name:` field is the macro identifier, followed by macro parameters when defining a function-like macro, and the `value` field is the token-string used to expand the macro identifier.

```
!2 = !DIMacro(macinfo: DW_MACINFO_define, line: 7, name: "foo(x)",
              value: "((x) + 1)")
!3 = !DIMacro(macinfo: DW_MACINFO_undef, line: 30, name: "foo")
```

DIMacroFile

`DIMacroFile` nodes represent inclusion of source files. The `nodes:` field is a list of `DIMacro` and `DIMacroFile` nodes that appear in the included source file.

```
!2 = !DIMacroFile(macinfo: DW_MACINFO_start_file, line: 7, file: !2,
                  nodes: !3)
```

‘tbaa’ Metadata

In LLVM IR, memory does not have types, so LLVM’s own type system is not suitable for doing type based alias analysis (TBAA). Instead, metadata is added to the IR to describe a type system of a higher level language. This can be used to implement C/C++ strict type aliasing rules, but it can also be used to implement custom alias analysis behavior for other languages.

This description of LLVM’s TBAA system is broken into two parts: *Semantics* talks about high level issues, and *Representation* talks about the metadata encoding of various entities.

It is always possible to trace any TBAA node to a “root” TBAA node (details in the *Representation* section). TBAA nodes with different roots have an unknown aliasing relationship, and LLVM conservatively infers `MayAlias` between them. The rules mentioned in this section only pertain to TBAA nodes living under the same root.

Semantics

The TBAA metadata system, referred to as “struct path TBAA” (not to be confused with `tbaa.struct`), consists of the following high level concepts: *Type Descriptors*, further subdivided into scalar type descriptors and struct type descriptors; and *Access Tags*.

Type descriptors describe the type system of the higher level language being compiled. **Scalar type descriptors** describe types that do not contain other types. Each scalar type has a parent type, which must also be a scalar type or the TBAA root. Via this parent relation, scalar types within a TBAA root form a tree. **Struct type descriptors** denote types that contain a sequence of other type descriptors, at known offsets. These contained type descriptors can either be struct type descriptors themselves or scalar type descriptors.

Access tags are metadata nodes attached to load and store instructions. Access tags use type descriptors to describe the *location* being accessed in terms of the type system of the higher level language. Access tags are tuples consisting of a base type, an access type and an offset. The base type is a scalar type descriptor or a struct type descriptor, the access type is a scalar type descriptor, and the offset is a constant integer.

The access tag (`BaseTy`, `AccessTy`, `Offset`) can describe one of two things:

- If `BaseTy` is a struct type, the tag describes a memory access (load or store) of a value of type `AccessTy` contained in the struct type `BaseTy` at offset `Offset`.

- If `BaseTy` is a scalar type, `Offset` must be 0 and `BaseTy` and `AccessTy` must be the same; and the access tag describes a scalar access with scalar type `AccessTy`.

We first define an `ImmediateParent` relation on `(BaseTy, Offset)` tuples this way:

- If `BaseTy` is a scalar type then `ImmediateParent(BaseTy, 0)` is `(ParentTy, 0)` where `ParentTy` is the parent of the scalar type as described in the TBAA metadata. `ImmediateParent(BaseTy, Offset)` is undefined if `Offset` is non-zero.
- If `BaseTy` is a struct type then `ImmediateParent(BaseTy, Offset)` is `(NewTy, NewOffset)` where `NewTy` is the type contained in `BaseTy` at offset `Offset` and `NewOffset` is `Offset` adjusted to be relative within that inner type.

A memory access with an access tag `(BaseTy1, AccessTy1, Offset1)` aliases a memory access with an access tag `(BaseTy2, AccessTy2, Offset2)` if either `(BaseTy1, Offset1)` is reachable from `(BaseTy2, Offset2)` via the `Parent` relation or vice versa.

As a concrete example, the type descriptor graph for the following program

```
struct Inner {
    int i;    // offset 0
    float f;  // offset 4
};

struct Outer {
    float f;  // offset 0
    double d; // offset 4
    struct Inner inner_a; // offset 12
};

void f(struct Outer* outer, struct Inner* inner, float* f, int* i, char* c) {
    outer->f = 0;           // tag0: (OuterStructTy, FloatScalarTy, 0)
    outer->inner_a.i = 0;    // tag1: (OuterStructTy, IntScalarTy, 12)
    outer->inner_a.f = 0.0;  // tag2: (OuterStructTy, IntScalarTy, 16)
    *f = 0.0;              // tag3: (FloatScalarTy, FloatScalarTy, 0)
}
```

is (note that in C and C++, `char` can be used to access any arbitrary type):

```
Root = "TBAA Root"
CharScalarTy = ("char", Root, 0)
FloatScalarTy = ("float", CharScalarTy, 0)
DoubleScalarTy = ("double", CharScalarTy, 0)
IntScalarTy = ("int", CharScalarTy, 0)
InnerStructTy = {"Inner" (IntScalarTy, 0), (FloatScalarTy, 4)}
OuterStructTy = {"Outer", (FloatScalarTy, 0), (DoubleScalarTy, 4),
                 (InnerStructTy, 12)}
```

with (e.g.) `ImmediateParent(OuterStructTy, 12) = (InnerStructTy, 0)`, `ImmediateParent(InnerStructTy, 0) = (IntScalarTy, 0)`, and `ImmediateParent(IntScalarTy, 0) = (CharScalarTy, 0)`.

Representation

The root node of a TBAA type hierarchy is an `MDNode` with 0 operands or with exactly one `MDString` operand.

Scalar type descriptors are represented as an `MDNode` with two operands. The first operand is an `MDString` denoting the name of the struct type. LLVM does not assign meaning to the value of this operand, it only cares about

it being an `MDString`. The second operand is an `MDNode` which points to the parent for said scalar type descriptor, which is either another scalar type descriptor or the TBAA root. Scalar type descriptors can have an optional third argument, but that must be the constant integer zero.

Struct type descriptors are represented as `MDNode`s with an odd number of operands greater than 1. The first operand is an `MDString` denoting the name of the struct type. Like in scalar type descriptors the actual value of this name operand is irrelevant to LLVM. After the name operand, the struct type descriptors have a sequence of alternating `MDNode` and `ConstantInt` operands. With N starting from 1, the $2N - 1$ th operand, an `MDNode`, denotes a contained field, and the $2N$ th operand, a `ConstantInt`, is the offset of the said contained field. The offsets must be in non-decreasing order.

Access tags are represented as `MDNode`s with either 3 or 4 operands. The first operand is an `MDNode` pointing to the node representing the base type. The second operand is an `MDNode` pointing to the node representing the access type. The third operand is a `ConstantInt` that states the offset of the access. If a fourth field is present, it must be a `ConstantInt` valued at 0 or 1. If it is 1 then the access tag states that the location being accessed is “constant” (meaning `pointsToConstantMemory` should return true; see [other useful AliasAnalysis methods](#)). The TBAA root of the access type and the base type of an access tag must be the same, and that is the TBAA root of the access tag.

‘tbaa.struct’ Metadata

The `llvm.memcpy` is often used to implement aggregate assignment operations in C and similar languages, however it is defined to copy a contiguous region of memory, which is more than strictly necessary for aggregate types which contain holes due to padding. Also, it doesn’t contain any TBAA information about the fields of the aggregate.

`!tbaa.struct` metadata can describe which memory subregions in a `memcpy` are padding and what the TBAA tags of the struct are.

The current metadata format is very simple. `!tbaa.struct` metadata nodes are a list of operands which are in conceptual groups of three. For each group of three, the first operand gives the byte offset of a field in bytes, the second gives its size in bytes, and the third gives its tbaa tag. e.g.:

```
!4 = !{ i64 0, i64 4, !1, i64 8, i64 4, !2 }
```

This describes a struct with two fields. The first is at offset 0 bytes with size 4 bytes, and has tbaa tag !1. The second is at offset 8 bytes and has size 4 bytes and has tbaa tag !2.

Note that the fields need not be contiguous. In this example, there is a 4 byte gap between the two fields. This gap represents padding which does not carry useful data and need not be preserved.

‘noalias’ and ‘alias.scope’ Metadata

`noalias` and `alias.scope` metadata provide the ability to specify generic `noalias` memory-access sets. This means that some collection of memory access instructions (loads, stores, memory-accessing calls, etc.) that carry `noalias` metadata can specifically be specified not to alias with some other collection of memory access instructions that carry `alias.scope` metadata. Each type of metadata specifies a list of scopes where each scope has an id and a domain.

When evaluating an aliasing query, if for some domain, the set of scopes with that domain in one instruction’s `alias.scope` list is a subset of (or equal to) the set of scopes for that domain in another instruction’s `noalias` list, then the two memory accesses are assumed not to alias.

Because scopes in one domain don’t affect scopes in other domains, separate domains can be used to compose multiple independent `noalias` sets. This is used for example during inlining. As the `noalias` function parameters are turned into `noalias scope` metadata, a new domain is used every time the function is inlined.

The metadata identifying each domain is itself a list containing one or two entries. The first entry is the name of the domain. Note that if the name is a string then it can be combined across functions and translation units. A self-reference can be used to create globally unique domain names. A descriptive string may optionally be provided as a second list entry.

The metadata identifying each scope is also itself a list containing two or three entries. The first entry is the name of the scope. Note that if the name is a string then it can be combined across functions and translation units. A self-reference can be used to create globally unique scope names. A metadata reference to the scope's domain is the second entry. A descriptive string may optionally be provided as a third list entry.

For example,

```
; Two scope domains:
!0 = !{!0}
!1 = !{!1}

; Some scopes in these domains:
!2 = !{!2, !0}
!3 = !{!3, !0}
!4 = !{!4, !1}

; Some scope lists:
!5 = !{!4} ; A list containing only scope !4
!6 = !{!4, !3, !2}
!7 = !{!3}

; These two instructions don't alias:
%0 = load float, float* %c, align 4, !alias.scope !5
store float %0, float* %arrayidx.i, align 4, !noalias !5

; These two instructions also don't alias (for domain !1, the set of scopes
; in the !alias.scope equals that in the !noalias list):
%2 = load float, float* %c, align 4, !alias.scope !5
store float %2, float* %arrayidx.i2, align 4, !noalias !6

; These two instructions may alias (for domain !0, the set of scopes in
; the !noalias list is not a superset of, or equal to, the scopes in the
; !alias.scope list):
%2 = load float, float* %c, align 4, !alias.scope !6
store float %0, float* %arrayidx.i, align 4, !noalias !7
```

'fpmath' Metadata

fpmath metadata may be attached to any instruction of floating point type. It can be used to express the maximum acceptable error in the result of that instruction, in ULPs, thus potentially allowing the compiler to use a more efficient but less accurate method of computing it. ULP is defined as follows:

If x is a real number that lies between two finite consecutive floating-point numbers a and b , without being equal to one of them, then $\text{ulp}(x) = |b - a|$, otherwise $\text{ulp}(x)$ is the distance between the two non-equal finite floating-point numbers nearest x . Moreover, $\text{ulp}(\text{NaN})$ is NaN.

The metadata node shall consist of a single positive float type number representing the maximum relative error, for example:

```
!0 = !{ float 2.5 } ; maximum acceptable inaccuracy is 2.5 ULPs
```

‘range’ Metadata

range metadata may be attached only to load, call and invoke of integer types. It expresses the possible ranges the loaded value or the value returned by the called function at this call site is in. The ranges are represented with a flattened list of integers. The loaded value or the value returned is known to be in the union of the ranges defined by each consecutive pair. Each pair has the following properties:

- The type must match the type loaded by the instruction.
- The pair a, b represents the range $[a, b)$.
- Both a and b are constants.
- The range is allowed to wrap.
- The range should not represent the full or empty set. That is, $a \neq b$.

In addition, the pairs must be in signed order of the lower bound and they must be non-contiguous.

Examples:

```
%a = load i8, i8* %x, align 1, !range !0 ; Can only be 0 or 1
%b = load i8, i8* %y, align 1, !range !1 ; Can only be 255 (-1), 0 or 1
%c = call i8 @foo(), !range !2 ; Can only be 0, 1, 3, 4 or 5
%d = invoke i8 @bar() to label %cont
      unwind label %lpad, !range !3 ; Can only be -2, -1, 3, 4 or 5
...
!0 = !{ i8 0, i8 2 }
!1 = !{ i8 255, i8 2 }
!2 = !{ i8 0, i8 2, i8 3, i8 6 }
!3 = !{ i8 -2, i8 0, i8 3, i8 6 }
```

‘absolute_symbol’ Metadata

absolute_symbol metadata may be attached to a global variable declaration. It marks the declaration as a reference to an absolute symbol, which causes the backend to use absolute relocations for the symbol even in position independent code, and expresses the possible ranges that the global variable’s address (not its value) is in, in the same format as range metadata, with the extension that the pair all-ones, all-ones may be used to represent the full set.

Example (assuming 64-bit pointers):

```
@a = external global i8, !absolute_symbol !0 ; Absolute symbol in range [0,256)
@b = external global i8, !absolute_symbol !1 ; Absolute symbol in range [0,2^64)
...
!0 = !{ i64 0, i64 256 }
!1 = !{ i64 -1, i64 -1 }
```

‘unpredictable’ Metadata

unpredictable metadata may be attached to any branch or switch instruction. It can be used to express the unpredictability of control flow. Similar to the llvm.expect intrinsic, it may be used to alter optimizations related to compare and branch instructions. The metadata is treated as a boolean value; if it exists, it signals that the branch or switch that it is attached to is completely unpredictable.

‘llvm.loop’

It is sometimes useful to attach information to loop constructs. Currently, loop metadata is implemented as metadata attached to the branch instruction in the loop latch block. This type of metadata refer to a metadata node that is guaranteed to be separate for each loop. The loop identifier metadata is specified with the name `llvm.loop`.

The loop identifier metadata is implemented using a metadata that refers to itself to avoid merging it with any other identifier metadata, e.g., during module linkage or function inlining. That is, each loop should refer to their own identification metadata even if they reside in separate functions. The following example contains loop identifier metadata for two separate loop constructs:

```
!0 = !{!0}
!1 = !{!1}
```

The loop identifier metadata can be used to specify additional per-loop metadata. Any operands after the first operand can be treated as user-defined metadata. For example the `llvm.loop.unroll.count` suggests an unroll factor to the loop unroller:

```
br i1 %exitcond, label %._crit_edge, label %l.r.ph, !llvm.loop !0
...
!0 = !{!0, !1}
!1 = !{"llvm.loop.unroll.count", i32 4}
```

‘llvm.loop.vectorize’ and ‘llvm.loop.interleave’

Metadata prefixed with `llvm.loop.vectorize` or `llvm.loop.interleave` are used to control per-loop vectorization and interleaving parameters such as vectorization width and interleave count. These metadata should be used in conjunction with `llvm.loop` loop identification metadata. The `llvm.loop.vectorize` and `llvm.loop.interleave` metadata are only optimization hints and the optimizer will only interleave and vectorize loops if it believes it is safe to do so. The `llvm.mem.parallel_loop_access` metadata which contains information about loop-carried memory dependencies can be helpful in determining the safety of these transformations.

‘llvm.loop.interleave.count’ Metadata

This metadata suggests an interleave count to the loop interleaver. The first operand is the string `llvm.loop.interleave.count` and the second operand is an integer specifying the interleave count. For example:

```
!0 = !{"llvm.loop.interleave.count", i32 4}
```

Note that setting `llvm.loop.interleave.count` to 1 disables interleaving multiple iterations of the loop. If `llvm.loop.interleave.count` is set to 0 then the interleave count will be determined automatically.

‘llvm.loop.vectorize.enable’ Metadata

This metadata selectively enables or disables vectorization for the loop. The first operand is the string `llvm.loop.vectorize.enable` and the second operand is a bit. If the bit operand value is 1 vectorization is enabled. A value of 0 disables vectorization:

```
!0 = !{"llvm.loop.vectorize.enable", i1 0}
!1 = !{"llvm.loop.vectorize.enable", i1 1}
```


'llvm.loop.vectorize.width' Metadata

This metadata sets the target width of the vectorizer. The first operand is the string `llvm.loop.vectorize.width` and the second operand is an integer specifying the width. For example:

```
!0 = !{"llvm.loop.vectorize.width", i32 4}
```

Note that setting `llvm.loop.vectorize.width` to 1 disables vectorization of the loop. If `llvm.loop.vectorize.width` is set to 0 or if the loop does not have this metadata the width will be determined automatically.

'llvm.loop.unroll'

Metadata prefixed with `llvm.loop.unroll` are loop unrolling optimization hints such as the unroll factor. `llvm.loop.unroll` metadata should be used in conjunction with `llvm.loop` loop identification metadata. The `llvm.loop.unroll` metadata are only optimization hints and the unrolling will only be performed if the optimizer believes it is safe to do so.

'llvm.loop.unroll.count' Metadata

This metadata suggests an unroll factor to the loop unroller. The first operand is the string `llvm.loop.unroll.count` and the second operand is a positive integer specifying the unroll factor. For example:

```
!0 = !{"llvm.loop.unroll.count", i32 4}
```

If the trip count of the loop is less than the unroll count the loop will be partially unrolled.

'llvm.loop.unroll.disable' Metadata

This metadata disables loop unrolling. The metadata has a single operand which is the string `llvm.loop.unroll.disable`. For example:

```
!0 = !{"llvm.loop.unroll.disable"}
```

'llvm.loop.unroll.runtime.disable' Metadata

This metadata disables runtime loop unrolling. The metadata has a single operand which is the string `llvm.loop.unroll.runtime.disable`. For example:

```
!0 = !{"llvm.loop.unroll.runtime.disable"}
```

'llvm.loop.unroll.enable' Metadata

This metadata suggests that the loop should be fully unrolled if the trip count is known at compile time and partially unrolled if the trip count is not known at compile time. The metadata has a single operand which is the string `llvm.loop.unroll.enable`. For example:

```
!0 = !{"llvm.loop.unroll.enable"}
```

'llvm.loop.unroll.full' Metadata

This metadata suggests that the loop should be unrolled fully. The metadata has a single operand which is the string `llvm.loop.unroll.full`. For example:

```
!0 = !{"llvm.loop.unroll.full"}
```

'llvm.loop.licm_versioning.disable' Metadata

This metadata indicates that the loop should not be versioned for the purpose of enabling loop-invariant code motion (LICM). The metadata has a single operand which is the string `llvm.loop.licm_versioning.disable`. For example:

```
!0 = !{"llvm.loop.licm_versioning.disable"}
```

'llvm.loop.distribute.enable' Metadata

Loop distribution allows splitting a loop into multiple loops. Currently, this is only performed if the entire loop cannot be vectorized due to unsafe memory dependencies. The transformation will attempt to isolate the unsafe dependencies into their own loop.

This metadata can be used to selectively enable or disable distribution of the loop. The first operand is the string `llvm.loop.distribute.enable` and the second operand is a bit. If the bit operand value is 1 distribution is enabled. A value of 0 disables distribution:

```
!0 = !{"llvm.loop.distribute.enable", i1 0}  
!1 = !{"llvm.loop.distribute.enable", i1 1}
```

This metadata should be used in conjunction with `llvm.loop` loop identification metadata.

'llvm.mem'

Metadata types used to annotate memory accesses with information helpful for optimizations are prefixed with `llvm.mem`.

'llvm.mem.parallel_loop_access' Metadata

The `llvm.mem.parallel_loop_access` metadata refers to a loop identifier, or metadata containing a list of loop identifiers for nested loops. The metadata is attached to memory accessing instructions and denotes that no loop carried memory dependence exist between it and other instructions denoted with the same loop identifier. The metadata on memory reads also implies that if conversion (i.e. speculative execution within a loop iteration) is safe.

Precisely, given two instructions `m1` and `m2` that both have the `llvm.mem.parallel_loop_access` metadata, with `L1` and `L2` being the set of loops associated with that metadata, respectively, then there is no loop carried dependence between `m1` and `m2` for loops in both `L1` and `L2`.

As a special case, if all memory accessing instructions in a loop have `llvm.mem.parallel_loop_access` metadata that refers to that loop, then the loop has no loop carried memory dependences and is considered to be a parallel loop.

Note that if not all memory access instructions have such metadata referring to the loop, then the loop is considered not being trivially parallel. Additional memory dependence analysis is required to make that determination. As a fail

safe mechanism, this causes loops that were originally parallel to be considered sequential (if optimization passes that are unaware of the parallel semantics insert new memory instructions into the loop body).

Example of a loop that is considered parallel due to its correct use of both `llvm.loop` and `llvm.mem.parallel_loop_access` metadata types that refer to the same loop identifier metadata.

```
for.body:
...
%val0 = load i32, i32* %arrayidx, !llvm.mem.parallel_loop_access !0
...
store i32 %val0, i32* %arrayidx1, !llvm.mem.parallel_loop_access !0
...
br i1 %exitcond, label %for.end, label %for.body, !llvm.loop !0

for.end:
...
!0 = !{!0}
```

It is also possible to have nested parallel loops. In that case the memory accesses refer to a list of loop identifier metadata nodes instead of the loop identifier metadata node directly:

```
outer.for.body:
...
%val1 = load i32, i32* %arrayidx3, !llvm.mem.parallel_loop_access !2
...
br label %inner.for.body

inner.for.body:
...
%val0 = load i32, i32* %arrayidx1, !llvm.mem.parallel_loop_access !0
...
store i32 %val0, i32* %arrayidx2, !llvm.mem.parallel_loop_access !0
...
br i1 %exitcond, label %inner.for.end, label %inner.for.body, !llvm.loop !1

inner.for.end:
...
store i32 %val1, i32* %arrayidx4, !llvm.mem.parallel_loop_access !2
...
br i1 %exitcond, label %outer.for.end, label %outer.for.body, !llvm.loop !2

outer.for.end:                                     ; preds = %for.body
...
!0 = !{!1, !2} ; a list of loop identifiers
!1 = !{!1} ; an identifier for the inner loop
!2 = !{!2} ; an identifier for the outer loop
```

‘invariant.group’ Metadata

The `invariant.group` metadata may be attached to load/store instructions. The existence of the `invariant.group` metadata on the instruction tells the optimizer that every load and store to the same pointer operand within the same invariant group can be assumed to load or store the same value (but see the `llvm.invariant.group.barrier` intrinsic which affects when two pointers are considered the same). Pointers returned by `bitcast` or `getelementptr` with only zero indices are considered the same.

Examples:

```
@unknownPtr = external global i8
...
%ptr = alloca i8
store i8 42, i8* %ptr, !invariant.group !0
call void @foo(i8* %ptr)

%a = load i8, i8* %ptr, !invariant.group !0 ; Can assume that value under %ptr didn't
↳change
call void @foo(i8* %ptr)
%b = load i8, i8* %ptr, !invariant.group !1 ; Can't assume anything, because group
↳changed

%newPtr = call i8* @getPointer(i8* %ptr)
%c = load i8, i8* %newPtr, !invariant.group !0 ; Can't assume anything, because we
↳only have information about %ptr

%unknownValue = load i8, i8* @unknownPtr
store i8 %unknownValue, i8* %ptr, !invariant.group !0 ; Can assume that %unknownValue
↳== 42

call void @foo(i8* %ptr)
%newPtr2 = call i8* @llvm.invariant.group.barrier(i8* %ptr)
%d = load i8, i8* %newPtr2, !invariant.group !0 ; Can't step through invariant.group.
↳barrier to get value of %ptr

...
declare void @foo(i8*)
declare i8* @getPointer(i8*)
declare i8* @llvm.invariant.group.barrier(i8*)

!0 = !{"magic ptr"}
!1 = !{"other ptr"}
```

The invariant.group metadata must be dropped when replacing one pointer by another based on aliasing information. This is because invariant.group is tied to the SSA value of the pointer operand.

```
%v = load i8, i8* %x, !invariant.group !0
; if %x must alias %y then we can replace the above instruction with
%v = load i8, i8* %y
```

‘type’ Metadata

See *Type Metadata*.

‘associated’ Metadata

The associated metadata may be attached to a global object declaration with a single argument that references another global object.

This metadata prevents discarding of the global object in linker GC unless the referenced object is also discarded. The linker support for this feature is spotty. For best compatibility, globals carrying this metadata may also:

- Be in a comdat with the referenced global.
- Be in @llvm.compiler.used.

- Have an explicit section with a name which is a valid C identifier.

It does not have any effect on non-ELF targets.

Example:

```
$a = comdat any
@a = global i32 1, comdat $a
@b = internal global i32 2, comdat $a, section "abc", !associated !0
!0 = !{i32* @a}
```

‘prof’ Metadata

The `prof` metadata is used to record profile data in the IR. The first operand of the metadata node indicates the profile metadata type. There are currently 3 types: *branch_weights*, *function_entry_count*, and *VP*.

branch_weights

Branch weight metadata attached to a branch, select, switch or call instruction represents the likeliness of the associated branch being taken. For more information, see *LLVM Branch Weight Metadata*.

function_entry_count

Function entry count metadata can be attached to function definitions to record the number of times the function is called. Used with BFI information, it is also used to derive the basic block profile count. For more information, see *LLVM Branch Weight Metadata*.

VP

VP (value profile) metadata can be attached to instructions that have value profile information. Currently this is indirect calls (where it records the hottest callees) and calls to memory intrinsics such as `memcpy`, `memmove`, and `memset` (where it records the hottest byte lengths).

Each VP metadata node contains “VP” string, then a `uint32_t` value for the value profiling kind, a `uint64_t` value for the total number of times the instruction is executed, followed by `uint64_t` value and execution count pairs. The value profiling kind is 0 for indirect call targets and 1 for memory operations. For indirect call targets, each profile value is a hash of the callee function name, and for memory operations each value is the byte length.

Note that the value counts do not need to add up to the total count listed in the third operand (in practice only the top hottest values are tracked and reported).

Indirect call example:

```
call void %f(), !prof !1
!1 = !{"VP", i32 0, i64 1600, i64 7651369219802541373, i64 1030, i64 -
↪4377547752858689819, i64 410}
```

Note that the VP type is 0 (the second operand), which indicates this is an indirect call value profile data. The third operand indicates that the indirect call executed 1600 times. The 4th and 6th operands give the hashes of the 2 hottest target functions’ names (this is the same hash used to represent function names in the profile database), and the 5th and 7th operands give the execution count that each of the respective prior target functions was called.

1.1.9 Module Flags Metadata

Information about the module as a whole is difficult to convey to LLVM's subsystems. The LLVM IR isn't sufficient to transmit this information. The `llvm.module.flags` named metadata exists in order to facilitate this. These flags are in the form of key / value pairs — much like a dictionary — making it easy for any subsystem who cares about a flag to look it up.

The `llvm.module.flags` metadata contains a list of metadata triplets. Each triplet has the following form:

- The first element is a *behavior* flag, which specifies the behavior when two (or more) modules are merged together, and it encounters two (or more) metadata with the same ID. The supported behaviors are described below.
- The second element is a metadata string that is a unique ID for the metadata. Each module may only have one flag entry for each unique ID (not including entries with the **Require** behavior).
- The third element is the value of the flag.

When two (or more) modules are merged together, the resulting `llvm.module.flags` metadata is the union of the modules' flags. That is, for each unique metadata ID string, there will be exactly one entry in the merged modules' `llvm.module.flags` metadata table, and the value for that entry will be determined by the merge behavior flag, as described below. The only exception is that entries with the *Require* behavior are always preserved.

The following behaviors are supported:

Value	Behavior
1	Error Emits an error if two values disagree, otherwise the resulting value is that of the operands.
2	Warning Emits a warning if two values disagree. The result value will be the operand for the flag from the first module being linked.
3	Require Adds a requirement that another module flag be present and have a specified value after linking is performed. The value must be a metadata pair, where the first element of the pair is the ID of the module flag to be restricted, and the second element of the pair is the value the module flag should be restricted to. This behavior can be used to restrict the allowable results (via triggering of an error) of linking IDs with the Override behavior.
4	Override Uses the specified value, regardless of the behavior or value of the other module. If both modules specify Override , but the values differ, an error will be emitted.
5	Append Appends the two values, which are required to be metadata nodes.
6	AppendUnique Appends the two values, which are required to be metadata nodes. However, duplicate entries in the second list are dropped during the append operation.
7	Max Takes the max of the two values, which are required to be integers.

It is an error for a particular unique flag ID to have multiple behaviors, except in the case of **Require** (which adds restrictions on another metadata value) or **Override**.

An example of module flags:

```
!0 = !{ i32 1, !"foo", i32 1 }
!1 = !{ i32 4, !"bar", i32 37 }
!2 = !{ i32 2, !"qux", i32 42 }
!3 = !{ i32 3, !"qux",
  !{
    !"foo", i32 1
  }
}
!llvm.module.flags = !{ !0, !1, !2, !3 }
```

- Metadata !0 has the ID !"foo" and the value '1'. The behavior if two or more !"foo" flags are seen is to emit an error if their values are not equal.
- Metadata !1 has the ID !"bar" and the value '37'. The behavior if two or more !"bar" flags are seen is to use the value '37'.
- Metadata !2 has the ID !"qux" and the value '42'. The behavior if two or more !"qux" flags are seen is to emit a warning if their values are not equal.
- Metadata !3 has the ID !"qux" and the value:

```
!{ !"foo", i32 1 }
```

The behavior is to emit an error if the `llvm.module.flags` does not contain a flag with the ID !"foo" that has the value '1' after linking is performed.

Objective-C Garbage Collection Module Flags Metadata

On the Mach-O platform, Objective-C stores metadata about garbage collection in a special section called "image info". The metadata consists of a version number and a bitmask specifying what types of garbage collection are supported (if any) by the file. If two or more modules are linked together their garbage collection metadata needs to be merged rather than appended together.

The Objective-C garbage collection module flags metadata consists of the following key-value pairs:

Key	Value
Objective-C Version	[Required] — The Objective-C ABI version. Valid values are 1 and 2.
Objective-C Image Info Version	[Required] — The version of the image info section. Currently always 0.
Objective-C Image Info Section	[Required] — The section to place the metadata. Valid values are " <code>__OBJC</code> ", " <code>__image_info</code> ", " <code>regular</code> " for Objective-C ABI version 1, and " <code>__DATA,__objc_imageinfo</code> ", " <code>regular</code> ", " <code>no_dead_strip</code> " for Objective-C ABI version 2.
Objective-C Garbage Collection	[Required] — Specifies whether garbage collection is supported or not. Valid values are 0, for no garbage collection, and 2, for garbage collection supported.
Objective-C GC Only	[Optional] — Specifies that only garbage collection is supported. If present, its value must be 6. This flag requires that the Objective-C Garbage Collection flag have the value 2.

Some important flag interactions:

- If a module with Objective-C Garbage Collection set to 0 is merged with a module with Objective-C Garbage Collection set to 2, then the resulting module has the Objective-C Garbage Collection flag set to 0.

- A module with Objective-C Garbage Collection set to 0 cannot be merged with a module with Objective-C GC Only set to 6.

C type width Module Flags Metadata

The ARM backend emits a section into each generated object file describing the options that it was compiled with (in a compiler-independent way) to prevent linking incompatible objects, and to allow automatic library selection. Some of these options are not visible at the IR level, namely `wchar_t` width and enum width.

To pass this information to the backend, these options are encoded in module flags metadata, using the following key-value pairs:

Key	Value
<code>short_wchar</code>	<ul style="list-style-type: none">• 0 — <code>sizeof(wchar_t) == 4</code>• 1 — <code>sizeof(wchar_t) == 2</code>
<code>short_enum</code>	<ul style="list-style-type: none">• 0 — Enums are at least as large as an <code>int</code>.• 1 — Enums are stored in the smallest integer type which can represent all of its values.

For example, the following metadata section specifies that the module was compiled with a `wchar_t` width of 4 bytes, and the underlying type of an enum is the smallest type which can represent all of its values:

```
!llvm.module.flags = !{!0, !1}
!0 = !{i32 1, !"short_wchar", i32 1}
!1 = !{i32 1, !"short_enum", i32 0}
```

1.1.10 Automatic Linker Flags Named Metadata

Some targets support embedding flags to the linker inside individual object files. Typically this is used in conjunction with language extensions which allow source files to explicitly declare the libraries they depend on, and have these automatically be transmitted to the linker via object files.

These flags are encoded in the IR using named metadata with the name `!llvm.linker.options`. Each operand is expected to be a metadata node which should be a list of other metadata nodes, each of which should be a list of metadata strings defining linker options.

For example, the following metadata section specifies two separate sets of linker options, presumably to link against `libz` and the Cocoa framework:

```
!0 = !{ !" -lz" },
!1 = !{ !" -framework", !"Cocoa" } } }
!llvm.linker.options = !{ !0, !1 }
```

The metadata encoding as lists of lists of options, as opposed to a collapsed list of options, is chosen so that the IR encoding can use multiple option strings to specify e.g., a single library, while still having that specifier be preserved as an atomic element that can be recognized by a target specific assembly writer or object file emitter.

Each individual option is required to be either a valid option for the target's linker, or an option that is reserved by the target specific assembly writer or object file emitter. No other aspect of these options is defined by the IR.

1.1.11 Intrinsic Global Variables

LLVM has a number of “magic” global variables that contain data that affect code generation or other IR semantics. These are documented here. All globals of this sort should have a section specified as “`llvm.metadata`”. This section and all globals that start with “`llvm.`” are reserved for use by LLVM.

The ‘`llvm.used`’ Global Variable

The `@llvm.used` global is an array which has *appending linkage*. This array contains a list of pointers to named global variables, functions and aliases which may optionally have a pointer cast formed of `bitcast` or `getelementptr`. For example, a legal use of it is:

```
@X = global i8 4
@Y = global i32 123

@llvm.used = appending global [2 x i8*] [
    i8* @X,
    i8* bitcast (i32* @Y to i8*)
], section "llvm.metadata"
```

If a symbol appears in the `@llvm.used` list, then the compiler, assembler, and linker are required to treat the symbol as if there is a reference to the symbol that it cannot see (which is why they have to be named). For example, if a variable has internal linkage and no references other than that from the `@llvm.used` list, it cannot be deleted. This is commonly used to represent references from inline asms and other things the compiler cannot “see”, and corresponds to “`attribute((used))`” in GNU C.

On some targets, the code generator must emit a directive to the assembler or object file to prevent the assembler and linker from molesting the symbol.

The ‘`llvm.compiler.used`’ Global Variable

The `@llvm.compiler.used` directive is the same as the `@llvm.used` directive, except that it only prevents the compiler from touching the symbol. On targets that support it, this allows an intelligent linker to optimize references to the symbol without being impeded as it would be by `@llvm.used`.

This is a rare construct that should only be used in rare circumstances, and should not be exposed to source languages.

The ‘`llvm.global_ctors`’ Global Variable

```
%0 = type { i32, void ()*, i8* }
@llvm.global_ctors = appending global [1 x %0] [%0 { i32 65535, void ()* @ctor, i8*_
  ↳@data }]
```

The `@llvm.global_ctors` array contains a list of constructor functions, priorities, and an optional associated global or function. The functions referenced by this array will be called in ascending order of priority (i.e. lowest first) when the module is loaded. The order of functions with the same priority is not defined.

If the third field is present, non-null, and points to a global variable or function, the initializer function will only run if the associated data from the current module is not discarded.

The ‘`llvm.global_dtors`’ Global Variable

```
%0 = type { i32, void ()*, i8* }
@llvm.global_dtors = appending global [1 x %0] [%0 { i32 65535, void ()* @dtor, i8*_
↳@data }]
```

The `@llvm.global_dtors` array contains a list of destructor functions, priorities, and an optional associated global or function. The functions referenced by this array will be called in descending order of priority (i.e. highest first) when the module is unloaded. The order of functions with the same priority is not defined.

If the third field is present, non-null, and points to a global variable or function, the destructor function will only run if the associated data from the current module is not discarded.

1.1.12 Instruction Reference

The LLVM instruction set consists of several different classifications of instructions: *terminator instructions*, *binary instructions*, *bitwise binary instructions*, *memory instructions*, and *other instructions*.

Terminator Instructions

As mentioned *previously*, every basic block in a program ends with a “Terminator” instruction, which indicates which block should be executed after the current block is finished. These terminator instructions typically yield a ‘void’ value: they produce control flow, not values (the one exception being the ‘*invoke*’ instruction).

The terminator instructions are: ‘*ret*’, ‘*br*’, ‘*switch*’, ‘*indirectbr*’, ‘*invoke*’, ‘*resume*’, ‘*catchswitch*’, ‘*catchret*’, ‘*cleanupret*’, and ‘*unreachable*’.

‘`ret`’ Instruction

Syntax:

```
ret <type> <value>      ; Return a value from a non-void function
ret void                 ; Return from void function
```

Overview:

The ‘`ret`’ instruction is used to return control flow (and optionally a value) from a function back to the caller.

There are two forms of the ‘`ret`’ instruction: one that returns a value and then causes control flow, and one that just causes control flow to occur.

Arguments:

The ‘`ret`’ instruction optionally accepts a single argument, the return value. The type of the return value must be a ‘*first class*’ type.

A function is not *well formed* if it has a non-void return type and contains a ‘`ret`’ instruction with no return value or a return value with a type that does not match its type, or if it has a void return type and contains a ‘`ret`’ instruction with a return value.

Semantics:

When the ‘ret’ instruction is executed, control flow returns back to the calling function’s context. If the caller is a “*call*” instruction, execution continues at the instruction after the call. If the caller was an “*invoke*” instruction, execution continues at the beginning of the “normal” destination block. If the instruction returns a value, that value shall set the call or invoke instruction’s return value.

Example:

```
ret i32 5           ; Return an integer value of 5
ret void           ; Return from a void function
ret { i32, i8 } { i32 4, i8 2 } ; Return a struct of values 4 and 2
```

‘br’ Instruction

Syntax:

```
br i1 <cond>, label <iftrue>, label <iffalse>
br label <dest>           ; Unconditional branch
```

Overview:

The ‘br’ instruction is used to cause control flow to transfer to a different basic block in the current function. There are two forms of this instruction, corresponding to a conditional branch and an unconditional branch.

Arguments:

The conditional branch form of the ‘br’ instruction takes a single ‘i1’ value and two ‘label’ values. The unconditional form of the ‘br’ instruction takes a single ‘label’ value as a target.

Semantics:

Upon execution of a conditional ‘br’ instruction, the ‘i1’ argument is evaluated. If the value is `true`, control flows to the ‘iftrue’ label argument. If “cond” is false, control flows to the ‘iffalse’ label argument.

Example:

```
Test:
  %cond = icmp eq i32 %a, %b
  br i1 %cond, label %IfEqual, label %IfUnequal
IfEqual:
  ret i32 1
IfUnequal:
  ret i32 0
```

‘switch’ Instruction

Syntax:

```
switch <intty> <value>, label <defaultdest> [ <intty> <val>, label <dest> ... ]
```

Overview:

The ‘switch’ instruction is used to transfer control flow to one of several different places. It is a generalization of the ‘br’ instruction, allowing a branch to occur to one of many possible destinations.

Arguments:

The ‘switch’ instruction uses three parameters: an integer comparison value ‘value’, a default ‘label’ destination, and an array of pairs of comparison value constants and ‘label’s. The table is not allowed to contain duplicate constant entries.

Semantics:

The switch instruction specifies a table of values and destinations. When the ‘switch’ instruction is executed, this table is searched for the given value. If the value is found, control flow is transferred to the corresponding destination; otherwise, control flow is transferred to the default destination.

Implementation:

Depending on properties of the target machine and the particular switch instruction, this instruction may be code generated in different ways. For example, it could be generated as a series of chained conditional branches or with a lookup table.

Example:

```
; Emulate a conditional br instruction
%Val = zext i1 %value to i32
switch i32 %Val, label %truedest [ i32 0, label %falsedest ]

; Emulate an unconditional br instruction
switch i32 0, label %dest [ ]

; Implement a jump table:
switch i32 %val, label %otherwise [ i32 0, label %onzero
                                   i32 1, label %onone
                                   i32 2, label %ontwo ]
```

‘indirectbr’ Instruction

Syntax:

```
indirectbr <someTy>* <address>, [ label <dest1>, label <dest2>, ... ]
```

Overview:

The ‘indirectbr’ instruction implements an indirect branch to a label within the current function, whose address is specified by “address”. Address must be derived from a *blockaddress* constant.

Arguments:

The ‘address’ argument is the address of the label to jump to. The rest of the arguments indicate the full set of possible destinations that the address may point to. Blocks are allowed to occur multiple times in the destination list, though this isn’t particularly useful.

This destination list is required so that dataflow analysis has an accurate understanding of the CFG.

Semantics:

Control transfers to the block specified in the address argument. All possible destination blocks must be listed in the label list, otherwise this instruction has undefined behavior. This implies that jumps to labels defined in other functions have undefined behavior as well.

Implementation:

This is typically implemented with a jump through a register.

Example:

```
indirectbr i8* %Addr, [ label %bb1, label %bb2, label %bb3 ]
```

‘invoke’ Instruction

Syntax:

```
<result> = invoke [cconv] [ret attrs] <ty>|<fnty> <fnptrval>(<function args>) [fn_
→attrs]
               [operand bundles] to label <normal label> unwind label <exception label>
```

Overview:

The ‘`invoke`’ instruction causes control to transfer to a specified function, with the possibility of control flow transfer to either the ‘`normal`’ label or the ‘`exception`’ label. If the callee function returns with the “`ret`” instruction, control flow will return to the “`normal`” label. If the callee (or any indirect callees) returns via the “*resume*” instruction or other exception handling mechanism, control is interrupted and continued at the dynamically nearest “`exception`” label.

The ‘`exception`’ label is a *landing pad* for the exception. As such, ‘`exception`’ label is required to have the “*landingpad*” instruction, which contains the information about the behavior of the program after unwinding happens, as its first non-PHI instruction. The restrictions on the “*landingpad*” instruction’s tightly couples it to the “`invoke`” instruction, so that the important information contained within the “*landingpad*” instruction can’t be lost through normal code motion.

Arguments:

This instruction requires several arguments:

1. The optional “`cconv`” marker indicates which *calling convention* the call should use. If none is specified, the call defaults to using C calling conventions.
2. The optional *Parameter Attributes* list for return values. Only ‘`zeroext`’, ‘`signext`’, and ‘`inreg`’ attributes are valid here.
3. ‘`ty`’: the type of the call instruction itself which is also the type of the return value. Functions that return no value are marked `void`.
4. ‘`fnty`’: shall be the signature of the function being invoked. The argument types must match the types implied by this signature. This type can be omitted if the function is not varargs.
5. ‘`fnptrval`’: An LLVM value containing a pointer to a function to be invoked. In most cases, this is a direct function invocation, but indirect `invoke`’s are just as possible, calling an arbitrary pointer to function value.
6. ‘`function args`’: argument list whose types match the function signature argument types and parameter attributes. All arguments must be of *first class* type. If the function signature indicates the function accepts a variable number of arguments, the extra arguments can be specified.
7. ‘`normal label`’: the label reached when the called function executes a ‘`ret`’ instruction.
8. ‘`exception label`’: the label reached when a callee returns via the *resume* instruction or other exception handling mechanism.
9. The optional *function attributes* list.
10. The optional *operand bundles* list.

Semantics:

This instruction is designed to operate as a standard ‘`call`’ instruction in most regards. The primary difference is that it establishes an association with a label, which is used by the runtime library to unwind the stack.

This instruction is used in languages with destructors to ensure that proper cleanup is performed in the case of either a `longjmp` or a thrown exception. Additionally, this is important for implementation of ‘`catch`’ clauses in high-level languages that support them.

For the purposes of the SSA form, the definition of the value returned by the ‘`invoke`’ instruction is deemed to occur on the edge from the current block to the “`normal`” label. If the callee unwinds then no return value is available.

Example:

```
%retval = invoke i32 @Test(i32 15) to label %Continue
           unwind label %TestCleanup           ; i32:retval set
%retval = invoke coldcc i32 @Testfnptr(i32 15) to label %Continue
           unwind label %TestCleanup           ; i32:retval set
```

‘resume’ Instruction**Syntax:**

```
resume <type> <value>
```

Overview:

The ‘resume’ instruction is a terminator instruction that has no successors.

Arguments:

The ‘resume’ instruction requires one argument, which must have the same type as the result of any ‘landingpad’ instruction in the same function.

Semantics:

The ‘resume’ instruction resumes propagation of an existing (in-flight) exception whose unwinding was interrupted with a *landingpad* instruction.

Example:

```
resume { i8*, i32 } %exn
```

‘catchswitch’ Instruction**Syntax:**

```
<resultval> = catchswitch within <parent> [ label <handler1>, label <handler2>, ... ]  
↳unwind to caller  
<resultval> = catchswitch within <parent> [ label <handler1>, label <handler2>, ... ]  
↳unwind label <default>
```

Overview:

The ‘catchswitch’ instruction is used by LLVM’s exception handling system to describe the set of possible catch handlers that may be executed by the *EH personality routine*.

Arguments:

The `parent` argument is the token of the funclet that contains the `catchswitch` instruction. If the `catchswitch` is not inside a funclet, this operand may be the token `none`.

The default argument is the label of another basic block beginning with either a `cleanuppad` or `catchswitch` instruction. This unwind destination must be a legal target with respect to the `parent` links, as described in the [exception handling documentation](#).

The `handlers` are a nonempty list of successor blocks that each begin with a *catchpad* instruction.

Semantics:

Executing this instruction transfers control to one of the successors in `handlers`, if appropriate, or continues to unwind via the unwind label if present.

The `catchswitch` is both a terminator and a “pad” instruction, meaning that it must be both the first non-phi instruction and last instruction in the basic block. Therefore, it must be the only non-phi instruction in the block.

Example:

```
dispatch1:
  %cs1 = catchswitch within none [label %handler0, label %handler1] unwind to caller
dispatch2:
  %cs2 = catchswitch within %parenthandler [label %handler0] unwind label %cleanup
```

‘catchret’ Instruction

Syntax:

```
catchret from <token> to label <normal>
```

Overview:

The ‘`catchret`’ instruction is a terminator instruction that has a single successor.

Arguments:

The first argument to a ‘`catchret`’ indicates which *catchpad* it exits. It must be a *catchpad*. The second argument to a ‘`catchret`’ specifies where control will transfer to next.

Semantics:

The ‘`catchret`’ instruction ends an existing (in-flight) exception whose unwinding was interrupted with a *catchpad* instruction. The *personality function* gets a chance to execute arbitrary code to, for example, destroy the active exception. Control then transfers to `normal`.

The `token` argument must be a token produced by a `catchpad` instruction. If the specified `catchpad` is not the most-recently-entered not-yet-exited funclet pad (as described in the [EH documentation](#)), the `catchret`'s behavior is undefined.

Example:

```
catchret from %catch label %continue
```

'cleanupret' Instruction

Syntax:

```
cleanupret from <value> unwind label <continue>
cleanupret from <value> unwind to caller
```

Overview:

The 'cleanupret' instruction is a terminator instruction that has an optional successor.

Arguments:

The 'cleanupret' instruction requires one argument, which indicates which `cleanuppadd` it exits, and must be a *cleanuppadd*. If the specified `cleanuppadd` is not the most-recently-entered not-yet-exited funclet pad (as described in the [EH documentation](#)), the `cleanupret`'s behavior is undefined.

The 'cleanupret' instruction also has an optional successor, `continue`, which must be the label of another basic block beginning with either a `cleanuppadd` or `catchswitch` instruction. This unwind destination must be a legal target with respect to the parent links, as described in the [exception handling documentation](#).

Semantics:

The 'cleanupret' instruction indicates to the *personality function* that one *cleanuppadd* it transferred control to has ended. It transfers control to `continue` or unwinds out of the function.

Example:

```
cleanupret from %cleanup unwind to caller
cleanupret from %cleanup unwind label %continue
```

'unreachable' Instruction

Syntax:

unreachable

Overview:

The ‘unreachable’ instruction has no defined semantics. This instruction is used to inform the optimizer that a particular portion of the code is not reachable. This can be used to indicate that the code after a no-return function cannot be reached, and other facts.

Semantics:

The ‘unreachable’ instruction has no defined semantics.

Binary Operations

Binary operators are used to do most of the computation in a program. They require two operands of the same type, execute an operation on them, and produce a single value. The operands might represent multiple data, as is the case with the *vector* data type. The result value has the same type as its operands.

There are several different binary operators:

‘add’ Instruction

Syntax:

```
<result> = add <ty> <op1>, <op2>           ; yields ty:result
<result> = add nuw <ty> <op1>, <op2>         ; yields ty:result
<result> = add nsw <ty> <op1>, <op2>         ; yields ty:result
<result> = add nuw nsw <ty> <op1>, <op2>     ; yields ty:result
```

Overview:

The ‘add’ instruction returns the sum of its two operands.

Arguments:

The two arguments to the ‘add’ instruction must be *integer* or *vector* of integer values. Both arguments must have identical types.

Semantics:

The value produced is the integer sum of the two operands.

If the sum has unsigned overflow, the result returned is the mathematical result modulo 2^n , where n is the bit width of the result.

Because LLVM integers use a two’s complement representation, this instruction is appropriate for both signed and unsigned integers.

`nuw` and `nsw` stand for “No Unsigned Wrap” and “No Signed Wrap”, respectively. If the `nuw` and/or `nsw` keywords are present, the result value of the `add` is a *poison value* if unsigned and/or signed overflow, respectively, occurs.

Example:

```
<result> = add i32 4, %var          ; yields i32:result = 4 + %var
```

‘fadd’ Instruction

Syntax:

```
<result> = fadd [fast-math flags]* <ty> <op1>, <op2>    ; yields ty:result
```

Overview:

The ‘fadd’ instruction returns the sum of its two operands.

Arguments:

The two arguments to the ‘fadd’ instruction must be *floating point* or *vector* of floating point values. Both arguments must have identical types.

Semantics:

The value produced is the floating point sum of the two operands. This instruction can also take any number of *fast-math flags*, which are optimization hints to enable otherwise unsafe floating point optimizations:

Example:

```
<result> = fadd float 4.0, %var      ; yields float:result = 4.0 + %var
```

‘sub’ Instruction

Syntax:

```
<result> = sub <ty> <op1>, <op2>      ; yields ty:result
<result> = sub nuw <ty> <op1>, <op2>   ; yields ty:result
<result> = sub nsw <ty> <op1>, <op2>   ; yields ty:result
<result> = sub nuw nsw <ty> <op1>, <op2> ; yields ty:result
```

Overview:

The ‘sub’ instruction returns the difference of its two operands.

Note that the ‘sub’ instruction is used to represent the ‘neg’ instruction present in most other intermediate representations.

Arguments:

The two arguments to the ‘sub’ instruction must be *integer* or *vector* of integer values. Both arguments must have identical types.

Semantics:

The value produced is the integer difference of the two operands.

If the difference has unsigned overflow, the result returned is the mathematical result modulo 2^n , where n is the bit width of the result.

Because LLVM integers use a two’s complement representation, this instruction is appropriate for both signed and unsigned integers.

`nuw` and `nsw` stand for “No Unsigned Wrap” and “No Signed Wrap”, respectively. If the `nuw` and/or `nsw` keywords are present, the result value of the `sub` is a *poison value* if unsigned and/or signed overflow, respectively, occurs.

Example:

```
<result> = sub i32 4, %var      ; yields i32:result = 4 - %var
<result> = sub i32 0, %val      ; yields i32:result = -%var
```

‘fsub’ Instruction

Syntax:

```
<result> = fsub [fast-math flags]* <ty> <op1>, <op2> ; yields ty:result
```

Overview:

The ‘fsub’ instruction returns the difference of its two operands.

Note that the ‘fsub’ instruction is used to represent the ‘fneg’ instruction present in most other intermediate representations.

Arguments:

The two arguments to the ‘fsub’ instruction must be *floating point* or *vector* of floating point values. Both arguments must have identical types.

Semantics:

The value produced is the floating point difference of the two operands. This instruction can also take any number of *fast-math flags*, which are optimization hints to enable otherwise unsafe floating point optimizations:

Example:

```
<result> = fsub float 4.0, %var      ; yields float:result = 4.0 - %var
<result> = fsub float -0.0, %val     ; yields float:result = -%var
```

‘mul’ Instruction**Syntax:**

```
<result> = mul <ty> <op1>, <op2>    ; yields ty:result
<result> = mul nuw <ty> <op1>, <op2> ; yields ty:result
<result> = mul nsw <ty> <op1>, <op2> ; yields ty:result
<result> = mul nuw nsw <ty> <op1>, <op2> ; yields ty:result
```

Overview:

The ‘mul’ instruction returns the product of its two operands.

Arguments:

The two arguments to the ‘mul’ instruction must be *integer* or *vector* of integer values. Both arguments must have identical types.

Semantics:

The value produced is the integer product of the two operands.

If the result of the multiplication has unsigned overflow, the result returned is the mathematical result modulo 2^n , where n is the bit width of the result.

Because LLVM integers use a two’s complement representation, and the result is the same width as the operands, this instruction returns the correct result for both signed and unsigned integers. If a full product (e.g. `i32 * i32 -> i64`) is needed, the operands should be sign-extended or zero-extended as appropriate to the width of the full product.

`nuw` and `nsw` stand for “No Unsigned Wrap” and “No Signed Wrap”, respectively. If the `nuw` and/or `nsw` keywords are present, the result value of the `mul` is a *poison value* if unsigned and/or signed overflow, respectively, occurs.

Example:

```
<result> = mul i32 4, %var           ; yields i32:result = 4 * %var
```

‘fmul’ Instruction

Syntax:

```
<result> = fmul [fast-math flags]* <ty> <op1>, <op2> ; yields ty:result
```

Overview:

The ‘fmul’ instruction returns the product of its two operands.

Arguments:

The two arguments to the ‘fmul’ instruction must be *floating point* or *vector* of floating point values. Both arguments must have identical types.

Semantics:

The value produced is the floating point product of the two operands. This instruction can also take any number of *fast-math flags*, which are optimization hints to enable otherwise unsafe floating point optimizations:

Example:

```
<result> = fmul float 4.0, %var ; yields float:result = 4.0 * %var
```

‘udiv’ Instruction

Syntax:

```
<result> = udiv <ty> <op1>, <op2> ; yields ty:result  
<result> = udiv exact <ty> <op1>, <op2> ; yields ty:result
```

Overview:

The ‘udiv’ instruction returns the quotient of its two operands.

Arguments:

The two arguments to the ‘udiv’ instruction must be *integer* or *vector* of integer values. Both arguments must have identical types.

Semantics:

The value produced is the unsigned integer quotient of the two operands.

Note that unsigned integer division and signed integer division are distinct operations; for signed integer division, use ‘sdiv’.

Division by zero is undefined behavior. For vectors, if any element of the divisor is zero, the operation has undefined behavior.

If the `exact` keyword is present, the result value of the `udiv` is a *poison value* if `%op1` is not a multiple of `%op2` (as such, “((a udiv exact b) mul b) == a”).

Example:

```
<result> = udiv i32 4, %var          ; yields i32:result = 4 / %var
```

‘sdiv’ Instruction**Syntax:**

```
<result> = sdiv <ty> <op1>, <op2>      ; yields ty:result
<result> = sdiv exact <ty> <op1>, <op2> ; yields ty:result
```

Overview:

The ‘sdiv’ instruction returns the quotient of its two operands.

Arguments:

The two arguments to the ‘sdiv’ instruction must be *integer* or *vector* of integer values. Both arguments must have identical types.

Semantics:

The value produced is the signed integer quotient of the two operands rounded towards zero.

Note that signed integer division and unsigned integer division are distinct operations; for unsigned integer division, use ‘udiv’.

Division by zero is undefined behavior. For vectors, if any element of the divisor is zero, the operation has undefined behavior. Overflow also leads to undefined behavior; this is a rare case, but can occur, for example, by doing a 32-bit division of -2147483648 by -1.

If the `exact` keyword is present, the result value of the `sdiv` is a *poison value* if the result would be rounded.

Example:

```
<result> = sdiv i32 4, %var          ; yields i32:result = 4 / %var
```

‘fdiv’ Instruction**Syntax:**

```
<result> = fdiv [fast-math flags]* <ty> <op1>, <op2>    ; yields ty:result
```

Overview:

The ‘fdiv’ instruction returns the quotient of its two operands.

Arguments:

The two arguments to the ‘fdiv’ instruction must be *floating point* or *vector* of floating point values. Both arguments must have identical types.

Semantics:

The value produced is the floating point quotient of the two operands. This instruction can also take any number of *fast-math flags*, which are optimization hints to enable otherwise unsafe floating point optimizations:

Example:

```
<result> = fdiv float 4.0, %var          ; yields float:result = 4.0 / %var
```

‘urem’ Instruction**Syntax:**

```
<result> = urem <ty> <op1>, <op2>      ; yields ty:result
```

Overview:

The ‘urem’ instruction returns the remainder from the unsigned division of its two arguments.

Arguments:

The two arguments to the ‘urem’ instruction must be *integer* or *vector* of integer values. Both arguments must have identical types.

Semantics:

This instruction returns the unsigned integer *remainder* of a division. This instruction always performs an unsigned division to get the remainder.

Note that unsigned integer remainder and signed integer remainder are distinct operations; for signed integer remainder, use ‘srem’.

Taking the remainder of a division by zero is undefined behavior. For vectors, if any element of the divisor is zero, the operation has undefined behavior.

Example:

```
<result> = urem i32 4, %var          ; yields i32:result = 4 % %var
```

‘srem’ Instruction**Syntax:**

```
<result> = srem <ty> <op1>, <op2>  ; yields ty:result
```

Overview:

The ‘srem’ instruction returns the remainder from the signed division of its two operands. This instruction can also take *vector* versions of the values in which case the elements must be integers.

Arguments:

The two arguments to the ‘srem’ instruction must be *integer* or *vector* of integer values. Both arguments must have identical types.

Semantics:

This instruction returns the *remainder* of a division (where the result is either zero or has the same sign as the dividend, op1), not the *modulo* operator (where the result is either zero or has the same sign as the divisor, op2) of a value. For more information about the difference, see [The Math Forum](#). For a table of how this is implemented in various languages, please see [Wikipedia: modulo operation](#).

Note that signed integer remainder and unsigned integer remainder are distinct operations; for unsigned integer remainder, use ‘urem’.

Taking the remainder of a division by zero is undefined behavior. For vectors, if any element of the divisor is zero, the operation has undefined behavior. Overflow also leads to undefined behavior; this is a rare case, but can occur, for example, by taking the remainder of a 32-bit division of -2147483648 by -1. (The remainder doesn’t actually overflow, but this rule lets srem be implemented using instructions that return both the result of the division and the remainder.)

Example:

```
<result> = srem i32 4, %var          ; yields i32:result = 4 % %var
```

‘frem’ Instruction**Syntax:**

```
<result> = frem [fast-math flags]* <ty> <op1>, <op2>    ; yields ty:result
```

Overview:

The ‘frem’ instruction returns the remainder from the division of its two operands.

Arguments:

The two arguments to the ‘frem’ instruction must be *floating point* or *vector* of floating point values. Both arguments must have identical types.

Semantics:

This instruction returns the *remainder* of a division. The remainder has the same sign as the dividend. This instruction can also take any number of *fast-math flags*, which are optimization hints to enable otherwise unsafe floating point optimizations:

Example:

```
<result> = frem float 4.0, %var      ; yields float:result = 4.0 % %var
```

Bitwise Binary Operations

Bitwise binary operators are used to do various forms of bit-twiddling in a program. They are generally very efficient instructions and can commonly be strength reduced from other instructions. They require two operands of the same type, execute an operation on them, and produce a single value. The resulting value is the same type as its operands.

‘shl’ Instruction**Syntax:**

```
<result> = shl <ty> <op1>, <op2>          ; yields ty:result  
<result> = shl nuw <ty> <op1>, <op2>      ; yields ty:result  
<result> = shl nsw <ty> <op1>, <op2>      ; yields ty:result  
<result> = shl nuw nsw <ty> <op1>, <op2>  ; yields ty:result
```

Overview:

The ‘shl’ instruction returns the first operand shifted to the left a specified number of bits.

Arguments:

Both arguments to the ‘shl’ instruction must be the same *integer* or *vector* of integer type. ‘op2’ is treated as an unsigned value.

Semantics:

The value produced is $op1 * 2^{op2} \bmod 2^n$, where n is the width of the result. If $op2$ is (statically or dynamically) equal to or larger than the number of bits in $op1$, this instruction returns a *poison value*. If the arguments are vectors, each vector element of $op1$ is shifted by the corresponding shift amount in $op2$.

If the `nsw` keyword is present, then the shift produces a poison value if it shifts out any non-zero bits. If the `nsw` keyword is present, then the shift produces a poison value if it shifts out any bits that disagree with the resultant sign bit.

Example:

```

<result> = shl i32 4, %var      ; yields i32: 4 << %var
<result> = shl i32 4, 2         ; yields i32: 16
<result> = shl i32 1, 10        ; yields i32: 1024
<result> = shl i32 1, 32        ; undefined
<result> = shl <2 x i32> <i32 1, i32 1>, <i32 1, i32 2> ; yields: result=<2 x i32>
↳ <i32 2, i32 4>

```

‘lshr’ Instruction

Syntax:

```

<result> = lshr <ty> <op1>, <op2>      ; yields ty:result
<result> = lshr exact <ty> <op1>, <op2> ; yields ty:result

```

Overview:

The ‘lshr’ instruction (logical shift right) returns the first operand shifted to the right a specified number of bits with zero fill.

Arguments:

Both arguments to the ‘lshr’ instruction must be the same *integer* or *vector* of integer type. ‘op2’ is treated as an unsigned value.

Semantics:

This instruction always performs a logical shift right operation. The most significant bits of the result will be filled with zero bits after the shift. If `op2` is (statically or dynamically) equal to or larger than the number of bits in `op1`, this instruction returns a *poison value*. If the arguments are vectors, each vector element of `op1` is shifted by the corresponding shift amount in `op2`.

If the `exact` keyword is present, the result value of the `lshr` is a poison value if any of the bits shifted out are non-zero.

Example:

```
<result> = lshr i32 4, 1    ; yields i32:result = 2
<result> = lshr i32 4, 2    ; yields i32:result = 1
<result> = lshr i8  4, 3    ; yields i8:result = 0
<result> = lshr i8 -2, 1    ; yields i8:result = 0x7F
<result> = lshr i32 1, 32   ; undefined
<result> = lshr <2 x i32> <i32 -2, i32 4>, <i32 1, i32 2> ; yields: result=<2 x i32> <i32 0x7FFFFFFF, i32 1>
```

‘ashr’ Instruction

Syntax:

```
<result> = ashr <ty> <op1>, <op2>      ; yields ty:result
<result> = ashr exact <ty> <op1>, <op2> ; yields ty:result
```

Overview:

The ‘`ashr`’ instruction (arithmetic shift right) returns the first operand shifted to the right a specified number of bits with sign extension.

Arguments:

Both arguments to the ‘`ashr`’ instruction must be the same *integer* or *vector* of integer type. ‘`op2`’ is treated as an unsigned value.

Semantics:

This instruction always performs an arithmetic shift right operation. The most significant bits of the result will be filled with the sign bit of `op1`. If `op2` is (statically or dynamically) equal to or larger than the number of bits in `op1`, this instruction returns a *poison value*. If the arguments are vectors, each vector element of `op1` is shifted by the corresponding shift amount in `op2`.

If the `exact` keyword is present, the result value of the `ashr` is a poison value if any of the bits shifted out are non-zero.

Example:

```

<result> = ashr i32 4, 1    ; yields i32:result = 2
<result> = ashr i32 4, 2    ; yields i32:result = 1
<result> = ashr i8  4, 3    ; yields i8:result = 0
<result> = ashr i8 -2, 1    ; yields i8:result = -1
<result> = ashr i32 1, 32   ; undefined
<result> = ashr <2 x i32> < i32 -2, i32 4>, < i32 1, i32 3> ; yields: result=<2 x i32> < i32 -1, i32 0>

```

‘and’ Instruction**Syntax:**

```

<result> = and <ty> <op1>, <op2>    ; yields ty:result

```

Overview:

The ‘and’ instruction returns the bitwise logical and of its two operands.

Arguments:

The two arguments to the ‘and’ instruction must be *integer* or *vector* of integer values. Both arguments must have identical types.

Semantics:

The truth table used for the ‘and’ instruction is:

In0	In1	Out
0	0	0
0	1	0
1	0	0
1	1	1

Example:

```

<result> = and i32 4, %var    ; yields i32:result = 4 & %var
<result> = and i32 15, 40     ; yields i32:result = 8
<result> = and i32 4, 8       ; yields i32:result = 0

```

‘or’ Instruction

Syntax:

```
<result> = or <ty> <op1>, <op2> ; yields ty:result
```

Overview:

The ‘or’ instruction returns the bitwise logical inclusive or of its two operands.

Arguments:

The two arguments to the ‘or’ instruction must be *integer* or *vector* of integer values. Both arguments must have identical types.

Semantics:

The truth table used for the ‘or’ instruction is:

In0	In1	Out
0	0	0
0	1	1
1	0	1
1	1	1

Example:

```
<result> = or i32 4, %var ; yields i32:result = 4 | %var
<result> = or i32 15, 40 ; yields i32:result = 47
<result> = or i32 4, 8 ; yields i32:result = 12
```

‘xor’ Instruction

Syntax:

```
<result> = xor <ty> <op1>, <op2> ; yields ty:result
```

Overview:

The ‘xor’ instruction returns the bitwise logical exclusive or of its two operands. The xor is used to implement the “one’s complement” operation, which is the “~” operator in C.

Arguments:

The two arguments to the ‘xor’ instruction must be *integer* or *vector* of integer values. Both arguments must have identical types.

Semantics:

The truth table used for the ‘xor’ instruction is:

In0	In1	Out
0	0	0
0	1	1
1	0	1
1	1	0

Example:

```
<result> = xor i32 4, %var      ; yields i32:result = 4 ^ %var
<result> = xor i32 15, 40       ; yields i32:result = 39
<result> = xor i32 4, 8         ; yields i32:result = 12
<result> = xor i32 %V, -1       ; yields i32:result = ~%V
```

Vector Operations

LLVM supports several instructions to represent vector operations in a target-independent manner. These instructions cover the element-access and vector-specific operations needed to process vectors effectively. While LLVM does directly support these vector operations, many sophisticated algorithms will want to use target-specific intrinsics to take full advantage of a specific target.

‘extractelement’ Instruction**Syntax:**

```
<result> = extractelement <n x <ty>> <val>, <ty2> <idx> ; yields <ty>
```

Overview:

The ‘extractelement’ instruction extracts a single scalar element from a vector at a specified index.

Arguments:

The first operand of an ‘extractelement’ instruction is a value of *vector* type. The second operand is an index indicating the position from which to extract the element. The index may be a variable of any integer type.

Semantics:

The result is a scalar of the same type as the element type of `val`. Its value is the value at position `idx` of `val`. If `idx` exceeds the length of `val`, the results are undefined.

Example:

```
<result> = extractelement <4 x i32> %vec, i32 0    ; yields i32
```

‘insertelement’ Instruction**Syntax:**

```
<result> = insertelement <n x <ty>> <val>, <ty> <elt>, <ty2> <idx>    ; yields <n x  
↪<ty>>
```

Overview:

The ‘insertelement’ instruction inserts a scalar element into a vector at a specified index.

Arguments:

The first operand of an ‘insertelement’ instruction is a value of *vector* type. The second operand is a scalar value whose type must equal the element type of the first operand. The third operand is an index indicating the position at which to insert the value. The index may be a variable of any integer type.

Semantics:

The result is a vector of the same type as `val`. Its element values are those of `val` except at position `idx`, where it gets the value `elt`. If `idx` exceeds the length of `val`, the results are undefined.

Example:

```
<result> = insertelement <4 x i32> %vec, i32 1, i32 0    ; yields <4 x i32>
```

‘shufflevector’ Instruction**Syntax:**

```
<result> = shufflevector <n x <ty>> <v1>, <n x <ty>> <v2>, <m x i32> <mask>    ;  
↪yields <m x <ty>>
```


Overview:

The ‘`shufflevector`’ instruction constructs a permutation of elements from two input vectors, returning a vector with the same element type as the input and length that is the same as the shuffle mask.

Arguments:

The first two operands of a ‘`shufflevector`’ instruction are vectors with the same type. The third argument is a shuffle mask whose element type is always ‘`i32`’. The result of the instruction is a vector whose length is the same as the shuffle mask and whose element type is the same as the element type of the first two operands.

The shuffle mask operand is required to be a constant vector with either constant integer or undef values.

Semantics:

The elements of the two input vectors are numbered from left to right across both of the vectors. The shuffle mask operand specifies, for each element of the result vector, which element of the two input vectors the result element gets. If the shuffle mask is undef, the result vector is undef. If any element of the mask operand is undef, that element of the result is undef. If the shuffle mask selects an undef element from one of the input vectors, the resulting element is undef.

Example:

```
<result> = shufflevector <4 x i32> %v1, <4 x i32> %v2,
                        <4 x i32> <i32 0, i32 4, i32 1, i32 5> ; yields <4 x i32>
<result> = shufflevector <4 x i32> %v1, <4 x i32> undef,
                        <4 x i32> <i32 0, i32 1, i32 2, i32 3> ; yields <4 x i32> -
↳ Identity shuffle.
<result> = shufflevector <8 x i32> %v1, <8 x i32> undef,
                        <4 x i32> <i32 0, i32 1, i32 2, i32 3> ; yields <4 x i32>
<result> = shufflevector <4 x i32> %v1, <4 x i32> %v2,
                        <8 x i32> <i32 0, i32 1, i32 2, i32 3, i32 4, i32 5, i32 6,
↳ i32 7 > ; yields <8 x i32>
```

Aggregate Operations

LLVM supports several instructions for working with *aggregate* values.

‘`extractvalue`’ Instruction

Syntax:

```
<result> = extractvalue <aggregate type> <val>, <idx>{, <idx>}*
```

Overview:

The ‘`extractvalue`’ instruction extracts the value of a member field from an *aggregate* value.

Arguments:

The first operand of an ‘extractvalue’ instruction is a value of *struct* or *array* type. The other operands are constant indices to specify which value to extract in a similar manner as indices in a ‘getelementptr’ instruction.

The major differences to getelementptr indexing are:

- Since the value being indexed is not a pointer, the first index is omitted and assumed to be zero.
- At least one index must be specified.
- Not only struct indices but also array indices must be in bounds.

Semantics:

The result is the value at the position in the aggregate specified by the index operands.

Example:

```
<result> = extractvalue {i32, float} %agg, 0      ; yields i32
```

‘insertvalue’ Instruction

Syntax:

```
<result> = insertvalue <aggregate type> <val>, <ty> <elt>, <idx>{, <idx>}*      ;  
→yields <aggregate type>
```

Overview:

The ‘insertvalue’ instruction inserts a value into a member field in an *aggregate* value.

Arguments:

The first operand of an ‘insertvalue’ instruction is a value of *struct* or *array* type. The second operand is a first-class value to insert. The following operands are constant indices indicating the position at which to insert the value in a similar manner as indices in a ‘extractvalue’ instruction. The value to insert must have the same type as the value identified by the indices.

Semantics:

The result is an aggregate of the same type as *val*. Its value is that of *val* except that the value at the position specified by the indices is that of *elt*.

Example:

```
%agg1 = insertvalue {i32, float} undef, i32 1, 0           ; yields {i32 1, float_
↳undef}
%agg2 = insertvalue {i32, float} %agg1, float %val, 1       ; yields {i32 1, float
↳%val}
%agg3 = insertvalue {i32, {float}} undef, float %val, 1, 0  ; yields {i32 undef,
↳{float %val}}
```

Memory Access and Addressing Operations

A key design point of an SSA-based representation is how it represents memory. In LLVM, no memory locations are in SSA form, which makes things very simple. This section describes how to read, write, and allocate memory in LLVM.

‘alloca’ Instruction**Syntax:**

```
<result> = alloca [inalloca] <type> [, <ty> <NumElements>] [, align <alignment>] [,
↳addrspace(<num>)] ; yields type addrspace(num)*:result
```

Overview:

The ‘alloca’ instruction allocates memory on the stack frame of the currently executing function, to be automatically released when this function returns to its caller. The object is always allocated in the address space for allocas indicated in the datalayout.

Arguments:

The ‘alloca’ instruction allocates `sizeof(<type>) * NumElements` bytes of memory on the runtime stack, returning a pointer of the appropriate type to the program. If “NumElements” is specified, it is the number of elements allocated, otherwise “NumElements” is defaulted to be one. If a constant alignment is specified, the value result of the allocation is guaranteed to be aligned to at least that boundary. The alignment may not be greater than `1 << 29`. If not specified, or if zero, the target can choose to align the allocation on any convenient boundary compatible with the type.

‘type’ may be any sized type.

Semantics:

Memory is allocated; a pointer is returned. The operation is undefined if there is insufficient stack space for the allocation. ‘alloca’d memory is automatically released when the function returns. The ‘alloca’ instruction is commonly used to represent automatic variables that must have an address available. When the function returns (either with the `ret` or `resume` instructions), the memory is reclaimed. Allocating zero bytes is legal, but the result is undefined. The order in which memory is allocated (ie., which way the stack grows) is not specified.

Example:

```
%ptr = alloca i32                                ; yields i32*:ptr
%ptr = alloca i32, i32 4                        ; yields i32*:ptr
%ptr = alloca i32, i32 4, align 1024            ; yields i32*:ptr
%ptr = alloca i32, align 1024                   ; yields i32*:ptr
```

‘load’ Instruction**Syntax:**

```
<result> = load [volatile] <ty>, <ty>* <pointer>[, align <alignment>][, !nontemporal !
↪<index>][, !invariant.load !<index>][, !invariant.group !<index>][, !nonnull !
↪<index>][, !dereferenceable !<deref_bytes_node>][, !dereferenceable_or_null !<deref_
↪bytes_node>][, !align !<align_node>]
<result> = load atomic [volatile] <ty>, <ty>* <pointer> [syncscope("<target-scope>")]
↪<ordering>, align <alignment> [, !invariant.group !<index>]
!<index> = !{ i32 1 }
!<deref_bytes_node> = !{i64 <dereferenceable_bytes>}
!<align_node> = !{ i64 <value_alignment> }
```

Overview:

The ‘load’ instruction is used to read from memory.

Arguments:

The argument to the `load` instruction specifies the memory address from which to load. The type specified must be a *first class* type of known size (i.e. not containing an *opaque structural type*). If the load is marked as *volatile*, then the optimizer is not allowed to modify the number or order of execution of this load with other *volatile operations*.

If the load is marked as *atomic*, it takes an extra *ordering* and optional `syncscope("<target-scope>")` argument. The `release` and `acq_rel` orderings are not valid on load instructions. Atomic loads produce *defined* results when they may see multiple atomic stores. The type of the pointee must be an integer, pointer, or floating-point type whose bit width is a power of two greater than or equal to eight and less than or equal to a target-specific size limit. `align` must be explicitly specified on atomic loads, and the load has undefined behavior if the alignment is not set to a value which is at least the size in bytes of the pointee. `!nontemporal` does not have any defined semantics for atomic loads.

The optional constant `align` argument specifies the alignment of the operation (that is, the alignment of the memory address). A value of 0 or an omitted `align` argument means that the operation has the ABI alignment for the target. It is the responsibility of the code emitter to ensure that the alignment information is correct. Overestimating the alignment results in undefined behavior. Underestimating the alignment may produce less efficient code. An alignment of 1 is always safe. The maximum possible alignment is $1 \ll 29$. An alignment value higher than the size of the loaded type implies memory up to the alignment value bytes can be safely loaded without trapping in the default address space. Access of the high bytes can interfere with debugging tools, so should not be accessed if the function has the `sanitize_thread` or `sanitize_address` attributes.

The optional `!nontemporal` metadata must reference a single metadata name `<index>` corresponding to a metadata node with one `i32` entry of value 1. The existence of the `!nontemporal` metadata on the instruction tells the

optimizer and code generator that this load is not expected to be reused in the cache. The code generator may select special instructions to save cache bandwidth, such as the MOVNT instruction on x86.

The optional `!invariant.load` metadata must reference a single metadata name `<index>` corresponding to a metadata node with no entries. If a load instruction tagged with the `!invariant.load` metadata is executed, the optimizer may assume the memory location referenced by the load contains the same value at all points in the program where the memory location is known to be dereferenceable.

The optional `!invariant.group` metadata must reference a single metadata name `<index>` corresponding to a metadata node. See `invariant.group` metadata.

The optional `!nonnull` metadata must reference a single metadata name `<index>` corresponding to a metadata node with no entries. The existence of the `!nonnull` metadata on the instruction tells the optimizer that the value loaded is known to never be null. This is analogous to the `nonnull` attribute on parameters and return values. This metadata can only be applied to loads of a pointer type.

The optional `!dereferenceable` metadata must reference a single metadata name `<deref_bytes_node>` corresponding to a metadata node with one `i64` entry. The existence of the `!dereferenceable` metadata on the instruction tells the optimizer that the value loaded is known to be dereferenceable. The number of bytes known to be dereferenceable is specified by the integer value in the metadata node. This is analogous to the “dereferenceable” attribute on parameters and return values. This metadata can only be applied to loads of a pointer type.

The optional `!dereferenceable_or_null` metadata must reference a single metadata name `<deref_bytes_node>` corresponding to a metadata node with one `i64` entry. The existence of the `!dereferenceable_or_null` metadata on the instruction tells the optimizer that the value loaded is known to be either dereferenceable or null. The number of bytes known to be dereferenceable is specified by the integer value in the metadata node. This is analogous to the “dereferenceable_or_null” attribute on parameters and return values. This metadata can only be applied to loads of a pointer type.

The optional `!align` metadata must reference a single metadata name `<align_node>` corresponding to a metadata node with one `i64` entry. The existence of the `!align` metadata on the instruction tells the optimizer that the value loaded is known to be aligned to a boundary specified by the integer value in the metadata node. The alignment must be a power of 2. This is analogous to the “align” attribute on parameters and return values. This metadata can only be applied to loads of a pointer type.

Semantics:

The location of memory pointed to is loaded. If the value being loaded is of scalar type then the number of bytes read does not exceed the minimum number of bytes needed to hold all bits of the type. For example, loading an `i24` reads at most three bytes. When loading a value of a type like `i20` with a size that is not an integral number of bytes, the result is undefined if the value was not originally written using a store of the same type.

Examples:

```
%ptr = alloca i32           ; yields i32*:ptr
store i32 3, i32* %ptr      ; yields void
%val = load i32, i32* %ptr   ; yields i32:val = i32 3
```

‘store’ Instruction

Syntax:

```
store [volatile] <ty> <value>, <ty>* <pointer>[, align <alignment>][, !nontemporal !
↪<index>][, !invariant.group !<index>] ; yields void
store atomic [volatile] <ty> <value>, <ty>* <pointer> [syncscope("<target-scope>")]
↪<ordering>, align <alignment> [, !invariant.group !<index>] ; yields void
```

Overview:

The ‘store’ instruction is used to write to memory.

Arguments:

There are two arguments to the `store` instruction: a value to store and an address at which to store it. The type of the `<pointer>` operand must be a pointer to the *first class* type of the `<value>` operand. If the store is marked as `volatile`, then the optimizer is not allowed to modify the number or order of execution of this store with other *volatile operations*. Only values of *first class* types of known size (i.e. not containing an *opaque structural type*) can be stored.

If the store is marked as `atomic`, it takes an extra *ordering* and optional `syncscope("<target-scope>")` argument. The `acquire` and `acq_rel` orderings aren’t valid on store instructions. Atomic loads produce *defined* results when they may see multiple atomic stores. The type of the pointee must be an integer, pointer, or floating-point type whose bit width is a power of two greater than or equal to eight and less than or equal to a target-specific size limit. `align` must be explicitly specified on atomic stores, and the store has undefined behavior if the alignment is not set to a value which is at least the size in bytes of the pointee. `!nontemporal` does not have any defined semantics for atomic stores.

The optional constant `align` argument specifies the alignment of the operation (that is, the alignment of the memory address). A value of 0 or an omitted `align` argument means that the operation has the ABI alignment for the target. It is the responsibility of the code emitter to ensure that the alignment information is correct. Overestimating the alignment results in undefined behavior. Underestimating the alignment may produce less efficient code. An alignment of 1 is always safe. The maximum possible alignment is $1 \ll 29$. An alignment value higher than the size of the stored type implies memory up to the alignment value bytes can be stored to without trapping in the default address space. Storing to the higher bytes however may result in data races if another thread can access the same address. Introducing a data race is not allowed. Storing to the extra bytes is not allowed even in situations where a data race is known to not exist if the function has the `sanitize_address` attribute.

The optional `!nontemporal` metadata must reference a single metadata name `<index>` corresponding to a metadata node with one `i32` entry of value 1. The existence of the `!nontemporal` metadata on the instruction tells the optimizer and code generator that this load is not expected to be reused in the cache. The code generator may select special instructions to save cache bandwidth, such as the `MOVNT` instruction on x86.

The optional `!invariant.group` metadata must reference a single metadata name `<index>`. See `invariant.group` metadata.

Semantics:

The contents of memory are updated to contain `<value>` at the location specified by the `<pointer>` operand. If `<value>` is of scalar type then the number of bytes written does not exceed the minimum number of bytes needed to hold all bits of the type. For example, storing an `i24` writes at most three bytes. When writing a value of a type

like `i20` with a size that is not an integral number of bytes, it is unspecified what happens to the extra bits that do not belong to the type, but they will typically be overwritten.

Example:

```
%ptr = alloca i32           ; yields i32*:ptr
store i32 3, i32* %ptr      ; yields void
%val = load i32, i32* %ptr   ; yields i32:val = i32 3
```

‘fence’ Instruction

Syntax:

```
fence [syncscope("<target-scope>")] <ordering> ; yields void
```

Overview:

The ‘fence’ instruction is used to introduce happens-before edges between operations.

Arguments:

‘fence’ instructions take an *ordering* argument which defines what *synchronizes-with* edges they add. They can only be given acquire, release, `acq_rel`, and `seq_cst` orderings.

Semantics:

A fence A which has (at least) release ordering semantics *synchronizes with* a fence B with (at least) acquire ordering semantics if and only if there exist atomic operations X and Y, both operating on some atomic object M, such that A is sequenced before X, X modifies M (either directly or through some side effect of a sequence headed by X), Y is sequenced before B, and Y observes M. This provides a *happens-before* dependency between A and B. Rather than an explicit fence, one (but not both) of the atomic operations X or Y might provide a release or acquire (resp.) ordering constraint and still *synchronize-with* the explicit fence and establish the *happens-before* edge.

A fence which has `seq_cst` ordering, in addition to having both acquire and release semantics specified above, participates in the global program order of other `seq_cst` operations and/or fences.

A fence instruction can also take an optional “*syncscope*” argument.

Example:

```
fence acquire                ; yields void
fence syncscope("singlethread") seq_cst ; yields void
fence syncscope("agent") seq_cst ; yields void
```

‘cmpxchg’ Instruction

Syntax:

```
cmpxchg [weak] [volatile] <ty>* <pointer>, <ty> <cmp>, <ty> <new> [syncscope("<target-  
↪scope>")] <success ordering> <failure ordering> ; yields { ty, i1 }
```

Overview:

The ‘cmpxchg’ instruction is used to atomically modify memory. It loads a value in memory and compares it to a given value. If they are equal, it tries to store a new value into the memory.

Arguments:

There are three arguments to the ‘cmpxchg’ instruction: an address to operate on, a value to compare to the value currently be at that address, and a new value to place at that address if the compared values are equal. The type of ‘<cmp>’ must be an integer or pointer type whose bit width is a power of two greater than or equal to eight and less than or equal to a target-specific size limit. ‘<cmp>’ and ‘<new>’ must have the same type, and the type of ‘<pointer>’ must be a pointer to that type. If the cmpxchg is marked as *volatile*, then the optimizer is not allowed to modify the number or order of execution of this cmpxchg with other *volatile operations*.

The success and failure *ordering* arguments specify how this cmpxchg synchronizes with other atomic operations. Both ordering parameters must be at least *monotonic*, the ordering constraint on failure must be no stronger than that on success, and the failure ordering cannot be either *release* or *acq_rel*.

A cmpxchg instruction can also take an optional “*syncscope*” argument.

The pointer passed into cmpxchg must have alignment greater than or equal to the size in memory of the operand.

Semantics:

The contents of memory at the location specified by the ‘<pointer>’ operand is read and compared to ‘<cmp>’; if the read value is the equal, the ‘<new>’ is written. The original value at the location is returned, together with a flag indicating success (true) or failure (false).

If the cmpxchg operation is marked as *weak* then a spurious failure is permitted: the operation may not write <new> even if the comparison matched.

If the cmpxchg operation is strong (the default), the i1 value is 1 if and only if the value loaded equals cmp.

A successful cmpxchg is a read-modify-write instruction for the purpose of identifying release sequences. A failed cmpxchg is equivalent to an atomic load with an ordering parameter determined the second ordering parameter.

Example:

```
entry:  
    %orig = load atomic i32, i32* %ptr unordered, align 4 ; yields ↪  
    ↪i32  
    br label %loop  
  
loop:  
    %cmp = phi i32 [ %orig, %entry ], [%value_loaded, %loop]
```



```

    %squared = mul i32 %cmp, %cmp
    %val_success = cmpxchg i32* %ptr, i32 %cmp, i32 %squared acq_rel monotonic ; yields_
    ↪ { i32, i1 }
    %value_loaded = extractvalue { i32, i1 } %val_success, 0
    %success = extractvalue { i32, i1 } %val_success, 1
    br i1 %success, label %done, label %loop

done:
    ...

```

‘atomicrmw’ Instruction

Syntax:

```

atomicrmw [volatile] <operation> <ty>* <pointer>, <ty> <value> [syncscope("<target-
    ↪scope>")] <ordering> ; yields ty

```

Overview:

The ‘atomicrmw’ instruction is used to atomically modify memory.

Arguments:

There are three arguments to the ‘atomicrmw’ instruction: an operation to apply, an address whose value to modify, an argument to the operation. The operation must be one of the following keywords:

- xchg
- add
- sub
- and
- nand
- or
- xor
- max
- min
- umax
- umin

The type of ‘<value>’ must be an integer type whose bit width is a power of two greater than or equal to eight and less than or equal to a target-specific size limit. The type of the ‘<pointer>’ operand must be a pointer to that type. If the atomicrmw is marked as volatile, then the optimizer is not allowed to modify the number or order of execution of this atomicrmw with other *volatile operations*.

A atomicrmw instruction can also take an optional “*syncscope*” argument.

Semantics:

The contents of memory at the location specified by the ‘<pointer>’ operand are atomically read, modified, and written back. The original value at the location is returned. The modification is specified by the operation argument:

- xchg: `*ptr = val`
- add: `*ptr = *ptr + val`
- sub: `*ptr = *ptr - val`
- and: `*ptr = *ptr & val`
- nand: `*ptr = ~(*ptr & val)`
- or: `*ptr = *ptr | val`
- xor: `*ptr = *ptr ^ val`
- max: `*ptr = *ptr > val ? *ptr : val` (using a signed comparison)
- min: `*ptr = *ptr < val ? *ptr : val` (using a signed comparison)
- umax: `*ptr = *ptr > val ? *ptr : val` (using an unsigned comparison)
- umin: `*ptr = *ptr < val ? *ptr : val` (using an unsigned comparison)

Example:

```
%old = atomicrmw add i32* %ptr, i32 1 acquire ; yields i32
```

‘getelementptr’ Instruction

Syntax:

```
<result> = getelementptr <ty>, <ty>* <ptrval>{, [inrange] <ty> <idx>}*  
<result> = getelementptr inbounds <ty>, <ty>* <ptrval>{, [inrange] <ty> <idx>}*  
<result> = getelementptr <ty>, <ptr vector> <ptrval>, [inrange] <vector index type>  
↪<idx>
```

Overview:

The ‘getelementptr’ instruction is used to get the address of a subelement of an *aggregate* data structure. It performs address calculation only and does not access memory. The instruction can also be used to calculate a vector of such addresses.

Arguments:

The first argument is always a type used as the basis for the calculations. The second argument is always a pointer or a vector of pointers, and is the base address to start from. The remaining arguments are indices that indicate which of the elements of the aggregate object are indexed. The interpretation of each index is dependent on the type being indexed into. The first index always indexes the pointer value given as the second argument, the second index indexes a value of the type pointed to (not necessarily the value directly pointed to, since the first index can be non-zero), etc. The first type indexed into must be a pointer value, subsequent types can be arrays, vectors, and structs. Note

that subsequent types being indexed into can never be pointers, since that would require loading the pointer before continuing calculation.

The type of each index argument depends on the type it is indexing into. When indexing into a (optionally packed) structure, only `i32` integer **constants** are allowed (when using a vector of indices they must all be the **same** `i32` integer constant). When indexing into an array, pointer or vector, integers of any width are allowed, and they are not required to be constant. These integers are treated as signed values where relevant.

For example, let's consider a C code fragment and how it gets compiled to LLVM:

```
struct RT {
    char A;
    int B[10][20];
    char C;
};
struct ST {
    int X;
    double Y;
    struct RT Z;
};

int *foo(struct ST *s) {
    return &s[1].Z.B[5][13];
}
```

The LLVM code generated by Clang is:

```
%struct.RT = type { i8, [10 x [20 x i32]], i8 }
%struct.ST = type { i32, double, %struct.RT }

define i32* @foo(%struct.ST* %s) nounwind uwtable readnone optsize ssp {
entry:
    %arrayidx = getelementptr inbounds %struct.ST, %struct.ST* %s, i64 1, i32 2, i32 1,
    ↪ i64 5, i64 13
    ret i32* %arrayidx
}
```

Semantics:

In the example above, the first index is indexing into the '`%struct.ST*`' type, which is a pointer, yielding a '`%struct.ST`' = '{ `i32`, `double`, `%struct.RT` }' type, a structure. The second index indexes into the third element of the structure, yielding a '`%struct.RT`' = '{ `i8`, `[10 x [20 x i32]]`, `i8` }' type, another structure. The third index indexes into the second element of the structure, yielding a '`[10 x [20 x i32]]`' type, an array. The two dimensions of the array are subscripted into, yielding an '`i32`' type. The '`getelementptr`' instruction returns a pointer to this element, thus computing a value of '`i32*`' type.

Note that it is perfectly legal to index partially through a structure, returning a pointer to an inner element. Because of this, the LLVM code for the given testcase is equivalent to:

```
define i32* @foo(%struct.ST* %s) {
    %t1 = getelementptr %struct.ST, %struct.ST* %s, i32 1 ;_
    ↪ yields %struct.ST*:%t1
    %t2 = getelementptr %struct.ST, %struct.ST* %t1, i32 0, i32 2 ;_
    ↪ yields %struct.RT*:%t2
    %t3 = getelementptr %struct.RT, %struct.RT* %t2, i32 0, i32 1 ;_
    ↪ yields [10 x [20 x i32]]*:%t3
    %t4 = getelementptr [10 x [20 x i32]], [10 x [20 x i32]]* %t3, i32 0, i32 5 ;_
    ↪ yields [20 x i32]*:%t4
```

```

    %t5 = getelementptr [20 x i32], [20 x i32]* %t4, i32 0, i32 13
    ↪ yields i32*:%t5
    ret i32* %t5
}

```

If the `inbounds` keyword is present, the result value of the `getelementptr` is a *poison value* if the base pointer is not an *in bounds* address of an allocated object, or if any of the addresses that would be formed by successive addition of the offsets implied by the indices to the base address with infinitely precise signed arithmetic are not an *in bounds* address of that allocated object. The *in bounds* addresses for an allocated object are all the addresses that point into the object, plus the address one byte past the end. The only *in bounds* address for a null pointer in the default address-space is the null pointer itself. In cases where the base is a vector of pointers the `inbounds` keyword applies to each of the computations element-wise.

If the `inbounds` keyword is not present, the offsets are added to the base address with silently-wrapping two's complement arithmetic. If the offsets have a different width from the pointer, they are sign-extended or truncated to the width of the pointer. The result value of the `getelementptr` may be outside the object pointed to by the base pointer. The result value may not necessarily be used to access memory though, even if it happens to point into allocated storage. See the [Pointer Aliasing Rules](#) section for more information.

If the `inrange` keyword is present before any index, loading from or storing to any pointer derived from the `getelementptr` has undefined behavior if the load or store would access memory outside of the bounds of the element selected by the index marked as `inrange`. The result of a pointer comparison or `ptrtoint` (including `ptrtoint`-like operations involving memory) involving a pointer derived from a `getelementptr` with the `inrange` keyword is undefined, with the exception of comparisons in the case where both operands are in the range of the element selected by the `inrange` keyword, inclusive of the address one past the end of that element. Note that the `inrange` keyword is currently only allowed in constant `getelementptr` expressions.

The `getelementptr` instruction is often confusing. For some more insight into how it works, see [the `getelementptr` FAQ](#).

Example:

```

; yields [12 x i8]*:aptr
%aptr = getelementptr {i32, [12 x i8]}, {i32, [12 x i8]}* %saptr, i64 0, i32 1
; yields i8*:vptr
%vptr = getelementptr {i32, <2 x i8>}, {i32, <2 x i8>}* %svptr, i64 0, i32 1, i32 1
; yields i8*:eptr
%eptr = getelementptr [12 x i8], [12 x i8]* %aptr, i64 0, i32 1
; yields i32*:iptr
%iptr = getelementptr [10 x i32], [10 x i32]* @arr, i16 0, i16 0

```

Vector of pointers:

The `getelementptr` returns a vector of pointers, instead of a single address, when one or more of its arguments is a vector. In such cases, all vector arguments should have the same number of elements, and every scalar argument will be effectively broadcast into a vector during address calculation.

```

; All arguments are vectors:
;   A[i] = ptrs[i] + offsets[i]*sizeof(i8)
%A = getelementptr i8, <4 x i8*> %ptrs, <4 x i64> %offsets

; Add the same scalar offset to each pointer of a vector:
;   A[i] = ptrs[i] + offset*sizeof(i8)
%A = getelementptr i8, <4 x i8*> %ptrs, i64 %offset

```

```
; Add distinct offsets to the same pointer:
;   A[i] = ptr + offsets[i]*sizeof(i8)
%A = getelementptr i8, i8* %ptr, <4 x i64> %offsets

; In all cases described above the type of the result is <4 x i8*>
```

The two following instructions are equivalent:

```
getelementptr %struct.ST, <4 x %struct.ST*> %s, <4 x i64> %ind1,
  <4 x i32> <i32 2, i32 2, i32 2, i32 2>,
  <4 x i32> <i32 1, i32 1, i32 1, i32 1>,
  <4 x i32> %ind4,
  <4 x i64> <i64 13, i64 13, i64 13, i64 13>

getelementptr %struct.ST, <4 x %struct.ST*> %s, <4 x i64> %ind1,
  i32 2, i32 1, <4 x i32> %ind4, i64 13
```

Let's look at the C code, where the vector version of `getelementptr` makes sense:

```
// Let's assume that we vectorize the following loop:
double *A, *B; int *C;
for (int i = 0; i < size; ++i) {
    A[i] = B[C[i]];
}
```

```
; get pointers for 8 elements from array B
%ptrs = getelementptr double, double* %B, <8 x i32> %C
; load 8 elements from array B into A
%A = call <8 x double> @llvm.masked.gather.v8f64.v8p0f64(<8 x double*> %ptrs,
  i32 8, <8 x i1> %mask, <8 x double> %passthru)
```

Conversion Operations

The instructions in this category are the conversion instructions (casting) which all take a single operand and a type. They perform various bit conversions on the operand.

'trunc .. to' Instruction

Syntax:

```
<result> = trunc <ty> <value> to <ty2>           ; yields ty2
```

Overview:

The 'trunc' instruction truncates its operand to the type `ty2`.

Arguments:

The 'trunc' instruction takes a value to trunc, and a type to trunc it to. Both types must be of *integer* types, or vectors of the same number of integers. The bit size of the value must be larger than the bit size of the destination type, `ty2`. Equal sized types are not allowed.

Semantics:

The ‘trunc’ instruction truncates the high order bits in `value` and converts the remaining bits to `ty2`. Since the source size must be larger than the destination size, `trunc` cannot be a *no-op cast*. It will always truncate bits.

Example:

```
%X = trunc i32 257 to i8           ; yields i8:1
%Y = trunc i32 123 to i1          ; yields i1:true
%Z = trunc i32 122 to i1          ; yields i1:false
%W = trunc <2 x i16> <i16 8, i16 7> to <2 x i8> ; yields <i8 8, i8 7>
```

‘zext .. to’ Instruction

Syntax:

```
<result> = zext <ty> <value> to <ty2>           ; yields ty2
```

Overview:

The ‘zext’ instruction zero extends its operand to type `ty2`.

Arguments:

The ‘zext’ instruction takes a value to cast, and a type to cast it to. Both types must be of *integer* types, or vectors of the same number of integers. The bit size of the `value` must be smaller than the bit size of the destination type, `ty2`.

Semantics:

The `zext` fills the high order bits of the `value` with zero bits until it reaches the size of the destination type, `ty2`.

When zero extending from `i1`, the result will always be either 0 or 1.

Example:

```
%X = zext i32 257 to i64           ; yields i64:257
%Y = zext i1 true to i32           ; yields i32:1
%Z = zext <2 x i16> <i16 8, i16 7> to <2 x i32> ; yields <i32 8, i32 7>
```

‘sext .. to’ Instruction

Syntax:

```
<result> = sext <ty> <value> to <ty2> ; yields ty2
```

Overview:

The ‘sext’ sign extends `value` to the type `ty2`.

Arguments:

The ‘sext’ instruction takes a value to cast, and a type to cast it to. Both types must be of *integer* types, or vectors of the same number of integers. The bit size of the `value` must be smaller than the bit size of the destination type, `ty2`.

Semantics:

The ‘sext’ instruction performs a sign extension by copying the sign bit (highest order bit) of the `value` until it reaches the bit size of the type `ty2`.

When sign extending from `i1`, the extension always results in -1 or 0.

Example:

```
%X = sext i8  -1 to i16 ; yields i16 :65535
%Y = sext i1 true to i32 ; yields i32:-1
%Z = sext <2 x i16> <i16 8, i16 7> to <2 x i32> ; yields <i32 8, i32 7>
```

‘fptrunc .. to’ Instruction

Syntax:

```
<result> = fptrunc <ty> <value> to <ty2> ; yields ty2
```

Overview:

The ‘fptrunc’ instruction truncates `value` to type `ty2`.

Arguments:

The ‘fptrunc’ instruction takes a *floating point* value to cast and a *floating point* type to cast it to. The size of `value` must be larger than the size of `ty2`. This implies that `fptrunc` cannot be used to make a *no-op cast*.

Semantics:

The ‘fptrunc’ instruction casts a value from a larger *floating point* type to a smaller *floating point* type. If the value cannot fit (i.e. overflows) within the destination type, `ty2`, then the results are undefined. If the cast produces an inexact result, how rounding is performed (e.g. truncation, also known as round to zero) is undefined.

Example:

```
%X = fptrunc double 123.0 to float      ; yields float:123.0
%Y = fptrunc double 1.0E+300 to float   ; yields undefined
```

‘fpext .. to’ Instruction**Syntax:**

```
<result> = fpext <ty> <value> to <ty2>      ; yields ty2
```

Overview:

The ‘fpext’ extends a floating point value to a larger floating point value.

Arguments:

The ‘fpext’ instruction takes a *floating point* value to cast, and a *floating point* type to cast it to. The source type must be smaller than the destination type.

Semantics:

The ‘fpext’ instruction extends the value from a smaller *floating point* type to a larger *floating point* type. The fpext cannot be used to make a *no-op cast* because it always changes bits. Use bitcast to make a *no-op cast* for a floating point cast.

Example:

```
%X = fpext float 3.125 to double          ; yields double:3.125000e+00
%Y = fpext double %X to fp128              ; yields_
↳ fp128:0xL00000000000000004000900000000000
```

‘fptoui .. to’ Instruction**Syntax:**

```
<result> = fptoui <ty> <value> to <ty2>      ; yields ty2
```

Overview:

The ‘fptoui’ converts a floating point value to its unsigned integer equivalent of type ty2.

Arguments:

The ‘fptoui’ instruction takes a value to cast, which must be a scalar or vector *floating point* value, and a type to cast it to ty_2 , which must be an *integer* type. If ty is a vector floating point type, ty_2 must be a vector integer type with the same number of elements as ty .

Semantics:

The ‘fptoui’ instruction converts its *floating point* operand into the nearest (rounding towards zero) unsigned integer value. If the value cannot fit in ty_2 , the results are undefined.

Example:

```
%X = fptoui double 123.0 to i32      ; yields i32:123
%Y = fptoui float 1.0E+300 to i1     ; yields undefined:1
%Z = fptoui float 1.04E+17 to i8     ; yields undefined:1
```

‘fptosi .. to’ Instruction

Syntax:

```
<result> = fptosi <ty> <value> to <ty2>          ; yields ty2
```

Overview:

The ‘fptosi’ instruction converts *floating point* value to type ty_2 .

Arguments:

The ‘fptosi’ instruction takes a value to cast, which must be a scalar or vector *floating point* value, and a type to cast it to ty_2 , which must be an *integer* type. If ty is a vector floating point type, ty_2 must be a vector integer type with the same number of elements as ty .

Semantics:

The ‘fptosi’ instruction converts its *floating point* operand into the nearest (rounding towards zero) signed integer value. If the value cannot fit in ty_2 , the results are undefined.

Example:

```
%X = fptosi double -123.0 to i32    ; yields i32:-123
%Y = fptosi float 1.0E-247 to i1    ; yields undefined:1
%Z = fptosi float 1.04E+17 to i8    ; yields undefined:1
```

'uitofp .. to' Instruction

Syntax:

```
<result> = uitofp <ty> <value> to <ty2>           ; yields ty2
```

Overview:

The 'uitofp' instruction regards `value` as an unsigned integer and converts that value to the `ty2` type.

Arguments:

The 'uitofp' instruction takes a value to cast, which must be a scalar or vector *integer* value, and a type to cast it to `ty2`, which must be an *floating point* type. If `ty` is a vector integer type, `ty2` must be a vector floating point type with the same number of elements as `ty`.

Semantics:

The 'uitofp' instruction interprets its operand as an unsigned integer quantity and converts it to the corresponding floating point value. If the value cannot fit in the floating point value, the results are undefined.

Example:

```
%X = uitofp i32 257 to float           ; yields float:257.0  
%Y = uitofp i8 -1 to double           ; yields double:255.0
```

'sitofp .. to' Instruction

Syntax:

```
<result> = sitofp <ty> <value> to <ty2>           ; yields ty2
```

Overview:

The 'sitofp' instruction regards `value` as a signed integer and converts that value to the `ty2` type.

Arguments:

The 'sitofp' instruction takes a value to cast, which must be a scalar or vector *integer* value, and a type to cast it to `ty2`, which must be an *floating point* type. If `ty` is a vector integer type, `ty2` must be a vector floating point type with the same number of elements as `ty`.

Semantics:

The ‘sitofp’ instruction interprets its operand as a signed integer quantity and converts it to the corresponding floating point value. If the value cannot fit in the floating point value, the results are undefined.

Example:

```
%X = sitofp i32 257 to float      ; yields float:257.0
%Y = sitofp i8 -1 to double      ; yields double:-1.0
```

‘ptrtoint .. to’ Instruction**Syntax:**

```
<result> = ptrtoint <ty> <value> to <ty2>      ; yields ty2
```

Overview:

The ‘ptrtoint’ instruction converts the pointer or a vector of pointers value to the integer (or vector of integers) type ty2.

Arguments:

The ‘ptrtoint’ instruction takes a value to cast, which must be a value of type *pointer* or a vector of pointers, and a type to cast it to ty2, which must be an *integer* or a vector of integers type.

Semantics:

The ‘ptrtoint’ instruction converts value to integer type ty2 by interpreting the pointer value as an integer and either truncating or zero extending that value to the size of the integer type. If value is smaller than ty2 then a zero extension is done. If value is larger than ty2 then a truncation is done. If they are the same size, then nothing is done (*no-op cast*) other than a type change.

Example:

```
%X = ptrtoint i32* %P to i8      ; yields truncation on 32-bit_
↪architecture
%Y = ptrtoint i32* %P to i64      ; yields zero extension on 32-bit_
↪architecture
%Z = ptrtoint <4 x i32*> %P to <4 x i64>; yields vector zero extension for a vector_
↪of addresses on 32-bit architecture
```

'inttoptr .. to' Instruction

Syntax:

```
<result> = inttoptr <ty> <value> to <ty2> ; yields ty2
```

Overview:

The 'inttoptr' instruction converts an integer value to a pointer type, `ty2`.

Arguments:

The 'inttoptr' instruction takes an *integer* value to cast, and a type to cast it to, which must be a *pointer* type.

Semantics:

The 'inttoptr' instruction converts `value` to type `ty2` by applying either a zero extension or a truncation depending on the size of the integer `value`. If `value` is larger than the size of a pointer then a truncation is done. If `value` is smaller than the size of a pointer then a zero extension is done. If they are the same size, nothing is done (*no-op cast*).

Example:

```
%X = inttoptr i32 255 to i32* ; yields zero extension on 64-bit architecture
%Y = inttoptr i32 255 to i32* ; yields no-op on 32-bit architecture
%Z = inttoptr i64 0 to i32* ; yields truncation on 32-bit architecture
%Z = inttoptr <4 x i32> %G to <4 x i8*>; yields truncation of vector G to four
↳pointers
```

'bitcast .. to' Instruction

Syntax:

```
<result> = bitcast <ty> <value> to <ty2> ; yields ty2
```

Overview:

The 'bitcast' instruction converts `value` to type `ty2` without changing any bits.

Arguments:

The 'bitcast' instruction takes a value to cast, which must be a non-aggregate first class value, and a type to cast it to, which must also be a non-aggregate *first class* type. The bit sizes of `value` and the destination type, `ty2`, must be identical. If the source type is a pointer, the destination type must also be a pointer of the same size. This instruction supports bitwise conversion of vectors to integers and to vectors of other types (as long as they have the same size).

Semantics:

The ‘bitcast’ instruction converts `value` to type `ty2`. It is always a *no-op cast* because no bits change with this conversion. The conversion is done as if the `value` had been stored to memory and read back as type `ty2`. Pointer (or vector of pointers) types may only be converted to other pointer (or vector of pointers) types with the same address space through this instruction. To convert pointers to other types, use the *inttoptr* or *ptrtoint* instructions first.

Example:

```
%X = bitcast i8 255 to i8          ; yields i8 :-1
%Y = bitcast i32* %x to sint*      ; yields sint*:%x
%Z = bitcast <2 x int> %V to i64;   ; yields i64: %V
%Z = bitcast <2 x i32*> %V to <2 x i64*> ; yields <2 x i64*>
```

‘addrspacecast .. to’ Instruction

Syntax:

```
<result> = addrspacecast <pty> <ptrval> to <pty2>      ; yields pty2
```

Overview:

The ‘addrspacecast’ instruction converts `ptrval` from `pty` in address space `n` to type `pty2` in address space `m`.

Arguments:

The ‘addrspacecast’ instruction takes a pointer or vector of pointer value to cast and a pointer type to cast it to, which must have a different address space.

Semantics:

The ‘addrspacecast’ instruction converts the pointer value `ptrval` to type `pty2`. It can be a *no-op cast* or a complex value modification, depending on the target and the address space pair. Pointer conversions within the same address space must be performed with the `bitcast` instruction. Note that if the address space conversion is legal then both result and operand refer to the same memory location.

Example:

```
%X = addrspacecast i32* %x to i32 addrspace(1)*      ; yields i32 addrspace(1)*:%x
%Y = addrspacecast i32 addrspace(1)* %y to i64 addrspace(2)*      ; yields i64_
↳addrspace(2)*:%y
%Z = addrspacecast <4 x i32*> %z to <4 x float addrspace(3)*>      ; yields <4 x float_
↳addrspace(3)*:%z
```

Other Operations

The instructions in this category are the “miscellaneous” instructions, which defy better classification.

‘icmp’ Instruction

Syntax:

`<result> = icmp <cond> <ty> <op1>, <op2> ; yields i1 or <N x i1>:result`

Overview:

The ‘icmp’ instruction returns a boolean value or a vector of boolean values based on comparison of its two integer, integer vector, pointer, or pointer vector operands.

Arguments:

The ‘icmp’ instruction takes three operands. The first operand is the condition code indicating the kind of comparison to perform. It is not a value, just a keyword. The possible condition codes are:

1. `eq`: equal
2. `ne`: not equal
3. `ugt`: unsigned greater than
4. `uge`: unsigned greater or equal
5. `ult`: unsigned less than
6. `ule`: unsigned less or equal
7. `sgt`: signed greater than
8. `sge`: signed greater or equal
9. `slt`: signed less than
10. `sle`: signed less or equal

The remaining two arguments must be *integer* or *pointer* or integer *vector* typed. They must also be identical types.

Semantics:

The ‘icmp’ compares `op1` and `op2` according to the condition code given as `cond`. The comparison performed always yields either an *il* or vector of *il* result, as follows:

1. `eq`: yields `true` if the operands are equal, `false` otherwise. No sign interpretation is necessary or performed.
2. `ne`: yields `true` if the operands are unequal, `false` otherwise. No sign interpretation is necessary or performed.
3. `ugt`: interprets the operands as unsigned values and yields `true` if `op1` is greater than `op2`.
4. `uge`: interprets the operands as unsigned values and yields `true` if `op1` is greater than or equal to `op2`.
5. `ult`: interprets the operands as unsigned values and yields `true` if `op1` is less than `op2`.

6. `ule`: interprets the operands as unsigned values and yields `true` if `op1` is less than or equal to `op2`.
7. `sgt`: interprets the operands as signed values and yields `true` if `op1` is greater than `op2`.
8. `sge`: interprets the operands as signed values and yields `true` if `op1` is greater than or equal to `op2`.
9. `slt`: interprets the operands as signed values and yields `true` if `op1` is less than `op2`.
10. `sle`: interprets the operands as signed values and yields `true` if `op1` is less than or equal to `op2`.

If the operands are *pointer* typed, the pointer values are compared as if they were integers.

If the operands are integer vectors, then they are compared element by element. The result is an `i1` vector with the same number of elements as the values being compared. Otherwise, the result is an `i1`.

Example:

```
<result> = icmp eq i32 4, 5           ; yields: result=false
<result> = icmp ne float* %X, %X     ; yields: result=false
<result> = icmp ult i16 4, 5          ; yields: result=true
<result> = icmp sgt i16 4, 5          ; yields: result=false
<result> = icmp ule i16 -4, 5         ; yields: result=false
<result> = icmp sge i16 4, 5          ; yields: result=false
```

'fcmp' Instruction

Syntax:

```
<result> = fcmp [fast-math flags]* <cond> <ty> <op1>, <op2> ; yields i1 or <N x i1>
→i1>:result
```

Overview:

The `'fcmp'` instruction returns a boolean value or vector of boolean values based on comparison of its operands.

If the operands are floating point scalars, then the result type is a boolean (*i1*).

If the operands are floating point vectors, then the result type is a vector of boolean with the same number of elements as the operands being compared.

Arguments:

The `'fcmp'` instruction takes three operands. The first operand is the condition code indicating the kind of comparison to perform. It is not a value, just a keyword. The possible condition codes are:

1. `false`: no comparison, always returns false
2. `oeq`: ordered and equal
3. `ogt`: ordered and greater than
4. `oge`: ordered and greater than or equal
5. `olt`: ordered and less than
6. `ole`: ordered and less than or equal

7. `one`: ordered and not equal
8. `ord`: ordered (no nans)
9. `ueq`: unordered or equal
10. `ugt`: unordered or greater than
11. `uge`: unordered or greater than or equal
12. `ult`: unordered or less than
13. `ule`: unordered or less than or equal
14. `une`: unordered or not equal
15. `uno`: unordered (either nans)
16. `true`: no comparison, always returns true

Ordered means that neither operand is a QNAN while *unordered* means that either operand may be a QNAN.

Each of `val1` and `val2` arguments must be either a *floating point* type or a *vector* of floating point type. They must have identical types.

Semantics:

The `fcmp` instruction compares `op1` and `op2` according to the condition code given as `cond`. If the operands are vectors, then the vectors are compared element by element. Each comparison performed always yields an *il* result, as follows:

1. `false`: always yields `false`, regardless of operands.
2. `oeq`: yields `true` if both operands are not a QNAN and `op1` is equal to `op2`.
3. `ogt`: yields `true` if both operands are not a QNAN and `op1` is greater than `op2`.
4. `oge`: yields `true` if both operands are not a QNAN and `op1` is greater than or equal to `op2`.
5. `olt`: yields `true` if both operands are not a QNAN and `op1` is less than `op2`.
6. `ole`: yields `true` if both operands are not a QNAN and `op1` is less than or equal to `op2`.
7. `one`: yields `true` if both operands are not a QNAN and `op1` is not equal to `op2`.
8. `ord`: yields `true` if both operands are not a QNAN.
9. `ueq`: yields `true` if either operand is a QNAN or `op1` is equal to `op2`.
10. `ugt`: yields `true` if either operand is a QNAN or `op1` is greater than `op2`.
11. `uge`: yields `true` if either operand is a QNAN or `op1` is greater than or equal to `op2`.
12. `ult`: yields `true` if either operand is a QNAN or `op1` is less than `op2`.
13. `ule`: yields `true` if either operand is a QNAN or `op1` is less than or equal to `op2`.
14. `une`: yields `true` if either operand is a QNAN or `op1` is not equal to `op2`.
15. `uno`: yields `true` if either operand is a QNAN.
16. `true`: always yields `true`, regardless of operands.

The `fcmp` instruction can also optionally take any number of *fast-math flags*, which are optimization hints to enable otherwise unsafe floating point optimizations.

Any set of fast-math flags are legal on an `fcmp` instruction, but the only flags that have any effect on its semantics are those that allow assumptions to be made about the values of input arguments; namely `nnan`, `ninf`, and `nsz`. See *Fast-Math Flags* for more information.

Example:

```
<result> = fcmp oeq float 4.0, 5.0    ; yields: result=false
<result> = fcmp one float 4.0, 5.0    ; yields: result=true
<result> = fcmp olt float 4.0, 5.0    ; yields: result=true
<result> = fcmp ueq double 1.0, 2.0   ; yields: result=false
```

‘phi’ Instruction

Syntax:

```
<result> = phi <ty> [ <val0>, <label0>], ...
```

Overview:

The ‘phi’ instruction is used to implement the ϕ node in the SSA graph representing the function.

Arguments:

The type of the incoming values is specified with the first type field. After this, the ‘phi’ instruction takes a list of pairs as arguments, with one pair for each predecessor basic block of the current block. Only values of *first class* type may be used as the value arguments to the PHI node. Only labels may be used as the label arguments.

There must be no non-phi instructions between the start of a basic block and the PHI instructions: i.e. PHI instructions must be first in a basic block.

For the purposes of the SSA form, the use of each incoming value is deemed to occur on the edge from the corresponding predecessor block to the current block (but after any definition of an ‘invoke’ instruction’s return value on the same edge).

Semantics:

At runtime, the ‘phi’ instruction logically takes on the value specified by the pair corresponding to the predecessor basic block that executed just prior to the current block.

Example:

```
Loop:           ; Infinite loop that counts from 0 on up...
  %indvar = phi i32 [ 0, %LoopHeader ], [ %nextindvar, %Loop ]
  %nextindvar = add i32 %indvar, 1
  br label %Loop
```

‘select’ Instruction

Syntax:

```
<result> = select selty <cond>, <ty> <val1>, <ty> <val2>           ; yields ty  
selty is either i1 or {<N x i1>}
```

Overview:

The ‘select’ instruction is used to choose one value based on a condition, without IR-level branching.

Arguments:

The ‘select’ instruction requires an ‘i1’ value or a vector of ‘i1’ values indicating the condition, and two values of the same *first class* type.

Semantics:

If the condition is an i1 and it evaluates to 1, the instruction returns the first value argument; otherwise, it returns the second value argument.

If the condition is a vector of i1, then the value arguments must be vectors of the same size, and the selection is done element by element.

If the condition is an i1 and the value arguments are vectors of the same size, then an entire vector is selected.

Example:

```
%X = select i1 true, i8 17, i8 42           ; yields i8:17
```

‘call’ Instruction

Syntax:

```
<result> = [tail | musttail | notail ] call [fast-math flags] [cconv] [ret attrs] <ty>  
→ |<fnty> <fnptrval>(<function args>) [fn attrs]  
   [ operand bundles ]
```

Overview:

The ‘call’ instruction represents a simple function call.

Arguments:

This instruction requires several arguments:

1. The optional `tail` and `musttail` markers indicate that the optimizers should perform tail call optimization. The `tail` marker is a hint that *can be ignored*. The `musttail` marker means that the call must be tail call optimized in order for the program to be correct. The `musttail` marker provides these guarantees:
 - (a) The call will not cause unbounded stack growth if it is part of a recursive cycle in the call graph.
 - (b) Arguments with the *inalloca* attribute are forwarded in place.

Both markers imply that the callee does not access allocas or varargs from the caller. Calls marked `musttail` must obey the following additional rules:

- The call must immediately precede a *ret* instruction, or a pointer bitcast followed by a *ret* instruction.
- The *ret* instruction must return the (possibly bitcasted) value produced by the call or void.
- The caller and callee prototypes must match. Pointer types of parameters or return types may differ in pointee type, but not in address space.
- The calling conventions of the caller and callee must match.
- All ABI-impacting function attributes, such as *sret*, *byval*, *inreg*, *returned*, and *inalloca*, must match.
- The callee must be varargs iff the caller is varargs. Bitcasting a non-varargs function to the appropriate varargs type is legal so long as the non-varargs prefixes obey the other rules.

Tail call optimization for calls marked `tail` is guaranteed to occur if the following conditions are met:

- Caller and callee both have the calling convention *fastcc*.
 - The call is in tail position (*ret* immediately follows call and *ret* uses value of call or is void).
 - Option `-tailcallopt` is enabled, or `llvm::GuaranteedTailCallOpt` is `true`.
 - *Platform-specific constraints are met*.
2. The optional `notail` marker indicates that the optimizers should not add `tail` or `musttail` markers to the call. It is used to prevent tail call optimization from being performed on the call.
 3. The optional `fast-math flags` marker indicates that the call has one or more *fast-math flags*, which are optimization hints to enable otherwise unsafe floating-point optimizations. Fast-math flags are only valid for calls that return a floating-point scalar or vector type.
 4. The optional “`cconv`” marker indicates which *calling convention* the call should use. If none is specified, the call defaults to using C calling conventions. The calling convention of the call must match the calling convention of the target function, or else the behavior is undefined.
 5. The optional *Parameter Attributes* list for return values. Only ‘*zeroext*’, ‘*signext*’, and ‘*inreg*’ attributes are valid here.
 6. ‘*ty*’: the type of the call instruction itself which is also the type of the return value. Functions that return no value are marked `void`.
 7. ‘*fnty*’: shall be the signature of the function being called. The argument types must match the types implied by this signature. This type can be omitted if the function is not varargs.
 8. ‘*fnptrval*’: An LLVM value containing a pointer to a function to be called. In most cases, this is a direct function call, but indirect `call`’s are just as possible, calling an arbitrary pointer to function value.
 9. ‘*function args*’: argument list whose types match the function signature argument types and parameter attributes. All arguments must be of *first class* type. If the function signature indicates the function accepts a variable number of arguments, the extra arguments can be specified.

10. The optional *function attributes* list.
11. The optional *operand bundles* list.

Semantics:

The ‘call’ instruction is used to cause control flow to transfer to a specified function, with its incoming arguments bound to the specified values. Upon a ‘ret’ instruction in the called function, control flow continues with the instruction after the function call, and the return value of the function is bound to the result argument.

Example:

```
%retval = call i32 @test(i32 %argc)
call i32 (i8*, ...) @printf(i8* %msg, i32 12, i8 42) ; yields i32
%X = tail call i32 @foo() ; yields i32
%Y = tail call fastcc i32 @foo() ; yields i32
call void @foo(i8 97 signext)

%struct.A = type { i32, i8 }
%r = call %struct.A @foo() ; yields { i32, i8 }
%gr = extractvalue %struct.A %r, 0 ; yields i32
%gr1 = extractvalue %struct.A %r, 1 ; yields i8
%Z = call void @foo() noreturn ; indicates that %foo never returns.
↳normally
%ZZ = call zeroext i32 @bar() ; Return value is %zero extended
```

llvm treats calls to some functions with names and arguments that match the standard C99 library as being the C99 library functions, and may perform optimizations or generate code for them under that assumption. This is something we’d like to change in the future to provide better support for freestanding environments and non-C-based languages.

‘va_arg’ Instruction

Syntax:

```
<resultval> = va_arg <va_list*> <arglist>, <argty>
```

Overview:

The ‘va_arg’ instruction is used to access arguments passed through the “variable argument” area of a function call. It is used to implement the `va_arg` macro in C.

Arguments:

This instruction takes a `va_list*` value and the type of the argument. It returns a value of the specified argument type and increments the `va_list` to point to the next argument. The actual type of `va_list` is target specific.

Semantics:

The ‘va_arg’ instruction loads an argument of the specified type from the specified va_list and causes the va_list to point to the next argument. For more information, see the variable argument handling *Intrinsic Functions*.

It is legal for this instruction to be called in a function which does not take a variable number of arguments, for example, the `vfprintf` function.

va_arg is an LLVM instruction instead of an *intrinsic function* because it takes a type as an argument.

Example:

See the *variable argument processing* section.

Note that the code generator does not yet fully support va_arg on many targets. Also, it does not currently support va_arg with aggregate types on any target.

‘landingpad’ Instruction

Syntax:

```
<resultval> = landingpad <resultty> <clause>+
<resultval> = landingpad <resultty> cleanup <clause>*

<clause> := catch <type> <value>
<clause> := filter <array constant type> <array constant>
```

Overview:

The ‘landingpad’ instruction is used by LLVM’s *exception handling system* to specify that a basic block is a landing pad — one where the exception lands, and corresponds to the code found in the catch portion of a try/catch sequence. It defines values supplied by the *personality function* upon re-entry to the function. The resultval has the type resultty.

Arguments:

The optional cleanup flag indicates that the landing pad block is a cleanup.

A clause begins with the clause type — catch or filter — and contains the global variable representing the “type” that may be caught or filtered respectively. Unlike the catch clause, the filter clause takes an array constant as its argument. Use “[0 x i8**] undef” for a filter which cannot throw. The ‘landingpad’ instruction must contain *at least* one clause or the cleanup flag.

Semantics:

The ‘landingpad’ instruction defines the values which are set by the *personality function* upon re-entry to the function, and therefore the “result type” of the landingpad instruction. As with calling conventions, how the personality function results are represented in LLVM IR is target specific.

The clauses are applied in order from top to bottom. If two landingpad instructions are merged together through inlining, the clauses from the calling function are appended to the list of clauses. When the call stack is being unwound

due to an exception being thrown, the exception is compared against each `clause` in turn. If it doesn't match any of the clauses, and the `cleanup` flag is not set, then unwinding continues further up the call stack.

The `landingpad` instruction has several restrictions:

- A landing pad block is a basic block which is the unwind destination of an 'invoke' instruction.
- A landing pad block must have a 'landingpad' instruction as its first non-PHI instruction.
- There can be only one 'landingpad' instruction within the landing pad block.
- A basic block that is not a landing pad block may not include a 'landingpad' instruction.

Example:

```
;; A landing pad which can catch an integer.
%res = landingpad { i8*, i32 }
      catch i8** @_ZTIi
;; A landing pad that is a cleanup.
%res = landingpad { i8*, i32 }
      cleanup
;; A landing pad which can catch an integer and can only throw a double.
%res = landingpad { i8*, i32 }
      catch i8** @_ZTIi
      filter [1 x i8**] [@_ZTId]
```

'catchpad' Instruction

Syntax:

```
<resultval> = catchpad within <catchswitch> [<args>*]
```

Overview:

The 'catchpad' instruction is used by LLVM's [exception handling system](#) to specify that a basic block begins a catch handler — one where a personality routine attempts to transfer control to catch an exception.

Arguments:

The `catchswitch` operand must always be a token produced by a [catchswitch](#) instruction in a predecessor block. This ensures that each `catchpad` has exactly one predecessor block, and it always terminates in a `catchswitch`.

The `args` correspond to whatever information the personality routine requires to know if this is an appropriate handler for the exception. Control will transfer to the `catchpad` if this is the first appropriate handler for the exception.

The `resultval` has the type [token](#) and is used to match the `catchpad` to corresponding [catchrets](#) and other nested EH pads.

Semantics:

When the call stack is being unwound due to an exception being thrown, the exception is compared against the `args`. If it doesn't match, control will not reach the `catchpad` instruction. The representation of `args` is entirely target and personality function-specific.

Like the *landingpad* instruction, the `catchpad` instruction must be the first non-phi of its parent basic block.

The meaning of the tokens produced and consumed by `catchpad` and other “pad” instructions is described in the [Windows exception handling documentation](#).

When a `catchpad` has been “entered” but not yet “exited” (as described in the [EH documentation](#)), it is undefined behavior to execute a *call* or *invoke* that does not carry an appropriate “*funclet*” *bundle*.

Example:

```
dispatch:
  %cs = catchswitch within none [label %handler0] unwind to caller
  ;; A catch block which can catch an integer.
handler0:
  %tok = catchpad within %cs [i8** @_ZTIi]
```

‘cleanuppad’ Instruction

Syntax:

```
<resultval> = cleanuppad within <parent> [<args>*]
```

Overview:

The ‘cleanuppad’ instruction is used by LLVM’s [exception handling system](#) to specify that a basic block is a cleanup block — one where a personality routine attempts to transfer control to run cleanup actions. The `args` correspond to whatever additional information the *personality function* requires to execute the cleanup. The `resultval` has the type *token* and is used to match the `cleanuppad` to corresponding *cleanuprets*. The `parent` argument is the token of the funclet that contains the `cleanuppad` instruction. If the `cleanuppad` is not inside a funclet, this operand may be the token `none`.

Arguments:

The instruction takes a list of arbitrary values which are interpreted by the *personality function*.

Semantics:

When the call stack is being unwound due to an exception being thrown, the *personality function* transfers control to the `cleanuppad` with the aid of the personality-specific arguments. As with calling conventions, how the personality function results are represented in LLVM IR is target specific.

The `cleanuppad` instruction has several restrictions:

- A cleanup block is a basic block which is the unwind destination of an exceptional instruction.

- A cleanup block must have a ‘`cleanuppad`’ instruction as its first non-PHI instruction.
- There can be only one ‘`cleanuppad`’ instruction within the cleanup block.
- A basic block that is not a cleanup block may not include a ‘`cleanuppad`’ instruction.

When a `cleanuppad` has been “entered” but not yet “exited” (as described in the [EH documentation](#)), it is undefined behavior to execute a *call* or *invoke* that does not carry an appropriate “*funclet*” *bundle*.

Example:

```
%tok = cleanuppad within %cs []
```

1.1.13 Intrinsic Functions

LLVM supports the notion of an “intrinsic function”. These functions have well known names and semantics and are required to follow certain restrictions. Overall, these intrinsics represent an extension mechanism for the LLVM language that does not require changing all of the transformations in LLVM when adding to the language (or the bitcode reader/writer, the parser, etc. ...).

Intrinsic function names must all start with an “`llvm.`” prefix. This prefix is reserved in LLVM for intrinsic names; thus, function names may not begin with this prefix. Intrinsic functions must always be external functions: you cannot define the body of intrinsic functions. Intrinsic functions may only be used in `call` or `invoke` instructions: it is illegal to take the address of an intrinsic function. Additionally, because intrinsic functions are part of the LLVM language, it is required if any are added that they be documented here.

Some intrinsic functions can be overloaded, i.e., the intrinsic represents a family of functions that perform the same operation but on different data types. Because LLVM can represent over 8 million different integer types, overloading is used commonly to allow an intrinsic function to operate on any integer type. One or more of the argument types or the result type can be overloaded to accept any integer type. Argument types may also be defined as exactly matching a previous argument’s type or the result type. This allows an intrinsic function which accepts multiple arguments, but needs all of them to be of the same type, to only be overloaded with respect to a single argument or the result.

Overloaded intrinsics will have the names of its overloaded argument types encoded into its function name, each preceded by a period. Only those types which are overloaded result in a name suffix. Arguments whose type is matched against another type do not. For example, the `llvm.ctpop` function can take an integer of any width and returns an integer of exactly the same integer width. This leads to a family of functions such as `i8 @llvm.ctpop.i8(i8 %val)` and `i29 @llvm.ctpop.i29(i29 %val)`. Only one type, the return type, is overloaded, and only one type suffix is required. Because the argument’s type is matched against the return type, it does not require its own name suffix.

To learn how to add an intrinsic function, please see the [Extending LLVM Guide](#).

Variable Argument Handling Intrinsics

Variable argument support is defined in LLVM with the *va_arg* instruction and these three intrinsic functions. These functions are related to the similarly named macros defined in the `<stdarg.h>` header file.

All of these functions operate on arguments that use a target-specific value type “`va_list`”. The LLVM assembly language reference manual does not define what this type is, so all transformations should be prepared to handle these functions regardless of the type used.

This example shows how the *va_arg* instruction and the variable argument handling intrinsic functions are used.


```

; This struct is different for every platform. For most platforms,
; it is merely an i8*.
%struct.va_list = type { i8* }

; For Unix x86_64 platforms, va_list is the following struct:
; %struct.va_list = type { i32, i32, i8*, i8* }

define i32 @test(i32 %X, ...) {
  ; Initialize variable argument processing
  %ap = alloca %struct.va_list
  %ap2 = bitcast %struct.va_list* %ap to i8*
  call void @llvm.va_start(i8* %ap2)

  ; Read a single integer argument
  %tmp = va_arg i8* %ap2, i32

  ; Demonstrate usage of llvm.va_copy and llvm.va_end
  %aq = alloca i8*
  %aq2 = bitcast i8** %aq to i8*
  call void @llvm.va_copy(i8* %aq2, i8* %ap2)
  call void @llvm.va_end(i8* %aq2)

  ; Stop processing of arguments.
  call void @llvm.va_end(i8* %ap2)
  ret i32 %tmp
}

declare void @llvm.va_start(i8*)
declare void @llvm.va_copy(i8*, i8*)
declare void @llvm.va_end(i8*)

```

‘llvm.va_start’ Intrinsic

Syntax:

```
declare void @llvm.va_start(i8* <arglist>)
```

Overview:

The ‘llvm.va_start’ intrinsic initializes *<arglist> for subsequent use by va_arg.

Arguments:

The argument is a pointer to a va_list element to initialize.

Semantics:

The ‘llvm.va_start’ intrinsic works just like the va_start macro available in C. In a target-dependent way, it initializes the va_list element to which the argument points, so that the next call to va_arg will produce the first variable argument passed to the function. Unlike the C va_start macro, this intrinsic does not need to know the last argument of the function as the compiler can figure that out.

‘llvm.va_end’ Intrinsic

Syntax:

```
declare void @llvm.va_end(i8* <arglist>)
```

Overview:

The ‘llvm.va_end’ intrinsic destroys *<arglist>, which has been initialized previously with `llvm.va_start` or `llvm.va_copy`.

Arguments:

The argument is a pointer to a `va_list` to destroy.

Semantics:

The ‘llvm.va_end’ intrinsic works just like the `va_end` macro available in C. In a target-dependent way, it destroys the `va_list` element to which the argument points. Calls to *llvm.va_start* and *llvm.va_copy* must be matched exactly with calls to `llvm.va_end`.

‘llvm.va_copy’ Intrinsic

Syntax:

```
declare void @llvm.va_copy(i8* <destarglist>, i8* <srcarglist>)
```

Overview:

The ‘llvm.va_copy’ intrinsic copies the current argument position from the source argument list to the destination argument list.

Arguments:

The first argument is a pointer to a `va_list` element to initialize. The second argument is a pointer to a `va_list` element to copy from.

Semantics:

The ‘llvm.va_copy’ intrinsic works just like the `va_copy` macro available in C. In a target-dependent way, it copies the source `va_list` element into the destination `va_list` element. This intrinsic is necessary because the “`llvm.va_start`” intrinsic may be arbitrarily complex and require, for example, memory allocation.

Accurate Garbage Collection Intrinsics

LLVM's support for [Accurate Garbage Collection](#) (GC) requires the frontend to generate code containing appropriate intrinsic calls and select an appropriate GC strategy which knows how to lower these intrinsics in a manner which is appropriate for the target collector.

These intrinsics allow identification of *GC roots on the stack*, as well as garbage collector implementations that require *read* and *write* barriers. Frontends for type-safe garbage collected languages should generate these intrinsics to make use of the LLVM garbage collectors. For more details, see [Garbage Collection with LLVM](#).

Experimental Statepoint Intrinsics

LLVM provides an second experimental set of intrinsics for describing garbage collection safe-points in compiled code. These intrinsics are an alternative to the `llvm.gcroot` intrinsics, but are compatible with the ones for *read* and *write* barriers. The differences in approach are covered in the [Garbage Collection with LLVM](#) documentation. The intrinsics themselves are described in [Garbage Collection Safe-points in LLVM](#).

'llvm.gcroot' Intrinsic

Syntax:

```
declare void @llvm.gcroot(i8** %ptrloc, i8* %metadata)
```

Overview:

The 'llvm.gcroot' intrinsic declares the existence of a GC root to the code generator, and allows some metadata to be associated with it.

Arguments:

The first argument specifies the address of a stack object that contains the root pointer. The second pointer (which must be either a constant or a global value address) contains the meta-data to be associated with the root.

Semantics:

At runtime, a call to this intrinsic stores a null pointer into the "ptrloc" location. At compile-time, the code generator generates information to allow the runtime to find the pointer at GC safe points. The 'llvm.gcroot' intrinsic may only be used in a function which *specifies a GC algorithm*.

'llvm.gcread' Intrinsic

Syntax:

```
declare i8* @llvm.gcread(i8* %ObjPtr, i8** %Ptr)
```

Overview:

The `'llvm.gcread'` intrinsic identifies reads of references from heap locations, allowing garbage collector implementations that require read barriers.

Arguments:

The second argument is the address to read from, which should be an address allocated from the garbage collector. The first object is a pointer to the start of the referenced object, if needed by the language runtime (otherwise null).

Semantics:

The `'llvm.gcread'` intrinsic has the same semantics as a load instruction, but may be replaced with substantially more complex code by the garbage collector runtime, as needed. The `'llvm.gcread'` intrinsic may only be used in a function which *specifies a GC algorithm*.

'llvm.gcwrite' Intrinsic**Syntax:**

```
declare void @llvm.gcwrite(i8* %P1, i8* %Obj, i8** %P2)
```

Overview:

The `'llvm.gcwrite'` intrinsic identifies writes of references to heap locations, allowing garbage collector implementations that require write barriers (such as generational or reference counting collectors).

Arguments:

The first argument is the reference to store, the second is the start of the object to store it to, and the third is the address of the field of Obj to store to. If the runtime does not require a pointer to the object, Obj may be null.

Semantics:

The `'llvm.gcwrite'` intrinsic has the same semantics as a store instruction, but may be replaced with substantially more complex code by the garbage collector runtime, as needed. The `'llvm.gcwrite'` intrinsic may only be used in a function which *specifies a GC algorithm*.

Code Generator Intrinsics

These intrinsics are provided by LLVM to expose special features that may only be implemented with code generator support.

'llvm.returnaddress' Intrinsic

Syntax:

```
declare i8* @llvm.returnaddress(i32 <level>)
```

Overview:

The 'llvm.returnaddress' intrinsic attempts to compute a target-specific value indicating the return address of the current function or one of its callers.

Arguments:

The argument to this intrinsic indicates which function to return the address for. Zero indicates the calling function, one indicates its caller, etc. The argument is **required** to be a constant integer value.

Semantics:

The 'llvm.returnaddress' intrinsic either returns a pointer indicating the return address of the specified call frame, or zero if it cannot be identified. The value returned by this intrinsic is likely to be incorrect or 0 for arguments other than zero, so it should only be used for debugging purposes.

Note that calling this intrinsic does not prevent function inlining or other aggressive transformations, so the value returned may not be that of the obvious source-language caller.

'llvm.addressofreturnaddress' Intrinsic

Syntax:

```
declare i8* @llvm.addressofreturnaddress()
```

Overview:

The 'llvm.addressofreturnaddress' intrinsic returns a target-specific pointer to the place in the stack frame where the return address of the current function is stored.

Semantics:

Note that calling this intrinsic does not prevent function inlining or other aggressive transformations, so the value returned may not be that of the obvious source-language caller.

This intrinsic is only implemented for x86.

'llvm.frameaddress' Intrinsic

Syntax:

```
declare i8* @llvm.frameaddress(i32 <level>)
```

Overview:

The 'llvm.frameaddress' intrinsic attempts to return the target-specific frame pointer value for the specified stack frame.

Arguments:

The argument to this intrinsic indicates which function to return the frame pointer for. Zero indicates the calling function, one indicates its caller, etc. The argument is **required** to be a constant integer value.

Semantics:

The 'llvm.frameaddress' intrinsic either returns a pointer indicating the frame address of the specified call frame, or zero if it cannot be identified. The value returned by this intrinsic is likely to be incorrect or 0 for arguments other than zero, so it should only be used for debugging purposes.

Note that calling this intrinsic does not prevent function inlining or other aggressive transformations, so the value returned may not be that of the obvious source-language caller.

'llvm.localescape' and 'llvm.localrecover' Intrinsics

Syntax:

```
declare void @llvm.localescape(...)
declare i8* @llvm.localrecover(i8* %func, i8* %fp, i32 %idx)
```

Overview:

The 'llvm.localescape' intrinsic escapes offsets of a collection of static allocas, and the 'llvm.localrecover' intrinsic applies those offsets to a live frame pointer to recover the address of the allocation. The offset is computed during frame layout of the caller of llvm.localescape.

Arguments:

All arguments to 'llvm.localescape' must be pointers to static allocas or casts of static allocas. Each function can only call 'llvm.localescape' once, and it can only do so from the entry block.

The func argument to 'llvm.localrecover' must be a constant bitcasted pointer to a function defined in the current module. The code generator cannot determine the frame allocation offset of functions defined in other modules.

The `fp` argument to `'llvm.localrecover'` must be a frame pointer of a call frame that is currently live. The return value of `'llvm.localaddress'` is one way to produce such a value, but various runtimes also expose a suitable pointer in platform-specific ways.

The `idx` argument to `'llvm.localrecover'` indicates which alloca passed to `'llvm.localescape'` to recover. It is zero-indexed.

Semantics:

These intrinsics allow a group of functions to share access to a set of local stack allocations of a one parent function. The parent function may call the `'llvm.localescape'` intrinsic once from the function entry block, and the child functions can use `'llvm.localrecover'` to access the escaped allocas. The `'llvm.localescape'` intrinsic blocks inlining, as inlining changes where the escaped allocas are allocated, which would break attempts to use `'llvm.localrecover'`.

`'llvm.read_register'` and `'llvm.write_register'` Intrinsics

Syntax:

```
declare i32 @llvm.read_register.i32(metadata)
declare i64 @llvm.read_register.i64(metadata)
declare void @llvm.write_register.i32(metadata, i32 @value)
declare void @llvm.write_register.i64(metadata, i64 @value)
!0 = !{"sp\00"}
```

Overview:

The `'llvm.read_register'` and `'llvm.write_register'` intrinsics provides access to the named register. The register must be valid on the architecture being compiled to. The type needs to be compatible with the register being read.

Semantics:

The `'llvm.read_register'` intrinsic returns the current value of the register, where possible. The `'llvm.write_register'` intrinsic sets the current value of the register, where possible.

This is useful to implement named register global variables that need to always be mapped to a specific register, as is common practice on bare-metal programs including OS kernels.

The compiler doesn't check for register availability or use of the used register in surrounding code, including inline assembly. Because of that, allocatable registers are not supported.

Warning: So far it only works with the stack pointer on selected architectures (ARM, AArch64, PowerPC and x86_64). Significant amount of work is needed to support other registers and even more so, allocatable registers.

'llvm.stacksave' Intrinsic

Syntax:

```
declare i8* @llvm.stacksave()
```

Overview:

The 'llvm.stacksave' intrinsic is used to remember the current state of the function stack, for use with [llvm.stackrestore](#). This is useful for implementing language features like scoped automatic variable sized arrays in C99.

Semantics:

This intrinsic returns an opaque pointer value that can be passed to [llvm.stackrestore](#). When an `llvm.stackrestore` intrinsic is executed with a value saved from `llvm.stacksave`, it effectively restores the state of the stack to the state it was in when the `llvm.stacksave` intrinsic executed. In practice, this pops any [alloca](#) blocks from the stack that were allocated after the `llvm.stacksave` was executed.

'llvm.stackrestore' Intrinsic

Syntax:

```
declare void @llvm.stackrestore(i8* %ptr)
```

Overview:

The 'llvm.stackrestore' intrinsic is used to restore the state of the function stack to the state it was in when the corresponding [llvm.stacksave](#) intrinsic executed. This is useful for implementing language features like scoped automatic variable sized arrays in C99.

Semantics:

See the description for [llvm.stacksave](#).

'llvm.get.dynamic.area.offset' Intrinsic

Syntax:

```
declare i32 @llvm.get.dynamic.area.offset.i32()
declare i64 @llvm.get.dynamic.area.offset.i64()
```


Overview:

The `'llvm.get.dynamic.area.offset.*'` intrinsic family is used to get the offset from native stack pointer to the address of the most recent dynamic alloca on the caller's stack. These intrinsics are intended for use in combination with [llvm.stacksave](#) to get a pointer to the most recent dynamic alloca. This is useful, for example, for AddressSanitizer's stack unpoisoning routines.

Semantics:

These intrinsics return a non-negative integer value that can be used to get the address of the most recent dynamic alloca, allocated by [alloca](#) on the caller's stack. In particular, for targets where stack grows downwards, adding this offset to the native stack pointer would get the address of the most recent dynamic alloca. For targets where stack grows upwards, the situation is a bit more complicated, because subtracting this value from stack pointer would get the address one past the end of the most recent dynamic alloca.

Although for most targets `llvm.get.dynamic.area.offset <int_get_dynamic_area_offset>` returns just a zero, for others, such as PowerPC and PowerPC64, it returns a compile-time-known constant value.

The return value type of `llvm.get.dynamic.area.offset` must match the target's default address space's (address space 0) pointer type.

'llvm.prefetch' Intrinsic**Syntax:**

```
declare void @llvm.prefetch(i8* <address>, i32 <rw>, i32 <locality>, i32 <cache type>)
```

Overview:

The `'llvm.prefetch'` intrinsic is a hint to the code generator to insert a prefetch instruction if supported; otherwise, it is a noop. Prefetches have no effect on the behavior of the program but can change its performance characteristics.

Arguments:

`address` is the address to be prefetched, `rw` is the specifier determining if the fetch should be for a read (0) or write (1), and `locality` is a temporal locality specifier ranging from (0) - no locality, to (3) - extremely local keep in cache. The `cache type` specifies whether the prefetch is performed on the data (1) or instruction (0) cache. The `rw`, `locality` and `cache type` arguments must be constant integers.

Semantics:

This intrinsic does not modify the behavior of the program. In particular, prefetches cannot trap and do not produce a value. On targets that support this intrinsic, the prefetch can provide hints to the processor cache for better performance.

'llvm.pcmarker' Intrinsic

Syntax:

```
declare void @llvm.pcmarker(i32 <id>)
```

Overview:

The `'llvm.pcmarker'` intrinsic is a method to export a Program Counter (PC) in a region of code to simulators and other tools. The method is target specific, but it is expected that the marker will use exported symbols to transmit the PC of the marker. The marker makes no guarantees that it will remain with any specific instruction after optimizations. It is possible that the presence of a marker will inhibit optimizations. The intended use is to be inserted after optimizations to allow correlations of simulation runs.

Arguments:

`id` is a numerical id identifying the marker.

Semantics:

This intrinsic does not modify the behavior of the program. Backends that do not support this intrinsic may ignore it.

'llvm.readcyclecounter' Intrinsic

Syntax:

```
declare i64 @llvm.readcyclecounter()
```

Overview:

The `'llvm.readcyclecounter'` intrinsic provides access to the cycle counter register (or similar low latency, high accuracy clocks) on those targets that support it. On X86, it should map to RDTSC. On Alpha, it should map to RPCC. As the backing counters overflow quickly (on the order of 9 seconds on alpha), this should only be used for small timings.

Semantics:

When directly supported, reading the cycle counter should not modify any memory. Implementations are allowed to either return a application specific value or a system wide value. On backends without support, this is lowered to a constant 0.

Note that runtime support may be conditional on the privilege-level code is running at and the host platform.

'llvm.clear_cache' Intrinsic

Syntax:

```
declare void @llvm.clear_cache(i8*, i8*)
```

Overview:

The 'llvm.clear_cache' intrinsic ensures visibility of modifications in the specified range to the execution unit of the processor. On targets with non-unified instruction and data cache, the implementation flushes the instruction cache.

Semantics:

On platforms with coherent instruction and data caches (e.g. x86), this intrinsic is a nop. On platforms with non-coherent instruction and data cache (e.g. ARM, MIPS), the intrinsic is lowered either to appropriate instructions or a system call, if cache flushing requires special privileges.

The default behavior is to emit a call to `__clear_cache` from the run time library.

This intrinsic does *not* empty the instruction pipeline. Modifications of the current function are outside the scope of the intrinsic.

'llvm.instrprof_increment' Intrinsic

Syntax:

```
declare void @llvm.instrprof_increment(i8* <name>, i64 <hash>,
                                       i32 <num-counters>, i32 <index>)
```

Overview:

The 'llvm.instrprof_increment' intrinsic can be emitted by a frontend for use with instrumentation based profiling. These will be lowered by the `-instrprof` pass to generate execution counts of a program at runtime.

Arguments:

The first argument is a pointer to a global variable containing the name of the entity being instrumented. This should generally be the (mangled) function name for a set of counters.

The second argument is a hash value that can be used by the consumer of the profile data to detect changes to the instrumented source, and the third is the number of counters associated with `name`. It is an error if `hash` or `num-counters` differ between two instances of `instrprof_increment` that refer to the same name.

The last argument refers to which of the counters for `name` should be incremented. It should be a value between 0 and `num-counters`.

Semantics:

This intrinsic represents an increment of a profiling counter. It will cause the `-instrprof` pass to generate the appropriate data structures and the code to increment the appropriate value, in a format that can be written out by a compiler runtime and consumed via the `llvm-profdata` tool.

`'llvm.instrprof_increment_step'` Intrinsic

Syntax:

```
declare void @llvm.instrprof_increment_step(i8* <name>, i64 <hash>,  
                                           i32 <num-counters>,  
                                           i32 <index>, i64 <step>)
```

Overview:

The `'llvm.instrprof_increment_step'` intrinsic is an extension to the `'llvm.instrprof_increment'` intrinsic with an additional fifth argument to specify the step of the increment.

Arguments:

The first four arguments are the same as `'llvm.instrprof_increment'` intrinsic.

The last argument specifies the value of the increment of the counter variable.

Semantics:

See description of `'llvm.instrprof_increment'` intrinsic.

`'llvm.instrprof_value_profile'` Intrinsic

Syntax:

```
declare void @llvm.instrprof_value_profile(i8* <name>, i64 <hash>,  
                                           i64 <value>, i32 <value_kind>,  
                                           i32 <index>)
```

Overview:

The `'llvm.instrprof_value_profile'` intrinsic can be emitted by a frontend for use with instrumentation based profiling. This will be lowered by the `-instrprof` pass to find out the target values, instrumented expressions take in a program at runtime.

Arguments:

The first argument is a pointer to a global variable containing the name of the entity being instrumented. `name` should generally be the (mangled) function name for a set of counters.

The second argument is a hash value that can be used by the consumer of the profile data to detect changes to the instrumented source. It is an error if `hash` differs between two instances of `llvm.instrprof_*` that refer to the same name.

The third argument is the value of the expression being profiled. The profiled expression's value should be representable as an unsigned 64-bit value. The fourth argument represents the kind of value profiling that is being done. The supported value profiling kinds are enumerated through the `InstrProfValueKind` type declared in the `<include/llvm/ProfileData/InstrProf.h>` header file. The last argument is the index of the instrumented expression within `name`. It should be `>= 0`.

Semantics:

This intrinsic represents the point where a call to a runtime routine should be inserted for value profiling of target expressions. `-instrprof` pass will generate the appropriate data structures and replace the `llvm.instrprof_value_profile` intrinsic with the call to the profile runtime library with proper arguments.

'llvm.thread.pointer' Intrinsic**Syntax:**

```
declare i8* @llvm.thread.pointer()
```

Overview:

The `'llvm.thread.pointer'` intrinsic returns the value of the thread pointer.

Semantics:

The `'llvm.thread.pointer'` intrinsic returns a pointer to the TLS area for the current thread. The exact semantics of this value are target specific: it may point to the start of TLS area, to the end, or somewhere in the middle. Depending on the target, this intrinsic may read a register, call a helper function, read from an alternate memory space, or perform other operations necessary to locate the TLS area. Not all targets support this intrinsic.

Standard C Library Intrinsics

LLVM provides intrinsics for a few important standard C library functions. These intrinsics allow source-language front-ends to pass information about the alignment of the pointer arguments to the code generator, providing opportunity for more efficient code generation.

‘llvm.memcpy’ Intrinsic

Syntax:

This is an overloaded intrinsic. You can use `llvm.memcpy` on any integer bit width and for different address spaces. Not all targets support all bit widths however.

```
declare void @llvm.memcpy.p0i8.p0i8.i32(i8* <dest>, i8* <src>,  
                                         i32 <len>, i32 <align>, i1 <isvolatile>)  
declare void @llvm.memcpy.p0i8.p0i8.i64(i8* <dest>, i8* <src>,  
                                         i64 <len>, i32 <align>, i1 <isvolatile>)
```

Overview:

The ‘`llvm.memcpy.*`’ intrinsics copy a block of memory from the source location to the destination location.

Note that, unlike the standard `libc` function, the `llvm.memcpy.*` intrinsics do not return a value, takes extra alignment/isvolatile arguments and the pointers can be in specified address spaces.

Arguments:

The first argument is a pointer to the destination, the second is a pointer to the source. The third argument is an integer argument specifying the number of bytes to copy, the fourth argument is the alignment of the source and destination locations, and the fifth is a boolean indicating a volatile access.

If the call to this intrinsic has an alignment value that is not 0 or 1, then the caller guarantees that both the source and destination pointers are aligned to that boundary.

If the `isvolatile` parameter is `true`, the `llvm.memcpy` call is a *volatile operation*. The detailed access behavior is not very cleanly specified and it is unwise to depend on it.

Semantics:

The ‘`llvm.memcpy.*`’ intrinsics copy a block of memory from the source location to the destination location, which are not allowed to overlap. It copies “len” bytes of memory over. If the argument is known to be aligned to some boundary, this can be specified as the fourth argument, otherwise it should be set to 0 or 1 (both meaning no alignment).

‘llvm.memmove’ Intrinsic

Syntax:

This is an overloaded intrinsic. You can use `llvm.memmove` on any integer bit width and for different address space. Not all targets support all bit widths however.

```
declare void @llvm.memmove.p0i8.p0i8.i32(i8* <dest>, i8* <src>,  
                                           i32 <len>, i32 <align>, i1 <isvolatile>)  
declare void @llvm.memmove.p0i8.p0i8.i64(i8* <dest>, i8* <src>,  
                                           i64 <len>, i32 <align>, i1 <isvolatile>)
```

Overview:

The `'llvm.memmove.*'` intrinsics move a block of memory from the source location to the destination location. It is similar to the `'llvm.memcpy'` intrinsic but allows the two memory locations to overlap.

Note that, unlike the standard libc function, the `llvm.memmove.*` intrinsics do not return a value, takes extra alignment/isvolatile arguments and the pointers can be in specified address spaces.

Arguments:

The first argument is a pointer to the destination, the second is a pointer to the source. The third argument is an integer argument specifying the number of bytes to copy, the fourth argument is the alignment of the source and destination locations, and the fifth is a boolean indicating a volatile access.

If the call to this intrinsic has an alignment value that is not 0 or 1, then the caller guarantees that the source and destination pointers are aligned to that boundary.

If the `isvolatile` parameter is `true`, the `llvm.memmove` call is a *volatile operation*. The detailed access behavior is not very cleanly specified and it is unwise to depend on it.

Semantics:

The `'llvm.memmove.*'` intrinsics copy a block of memory from the source location to the destination location, which may overlap. It copies “len” bytes of memory over. If the argument is known to be aligned to some boundary, this can be specified as the fourth argument, otherwise it should be set to 0 or 1 (both meaning no alignment).

'llvm.memset.*' Intrinsics

Syntax:

This is an overloaded intrinsic. You can use `llvm.memset` on any integer bit width and for different address spaces. However, not all targets support all bit widths.

```
declare void @llvm.memset.p0i8.i32(i8* <dest>, i8 <val>,
                                   i32 <len>, i32 <align>, i1 <isvolatile>)
declare void @llvm.memset.p0i8.i64(i8* <dest>, i8 <val>,
                                   i64 <len>, i32 <align>, i1 <isvolatile>)
```

Overview:

The `'llvm.memset.*'` intrinsics fill a block of memory with a particular byte value.

Note that, unlike the standard libc function, the `llvm.memset` intrinsic does not return a value and takes extra alignment/volatile arguments. Also, the destination can be in an arbitrary address space.

Arguments:

The first argument is a pointer to the destination to fill, the second is the byte value with which to fill it, the third argument is an integer argument specifying the number of bytes to fill, and the fourth argument is the known alignment of the destination location.

If the call to this intrinsic has an alignment value that is not 0 or 1, then the caller guarantees that the destination pointer is aligned to that boundary.

If the `isvolatile` parameter is `true`, the `llvm.memset` call is a *volatile operation*. The detailed access behavior is not very cleanly specified and it is unwise to depend on it.

Semantics:

The `'llvm.memset.*'` intrinsics fill “len” bytes of memory starting at the destination location. If the argument is known to be aligned to some boundary, this can be specified as the fourth argument, otherwise it should be set to 0 or 1 (both meaning no alignment).

`'llvm.sqrt.*'` Intrinsic

Syntax:

This is an overloaded intrinsic. You can use `llvm.sqrt` on any floating point or vector of floating point type. Not all targets support all types however.

```
declare float      @llvm.sqrt.f32(float %Val)
declare double     @llvm.sqrt.f64(double %Val)
declare x86_fp80   @llvm.sqrt.f80(x86_fp80 %Val)
declare fp128      @llvm.sqrt.f128(fp128 %Val)
declare ppc_fp128  @llvm.sqrt.ppcfp128(ppcfp128 %Val)
```

Overview:

The `'llvm.sqrt'` intrinsics return the square root of the specified value, returning the same value as the libm `'sqrt'` functions would, but without trapping or setting `errno`.

Arguments:

The argument and return value are floating point numbers of the same type.

Semantics:

This function returns the square root of the operand if it is a nonnegative floating point number.

`'llvm.powi.*'` Intrinsic

Syntax:

This is an overloaded intrinsic. You can use `llvm.powi` on any floating point or vector of floating point type. Not all targets support all types however.


```
declare float @llvm.powi.f32(float %Val, i32 %power)
declare double @llvm.powi.f64(double %Val, i32 %power)
declare x86_fp80 @llvm.powi.f80(x86_fp80 %Val, i32 %power)
declare fp128 @llvm.powi.f128(fp128 %Val, i32 %power)
declare ppc_fp128 @llvm.powi.ppcf128(ppc_fp128 %Val, i32 %power)
```

Overview:

The ‘`llvm.powi.*`’ intrinsics return the first operand raised to the specified (positive or negative) power. The order of evaluation of multiplications is not defined. When a vector of floating point type is used, the second argument remains a scalar integer value.

Arguments:

The second argument is an integer power, and the first is a value to raise to that power.

Semantics:

This function returns the first value raised to the second power with an unspecified sequence of rounding operations.

‘`llvm.sin.*`’ Intrinsic**Syntax:**

This is an overloaded intrinsic. You can use `llvm.sin` on any floating point or vector of floating point type. Not all targets support all types however.

```
declare float @llvm.sin.f32(float %Val)
declare double @llvm.sin.f64(double %Val)
declare x86_fp80 @llvm.sin.f80(x86_fp80 %Val)
declare fp128 @llvm.sin.f128(fp128 %Val)
declare ppc_fp128 @llvm.sin.ppcf128(ppc_fp128 %Val)
```

Overview:

The ‘`llvm.sin.*`’ intrinsics return the sine of the operand.

Arguments:

The argument and return value are floating point numbers of the same type.

Semantics:

This function returns the sine of the specified operand, returning the same values as the `libm sin` functions would, and handles error conditions in the same way.

'llvm.cos.*' Intrinsic

Syntax:

This is an overloaded intrinsic. You can use `llvm.cos` on any floating point or vector of floating point type. Not all targets support all types however.

```
declare float      @llvm.cos.f32(float  %Val)
declare double     @llvm.cos.f64(double %Val)
declare x86_fp80   @llvm.cos.f80(x86_fp80 %Val)
declare fp128      @llvm.cos.f128(fp128 %Val)
declare ppc_fp128  @llvm.cos.ppcf128(ppc_fp128 %Val)
```

Overview:

The `'llvm.cos.*'` intrinsics return the cosine of the operand.

Arguments:

The argument and return value are floating point numbers of the same type.

Semantics:

This function returns the cosine of the specified operand, returning the same values as the `libm cos` functions would, and handles error conditions in the same way.

'llvm.pow.*' Intrinsic

Syntax:

This is an overloaded intrinsic. You can use `llvm.pow` on any floating point or vector of floating point type. Not all targets support all types however.

```
declare float      @llvm.pow.f32(float  %Val, float %Power)
declare double     @llvm.pow.f64(double %Val, double %Power)
declare x86_fp80   @llvm.pow.f80(x86_fp80 %Val, x86_fp80 %Power)
declare fp128      @llvm.pow.f128(fp128 %Val, fp128 %Power)
declare ppc_fp128  @llvm.pow.ppcf128(ppc_fp128 %Val, ppc_fp128 %Power)
```

Overview:

The `'llvm.pow.*'` intrinsics return the first operand raised to the specified (positive or negative) power.

Arguments:

The second argument is a floating point power, and the first is a value to raise to that power.

Semantics:

This function returns the first value raised to the second power, returning the same values as the `libm pow` functions would, and handles error conditions in the same way.

'llvm.exp.*' Intrinsic**Syntax:**

This is an overloaded intrinsic. You can use `llvm.exp` on any floating point or vector of floating point type. Not all targets support all types however.

```
declare float      @llvm.exp.f32(float  %Val)
declare double     @llvm.exp.f64(double %Val)
declare x86_fp80   @llvm.exp.f80(x86_fp80 %Val)
declare fp128      @llvm.exp.f128(fp128 %Val)
declare ppc_fp128  @llvm.exp.ppcfp128(ppc_fp128 %Val)
```

Overview:

The `'llvm.exp.*'` intrinsics compute the base-e exponential of the specified value.

Arguments:

The argument and return value are floating point numbers of the same type.

Semantics:

This function returns the same values as the `libm exp` functions would, and handles error conditions in the same way.

'llvm.exp2.*' Intrinsic**Syntax:**

This is an overloaded intrinsic. You can use `llvm.exp2` on any floating point or vector of floating point type. Not all targets support all types however.

```
declare float      @llvm.exp2.f32(float  %Val)
declare double     @llvm.exp2.f64(double %Val)
declare x86_fp80   @llvm.exp2.f80(x86_fp80 %Val)
declare fp128      @llvm.exp2.f128(fp128 %Val)
declare ppc_fp128  @llvm.exp2.ppcfp128(ppc_fp128 %Val)
```

Overview:

The `'llvm.exp2.*'` intrinsics compute the base-2 exponential of the specified value.

Arguments:

The argument and return value are floating point numbers of the same type.

Semantics:

This function returns the same values as the `libm exp2` functions would, and handles error conditions in the same way.

'llvm.log.*' Intrinsic**Syntax:**

This is an overloaded intrinsic. You can use `llvm.log` on any floating point or vector of floating point type. Not all targets support all types however.

```
declare float      @llvm.log.f32(float  %Val)
declare double     @llvm.log.f64(double %Val)
declare x86_fp80   @llvm.log.f80(x86_fp80 %Val)
declare fp128      @llvm.log.f128(fp128 %Val)
declare ppc_fp128  @llvm.log.ppcfp128(ppc_fp128 %Val)
```

Overview:

The `'llvm.log.*'` intrinsics compute the base-e logarithm of the specified value.

Arguments:

The argument and return value are floating point numbers of the same type.

Semantics:

This function returns the same values as the `libm log` functions would, and handles error conditions in the same way.

'llvm.log10.*' Intrinsic**Syntax:**

This is an overloaded intrinsic. You can use `llvm.log10` on any floating point or vector of floating point type. Not all targets support all types however.

```
declare float      @llvm.log10.f32(float  %Val)
declare double     @llvm.log10.f64(double %Val)
declare x86_fp80   @llvm.log10.f80(x86_fp80 %Val)
declare fp128      @llvm.log10.f128(fp128 %Val)
declare ppc_fp128  @llvm.log10.ppcfp128(ppc_fp128 %Val)
```

Overview:

The `'llvm.log10.*'` intrinsics compute the base-10 logarithm of the specified value.

Arguments:

The argument and return value are floating point numbers of the same type.

Semantics:

This function returns the same values as the `libm log10` functions would, and handles error conditions in the same way.

'llvm.log2.*' Intrinsic**Syntax:**

This is an overloaded intrinsic. You can use `llvm.log2` on any floating point or vector of floating point type. Not all targets support all types however.

```
declare float      @llvm.log2.f32(float  %Val)
declare double     @llvm.log2.f64(double %Val)
declare x86_fp80   @llvm.log2.f80(x86_fp80 %Val)
declare fp128      @llvm.log2.f128(fp128 %Val)
declare ppc_fp128  @llvm.log2.ppcfp128(ppc_fp128 %Val)
```

Overview:

The `'llvm.log2.*'` intrinsics compute the base-2 logarithm of the specified value.

Arguments:

The argument and return value are floating point numbers of the same type.

Semantics:

This function returns the same values as the `libm log2` functions would, and handles error conditions in the same way.

'llvm.fma.*' Intrinsic**Syntax:**

This is an overloaded intrinsic. You can use `llvm.fma` on any floating point or vector of floating point type. Not all targets support all types however.

```
declare float @llvm.fma.f32(float %a, float %b, float %c)
declare double @llvm.fma.f64(double %a, double %b, double %c)
declare x86_fp80 @llvm.fma.f80(x86_fp80 %a, x86_fp80 %b, x86_fp80 %c)
declare fp128 @llvm.fma.f128(fp128 %a, fp128 %b, fp128 %c)
declare ppc_fp128 @llvm.fma.ppcf128(ppc_fp128 %a, ppc_fp128 %b, ppc_fp128 %c)
```

Overview:

The `'llvm.fma.*'` intrinsics perform the fused multiply-add operation.

Arguments:

The argument and return value are floating point numbers of the same type.

Semantics:

This function returns the same values as the `libm fma` functions would, and does not set `errno`.

`'llvm.fabs.*'` Intrinsic

Syntax:

This is an overloaded intrinsic. You can use `llvm.fabs` on any floating point or vector of floating point type. Not all targets support all types however.

```
declare float @llvm.fabs.f32(float %Val)
declare double @llvm.fabs.f64(double %Val)
declare x86_fp80 @llvm.fabs.f80(x86_fp80 %Val)
declare fp128 @llvm.fabs.f128(fp128 %Val)
declare ppc_fp128 @llvm.fabs.ppcf128(ppc_fp128 %Val)
```

Overview:

The `'llvm.fabs.*'` intrinsics return the absolute value of the operand.

Arguments:

The argument and return value are floating point numbers of the same type.

Semantics:

This function returns the same values as the `libm fabs` functions would, and handles error conditions in the same way.

'llvm.minnum.*' Intrinsic

Syntax:

This is an overloaded intrinsic. You can use `llvm.minnum` on any floating point or vector of floating point type. Not all targets support all types however.

```
declare float      @llvm.minnum.f32(float %Val0, float %Val1)
declare double     @llvm.minnum.f64(double %Val0, double %Val1)
declare x86_fp80   @llvm.minnum.f80(x86_fp80 %Val0, x86_fp80 %Val1)
declare fp128      @llvm.minnum.f128(fp128 %Val0, fp128 %Val1)
declare ppc_fp128  @llvm.minnum.ppcf128(ppc_fp128 %Val0, ppc_fp128 %Val1)
```

Overview:

The `'llvm.minnum.*'` intrinsics return the minimum of the two arguments.

Arguments:

The arguments and return value are floating point numbers of the same type.

Semantics:

Follows the IEEE-754 semantics for `minNum`, which also match for `libm's fmin`.

If either operand is a NaN, returns the other non-NaN operand. Returns NaN only if both operands are NaN. If the operands compare equal, returns a value that compares equal to both operands. This means that `fmin(+/-0.0, +/-0.0)` could return either `-0.0` or `0.0`.

'llvm.maxnum.*' Intrinsic

Syntax:

This is an overloaded intrinsic. You can use `llvm.maxnum` on any floating point or vector of floating point type. Not all targets support all types however.

```
declare float      @llvm.maxnum.f32(float %Val0, float %Val1)
declare double     @llvm.maxnum.f64(double %Val0, double %Val1)
declare x86_fp80   @llvm.maxnum.f80(x86_fp80 %Val0, x86_fp80 %Val1)
declare fp128      @llvm.maxnum.f128(fp128 %Val0, fp128 %Val1)
declare ppc_fp128  @llvm.maxnum.ppcf128(ppc_fp128 %Val0, ppc_fp128 %Val1)
```

Overview:

The `'llvm.maxnum.*'` intrinsics return the maximum of the two arguments.

Arguments:

The arguments and return value are floating point numbers of the same type.

Semantics:

Follows the IEEE-754 semantics for `maxNum`, which also match for `libm`'s `fmax`.

If either operand is a NaN, returns the other non-NaN operand. Returns NaN only if both operands are NaN. If the operands compare equal, returns a value that compares equal to both operands. This means that `fmax(+/-0.0, +/-0.0)` could return either -0.0 or 0.0.

'llvm.copysign.*' Intrinsic**Syntax:**

This is an overloaded intrinsic. You can use `llvm.copysign` on any floating point or vector of floating point type. Not all targets support all types however.

```
declare float      @llvm.copysign.f32(float %Mag, float %Sgn)
declare double     @llvm.copysign.f64(double %Mag, double %Sgn)
declare x86_fp80   @llvm.copysign.f80(x86_fp80 %Mag, x86_fp80 %Sgn)
declare fp128      @llvm.copysign.f128(fp128 %Mag, fp128 %Sgn)
declare ppc_fp128  @llvm.copysign.ppcf128(ppc_fp128 %Mag, ppc_fp128 %Sgn)
```

Overview:

The `'llvm.copysign.*'` intrinsics return a value with the magnitude of the first operand and the sign of the second operand.

Arguments:

The arguments and return value are floating point numbers of the same type.

Semantics:

This function returns the same values as the `libm` `copysign` functions would, and handles error conditions in the same way.

'llvm.floor.*' Intrinsic**Syntax:**

This is an overloaded intrinsic. You can use `llvm.floor` on any floating point or vector of floating point type. Not all targets support all types however.


```
declare float      @llvm.floor.f32(float  %Val)
declare double     @llvm.floor.f64(double %Val)
declare x86_fp80   @llvm.floor.f80(x86_fp80 %Val)
declare fp128      @llvm.floor.f128(fp128 %Val)
declare ppc_fp128  @llvm.floor.ppcf128(ppc_fp128 %Val)
```

Overview:

The ‘`llvm.floor.*`’ intrinsics return the floor of the operand.

Arguments:

The argument and return value are floating point numbers of the same type.

Semantics:

This function returns the same values as the libm `floor` functions would, and handles error conditions in the same way.

‘`llvm.ceil.*`’ Intrinsic**Syntax:**

This is an overloaded intrinsic. You can use `llvm.ceil` on any floating point or vector of floating point type. Not all targets support all types however.

```
declare float      @llvm.ceil.f32(float  %Val)
declare double     @llvm.ceil.f64(double %Val)
declare x86_fp80   @llvm.ceil.f80(x86_fp80 %Val)
declare fp128      @llvm.ceil.f128(fp128 %Val)
declare ppc_fp128  @llvm.ceil.ppcf128(ppc_fp128 %Val)
```

Overview:

The ‘`llvm.ceil.*`’ intrinsics return the ceiling of the operand.

Arguments:

The argument and return value are floating point numbers of the same type.

Semantics:

This function returns the same values as the libm `ceil` functions would, and handles error conditions in the same way.

`'llvm.trunc.*'` Intrinsic

Syntax:

This is an overloaded intrinsic. You can use `llvm.trunc` on any floating point or vector of floating point type. Not all targets support all types however.

```
declare float      @llvm.trunc.f32(float  %Val)
declare double     @llvm.trunc.f64(double %Val)
declare x86_fp80    @llvm.trunc.f80(x86_fp80 %Val)
declare fp128      @llvm.trunc.f128(fp128 %Val)
declare ppc_fp128   @llvm.trunc.ppcf128(ppc_fp128 %Val)
```

Overview:

The `'llvm.trunc.*'` intrinsics returns the operand rounded to the nearest integer not larger in magnitude than the operand.

Arguments:

The argument and return value are floating point numbers of the same type.

Semantics:

This function returns the same values as the `libm trunc` functions would, and handles error conditions in the same way.

`'llvm rint.*'` Intrinsic

Syntax:

This is an overloaded intrinsic. You can use `llvm.rint` on any floating point or vector of floating point type. Not all targets support all types however.

```
declare float      @llvm.rint.f32(float  %Val)
declare double     @llvm.rint.f64(double %Val)
declare x86_fp80    @llvm.rint.f80(x86_fp80 %Val)
declare fp128      @llvm.rint.f128(fp128 %Val)
declare ppc_fp128   @llvm.rint.ppcf128(ppc_fp128 %Val)
```

Overview:

The `'llvm.rint.*'` intrinsics returns the operand rounded to the nearest integer. It may raise an inexact floating-point exception if the operand isn't an integer.

Arguments:

The argument and return value are floating point numbers of the same type.

Semantics:

This function returns the same values as the `libm rint` functions would, and handles error conditions in the same way.

‘`llvm.nearbyint.*`’ Intrinsic**Syntax:**

This is an overloaded intrinsic. You can use `llvm.nearbyint` on any floating point or vector of floating point type. Not all targets support all types however.

```
declare float      @llvm.nearbyint.f32(float %Val)
declare double     @llvm.nearbyint.f64(double %Val)
declare x86_fp80   @llvm.nearbyint.f80(x86_fp80 %Val)
declare fp128      @llvm.nearbyint.f128(fp128 %Val)
declare ppc_fp128  @llvm.nearbyint.ppcfp128(ppc_fp128 %Val)
```

Overview:

The ‘`llvm.nearbyint.*`’ intrinsics returns the operand rounded to the nearest integer.

Arguments:

The argument and return value are floating point numbers of the same type.

Semantics:

This function returns the same values as the `libm nearbyint` functions would, and handles error conditions in the same way.

‘`llvm.round.*`’ Intrinsic**Syntax:**

This is an overloaded intrinsic. You can use `llvm.round` on any floating point or vector of floating point type. Not all targets support all types however.

```
declare float      @llvm.round.f32(float %Val)
declare double     @llvm.round.f64(double %Val)
declare x86_fp80   @llvm.round.f80(x86_fp80 %Val)
declare fp128      @llvm.round.f128(fp128 %Val)
declare ppc_fp128  @llvm.round.ppcfp128(ppc_fp128 %Val)
```

Overview:

The ‘`llvm.round.*`’ intrinsics returns the operand rounded to the nearest integer.

Arguments:

The argument and return value are floating point numbers of the same type.

Semantics:

This function returns the same values as the `libm` `round` functions would, and handles error conditions in the same way.

Bit Manipulation Intrinsics

LLVM provides intrinsics for a few important bit manipulation operations. These allow efficient code generation for some algorithms.

'llvm.bitreverse.*' Intrinsics**Syntax:**

This is an overloaded intrinsic function. You can use `bitreverse` on any integer type.

```
declare i16 @llvm.bitreverse.i16(i16 <id>)  
declare i32 @llvm.bitreverse.i32(i32 <id>)  
declare i64 @llvm.bitreverse.i64(i64 <id>)
```

Overview:

The `'llvm.bitreverse'` family of intrinsics is used to reverse the bitpattern of an integer value; for example `0b10110110` becomes `0b01101101`.

Semantics:

The `llvm.bitreverse.iN` intrinsic returns an `iN` value that has bit `M` in the input moved to bit `N-M` in the output.

'llvm.bswap.*' Intrinsics**Syntax:**

This is an overloaded intrinsic function. You can use `bswap` on any integer type that is an even number of bytes (i.e. `BitWidth % 16 == 0`).

```
declare i16 @llvm.bswap.i16(i16 <id>)  
declare i32 @llvm.bswap.i32(i32 <id>)  
declare i64 @llvm.bswap.i64(i64 <id>)
```

Overview:

The `llvm.bswap` family of intrinsics is used to byte swap integer values with an even number of bytes (positive multiple of 16 bits). These are useful for performing operations on data that is not in the target's native byte order.

Semantics:

The `llvm.bswap.i16` intrinsic returns an `i16` value that has the high and low byte of the input `i16` swapped. Similarly, the `llvm.bswap.i32` intrinsic returns an `i32` value that has the four bytes of the input `i32` swapped, so that if the input bytes are numbered 0, 1, 2, 3 then the returned `i32` will have its bytes in 3, 2, 1, 0 order. The `llvm.bswap.i48`, `llvm.bswap.i64` and other intrinsics extend this concept to additional even-byte lengths (6 bytes, 8 bytes and more, respectively).

'llvm.ctpop.*' Intrinsic**Syntax:**

This is an overloaded intrinsic. You can use `llvm.ctpop` on any integer bit width, or on any vector with integer elements. Not all targets support all bit widths or vector types, however.

```
declare i8 @llvm.ctpop.i8(i8 <src>)
declare i16 @llvm.ctpop.i16(i16 <src>)
declare i32 @llvm.ctpop.i32(i32 <src>)
declare i64 @llvm.ctpop.i64(i64 <src>)
declare i256 @llvm.ctpop.i256(i256 <src>)
declare <2 x i32> @llvm.ctpop.v2i32(<2 x i32> <src>)
```

Overview:

The `llvm.ctpop` family of intrinsics counts the number of bits set in a value.

Arguments:

The only argument is the value to be counted. The argument may be of any integer type, or a vector with integer elements. The return type must match the argument type.

Semantics:

The `llvm.ctpop` intrinsic counts the 1's in a variable, or within each element of a vector.

'llvm.ctlz.*' Intrinsic**Syntax:**

This is an overloaded intrinsic. You can use `llvm.ctlz` on any integer bit width, or any vector whose elements are integers. Not all targets support all bit widths or vector types, however.

```
declare i8 @llvm.ctlz.i8 (i8 <src>, i1 <is_zero_undef>)  
declare i16 @llvm.ctlz.i16 (i16 <src>, i1 <is_zero_undef>)  
declare i32 @llvm.ctlz.i32 (i32 <src>, i1 <is_zero_undef>)  
declare i64 @llvm.ctlz.i64 (i64 <src>, i1 <is_zero_undef>)  
declare i256 @llvm.ctlz.i256(i256 <src>, i1 <is_zero_undef>)  
declare <2 x i32> @llvm.ctlz.v2i32(<2 x i32> <src>, i1 <is_zero_undef>)
```

Overview:

The ‘`llvm.ctlz`’ family of intrinsic functions counts the number of leading zeros in a variable.

Arguments:

The first argument is the value to be counted. This argument may be of any integer type, or a vector with integer element type. The return type must match the first argument type.

The second argument must be a constant and is a flag to indicate whether the intrinsic should ensure that a zero as the first argument produces a defined result. Historically some architectures did not provide a defined result for zero values as efficiently, and many algorithms are now predicated on avoiding zero-value inputs.

Semantics:

The ‘`llvm.ctlz`’ intrinsic counts the leading (most significant) zeros in a variable, or within each element of the vector. If `src == 0` then the result is the size in bits of the type of `src` if `is_zero_undef == 0` and `undef` otherwise. For example, `llvm.ctlz(i32 2) = 30`.

‘`llvm.cttz.*`’ Intrinsic

Syntax:

This is an overloaded intrinsic. You can use `llvm.cttz` on any integer bit width, or any vector of integer elements. Not all targets support all bit widths or vector types, however.

```
declare i8 @llvm.cttz.i8 (i8 <src>, i1 <is_zero_undef>)  
declare i16 @llvm.cttz.i16 (i16 <src>, i1 <is_zero_undef>)  
declare i32 @llvm.cttz.i32 (i32 <src>, i1 <is_zero_undef>)  
declare i64 @llvm.cttz.i64 (i64 <src>, i1 <is_zero_undef>)  
declare i256 @llvm.cttz.i256(i256 <src>, i1 <is_zero_undef>)  
declare <2 x i32> @llvm.cttz.v2i32(<2 x i32> <src>, i1 <is_zero_undef>)
```

Overview:

The ‘`llvm.cttz`’ family of intrinsic functions counts the number of trailing zeros.

Arguments:

The first argument is the value to be counted. This argument may be of any integer type, or a vector with integer element type. The return type must match the first argument type.

The second argument must be a constant and is a flag to indicate whether the intrinsic should ensure that a zero as the first argument produces a defined result. Historically some architectures did not provide a defined result for zero values as efficiently, and many algorithms are now predicated on avoiding zero-value inputs.

Semantics:

The `'llvm.cttz'` intrinsic counts the trailing (least significant) zeros in a variable, or within each element of a vector. If `src == 0` then the result is the size in bits of the type of `src` if `is_zero_undef == 0` and `undef` otherwise. For example, `llvm.cttz(2) = 1`.

Arithmetic with Overflow Intrinsics

LLVM provides intrinsics for fast arithmetic overflow checking.

Each of these intrinsics returns a two-element struct. The first element of this struct contains the result of the corresponding arithmetic operation modulo 2^n , where n is the bit width of the result. Therefore, for example, the first element of the struct returned by `llvm.sadd.with.overflow.i32` is always the same as the result of a 32-bit add instruction with the same operands, where the add is *not* modified by an `nsw` or `nuw` flag.

The second element of the result is an `i1` that is 1 if the arithmetic operation overflowed and 0 otherwise. An operation overflows if, for any values of its operands A and B and for any N larger than the operands' width, $\text{ext}(A \text{ op } B) \text{ to } iN$ is not equal to $(\text{ext}(A) \text{ to } iN) \text{ op } (\text{ext}(B) \text{ to } iN)$ where `ext` is `sext` for signed overflow and `zext` for unsigned overflow, and `op` is the underlying arithmetic operation.

The behavior of these intrinsics is well-defined for all argument values.

'llvm.sadd.with.overflow.*' Intrinsics

Syntax:

This is an overloaded intrinsic. You can use `llvm.sadd.with.overflow` on any integer bit width.

```
declare {i16, i1} @llvm.sadd.with.overflow.i16(i16 %a, i16 %b)
declare {i32, i1} @llvm.sadd.with.overflow.i32(i32 %a, i32 %b)
declare {i64, i1} @llvm.sadd.with.overflow.i64(i64 %a, i64 %b)
```

Overview:

The `'llvm.sadd.with.overflow'` family of intrinsic functions perform a signed addition of the two arguments, and indicate whether an overflow occurred during the signed summation.

Arguments:

The arguments (`%a` and `%b`) and the first element of the result structure may be of integer types of any bit width, but they must have the same bit width. The second element of the result structure must be of type `i1`. `%a` and `%b` are the two values that will undergo signed addition.

Semantics:

The ‘`llvm.sadd.with.overflow`’ family of intrinsic functions perform a signed addition of the two variables. They return a structure — the first element of which is the signed summation, and the second element of which is a bit specifying if the signed summation resulted in an overflow.

Examples:

```
%res = call {i32, i1} @llvm.sadd.with.overflow.i32(i32 %a, i32 %b)
%sum = extractvalue {i32, i1} %res, 0
%obit = extractvalue {i32, i1} %res, 1
br i1 %obit, label %overflow, label %normal
```

‘`llvm.uadd.with.overflow.*`’ Intrinsics

Syntax:

This is an overloaded intrinsic. You can use `llvm.uadd.with.overflow` on any integer bit width.

```
declare {i16, i1} @llvm.uadd.with.overflow.i16(i16 %a, i16 %b)
declare {i32, i1} @llvm.uadd.with.overflow.i32(i32 %a, i32 %b)
declare {i64, i1} @llvm.uadd.with.overflow.i64(i64 %a, i64 %b)
```

Overview:

The ‘`llvm.uadd.with.overflow`’ family of intrinsic functions perform an unsigned addition of the two arguments, and indicate whether a carry occurred during the unsigned summation.

Arguments:

The arguments (`%a` and `%b`) and the first element of the result structure may be of integer types of any bit width, but they must have the same bit width. The second element of the result structure must be of type `i1`. `%a` and `%b` are the two values that will undergo unsigned addition.

Semantics:

The ‘`llvm.uadd.with.overflow`’ family of intrinsic functions perform an unsigned addition of the two arguments. They return a structure — the first element of which is the sum, and the second element of which is a bit specifying if the unsigned summation resulted in a carry.

Examples:

```
%res = call {i32, i1} @llvm.uadd.with.overflow.i32(i32 %a, i32 %b)
%sum = extractvalue {i32, i1} %res, 0
%obit = extractvalue {i32, i1} %res, 1
br i1 %obit, label %carry, label %normal
```


`'llvm.ssub.with.overflow.*'` Intrinsics

Syntax:

This is an overloaded intrinsic. You can use `llvm.ssub.with.overflow` on any integer bit width.

```
declare {i16, i1} @llvm.ssub.with.overflow.i16(i16 %a, i16 %b)
declare {i32, i1} @llvm.ssub.with.overflow.i32(i32 %a, i32 %b)
declare {i64, i1} @llvm.ssub.with.overflow.i64(i64 %a, i64 %b)
```

Overview:

The `'llvm.ssub.with.overflow'` family of intrinsic functions perform a signed subtraction of the two arguments, and indicate whether an overflow occurred during the subtraction.

Arguments:

The arguments (`%a` and `%b`) and the first element of the result structure may be of integer types of any bit width, but they must have the same bit width. The second element of the result structure must be of type `i1`. `%a` and `%b` are the two values that will undergo signed subtraction.

Semantics:

The `'llvm.ssub.with.overflow'` family of intrinsic functions perform a signed subtraction of the two arguments. They return a structure — the first element of which is the subtraction, and the second element of which is a bit specifying if the signed subtraction resulted in an overflow.

Examples:

```
%res = call {i32, i1} @llvm.ssub.with.overflow.i32(i32 %a, i32 %b)
%sum = extractvalue {i32, i1} %res, 0
%obit = extractvalue {i32, i1} %res, 1
br i1 %obit, label %overflow, label %normal
```

`'llvm.usub.with.overflow.*'` Intrinsics

Syntax:

This is an overloaded intrinsic. You can use `llvm.usub.with.overflow` on any integer bit width.

```
declare {i16, i1} @llvm.usub.with.overflow.i16(i16 %a, i16 %b)
declare {i32, i1} @llvm.usub.with.overflow.i32(i32 %a, i32 %b)
declare {i64, i1} @llvm.usub.with.overflow.i64(i64 %a, i64 %b)
```

Overview:

The `'llvm.usub.with.overflow'` family of intrinsic functions perform an unsigned subtraction of the two arguments, and indicate whether an overflow occurred during the unsigned subtraction.

Arguments:

The arguments (`%a` and `%b`) and the first element of the result structure may be of integer types of any bit width, but they must have the same bit width. The second element of the result structure must be of type `i1`. `%a` and `%b` are the two values that will undergo unsigned subtraction.

Semantics:

The `'llvm.usub.with.overflow'` family of intrinsic functions perform an unsigned subtraction of the two arguments. They return a structure — the first element of which is the subtraction, and the second element of which is a bit specifying if the unsigned subtraction resulted in an overflow.

Examples:

```
%res = call {i32, i1} @llvm.usub.with.overflow.i32(i32 %a, i32 %b)
%sum = extractvalue {i32, i1} %res, 0
%obit = extractvalue {i32, i1} %res, 1
br i1 %obit, label %overflow, label %normal
```

'llvm.smul.with.overflow.*' Intrinsics

Syntax:

This is an overloaded intrinsic. You can use `llvm.smul.with.overflow` on any integer bit width.

```
declare {i16, i1} @llvm.smul.with.overflow.i16(i16 %a, i16 %b)
declare {i32, i1} @llvm.smul.with.overflow.i32(i32 %a, i32 %b)
declare {i64, i1} @llvm.smul.with.overflow.i64(i64 %a, i64 %b)
```

Overview:

The `'llvm.smul.with.overflow'` family of intrinsic functions perform a signed multiplication of the two arguments, and indicate whether an overflow occurred during the signed multiplication.

Arguments:

The arguments (`%a` and `%b`) and the first element of the result structure may be of integer types of any bit width, but they must have the same bit width. The second element of the result structure must be of type `i1`. `%a` and `%b` are the two values that will undergo signed multiplication.

Semantics:

The ‘`llvm.smul.with.overflow`’ family of intrinsic functions perform a signed multiplication of the two arguments. They return a structure — the first element of which is the multiplication, and the second element of which is a bit specifying if the signed multiplication resulted in an overflow.

Examples:

```
%res = call {i32, i1} @llvm.smul.with.overflow.i32(i32 %a, i32 %b)
%sum = extractvalue {i32, i1} %res, 0
%obit = extractvalue {i32, i1} %res, 1
br i1 %obit, label %overflow, label %normal
```

‘`llvm.umul.with.overflow.*`’ Intrinsics

Syntax:

This is an overloaded intrinsic. You can use `llvm.umul.with.overflow` on any integer bit width.

```
declare {i16, i1} @llvm.umul.with.overflow.i16(i16 %a, i16 %b)
declare {i32, i1} @llvm.umul.with.overflow.i32(i32 %a, i32 %b)
declare {i64, i1} @llvm.umul.with.overflow.i64(i64 %a, i64 %b)
```

Overview:

The ‘`llvm.umul.with.overflow`’ family of intrinsic functions perform a unsigned multiplication of the two arguments, and indicate whether an overflow occurred during the unsigned multiplication.

Arguments:

The arguments (`%a` and `%b`) and the first element of the result structure may be of integer types of any bit width, but they must have the same bit width. The second element of the result structure must be of type `i1`. `%a` and `%b` are the two values that will undergo unsigned multiplication.

Semantics:

The ‘`llvm.umul.with.overflow`’ family of intrinsic functions perform an unsigned multiplication of the two arguments. They return a structure — the first element of which is the multiplication, and the second element of which is a bit specifying if the unsigned multiplication resulted in an overflow.

Examples:

```
%res = call {i32, i1} @llvm.umul.with.overflow.i32(i32 %a, i32 %b)
%sum = extractvalue {i32, i1} %res, 0
%obit = extractvalue {i32, i1} %res, 1
br i1 %obit, label %overflow, label %normal
```

Specialised Arithmetic Intrinsics

'llvm.canonicalize.*' Intrinsic

Syntax:

```
declare float @llvm.canonicalize.f32(float %a)
declare double @llvm.canonicalize.f64(double %b)
```

Overview:

The 'llvm.canonicalize.*' intrinsic returns the platform specific canonical encoding of a floating point number. This canonicalization is useful for implementing certain numeric primitives such as frexp. The canonical encoding is defined by IEEE-754-2008 to be:

```
2.1.8 canonical encoding: The preferred encoding of a floating-point
representation in a format. Applied to declets, significands of finite
numbers, infinities, and NaNs, especially in decimal formats.
```

This operation can also be considered equivalent to the IEEE-754-2008 conversion of a floating-point value to the same format. NaNs are handled according to section 6.2.

Examples of non-canonical encodings:

- x87 pseudo denormals, pseudo NaNs, pseudo Infinity, Unnormals. These are converted to a canonical representation per hardware-specific protocol.
- Many normal decimal floating point numbers have non-canonical alternative encodings.
- Some machines, like GPUs or ARMv7 NEON, do not support subnormal values. These are treated as non-canonical encodings of zero and will be flushed to a zero of the same sign by this operation.

Note that per IEEE-754-2008 6.2, systems that support signaling NaNs with default exception handling must signal an invalid exception, and produce a quiet NaN result.

This function should always be implementable as multiplication by 1.0, provided that the compiler does not constant fold the operation. Likewise, division by 1.0 and `llvm.minnum(x, x)` are possible implementations. Addition with -0.0 is also sufficient provided that the rounding mode is not -Infinity.

`@llvm.canonicalize` must preserve the equality relation. That is:

- `(@llvm.canonicalize(x) == x)` is equivalent to `(x == x)`
- `(@llvm.canonicalize(x) == @llvm.canonicalize(y))` is equivalent to `(x == y)`

Additionally, the sign of zero must be conserved: `@llvm.canonicalize(-0.0) = -0.0` and `@llvm.canonicalize(+0.0) = +0.0`

The payload bits of a NaN must be conserved, with two exceptions. First, environments which use only a single canonical representation of NaN must perform said canonicalization. Second, SNaNs must be quieted per the usual methods.

The canonicalization operation may be optimized away if:

- The input is known to be canonical. For example, it was produced by a floating-point operation that is required by the standard to be canonical.
- The result is consumed only by (or fused with) other floating-point operations. That is, the bits of the floating point value are not examined.

'llvm.fmuladd.*' Intrinsic

Syntax:

```
declare float @llvm.fmuladd.f32(float %a, float %b, float %c)
declare double @llvm.fmuladd.f64(double %a, double %b, double %c)
```

Overview:

The 'llvm.fmuladd.*' intrinsic functions represent multiply-add expressions that can be fused if the code generator determines that (a) the target instruction set has support for a fused operation, and (b) that the fused operation is more efficient than the equivalent, separate pair of mul and add instructions.

Arguments:

The 'llvm.fmuladd.*' intrinsics each take three arguments: two multiplicands, a and b, and an addend c.

Semantics:

The expression:

```
%0 = call float @llvm.fmuladd.f32(%a, %b, %c)
```

is equivalent to the expression $a * b + c$, except that rounding will not be performed between the multiplication and addition steps if the code generator fuses the operations. Fusion is not guaranteed, even if the target platform supports it. If a fused multiply-add is required the corresponding `llvm.fma.*` intrinsic function should be used instead. This never sets `errno`, just as 'llvm.fma.*'.

Examples:

```
%r2 = call float @llvm.fmuladd.f32(float %a, float %b, float %c) ; yields float:r2 =
↳ (a * b) + c
```

Experimental Vector Reduction Intrinsics

Horizontal reductions of vectors can be expressed using the following intrinsics. Each one takes a vector operand as an input and applies its respective operation across all elements of the vector, returning a single scalar result of the same element type.

'llvm.experimental.vector.reduce.add.*' Intrinsic

Syntax:

```
declare i32 @llvm.experimental.vector.reduce.add.i32.v4i32(<4 x i32> %a)
declare i64 @llvm.experimental.vector.reduce.add.i64.v2i64(<2 x i64> %a)
```

Overview:

The `'llvm.experimental.vector.reduce.add.*'` intrinsics do an integer ADD reduction of a vector, returning the result as a scalar. The return type matches the element-type of the vector input.

Arguments:

The argument to this intrinsic must be a vector of integer values.

`'llvm.experimental.vector.reduce.fadd.*'` Intrinsic

Syntax:

```
declare float @llvm.experimental.vector.reduce.fadd.f32.v4f32(float %acc, <4 x float>
↳ %a)
declare double @llvm.experimental.vector.reduce.fadd.f64.v2f64(double %acc, <2 x
↳ double> %a)
```

Overview:

The `'llvm.experimental.vector.reduce.fadd.*'` intrinsics do a floating point ADD reduction of a vector, returning the result as a scalar. The return type matches the element-type of the vector input.

If the intrinsic call has fast-math flags, then the reduction will not preserve the associativity of an equivalent scalarized counterpart. If it does not have fast-math flags, then the reduction will be *ordered*, implying that the operation respects the associativity of a scalarized reduction.

Arguments:

The first argument to this intrinsic is a scalar accumulator value, which is only used when there are no fast-math flags attached. This argument may be undef when fast-math flags are used.

The second argument must be a vector of floating point values.

Examples:

```
%fast = call fast float @llvm.experimental.vector.reduce.fadd.f32.v4f32(float undef,
↳ <4 x float> %input) ; fast reduction
%ord = call float @llvm.experimental.vector.reduce.fadd.f32.v4f32(float %acc, <4 x
↳ float> %input) ; ordered reduction
```

`'llvm.experimental.vector.reduce.mul.*'` Intrinsic

Syntax:

```
declare i32 @llvm.experimental.vector.reduce.mul.i32.v4i32(<4 x i32> %a)
declare i64 @llvm.experimental.vector.reduce.mul.i64.v2i64(<2 x i64> %a)
```

Overview:

The `'llvm.experimental.vector.reduce.mul.*'` intrinsics do an integer MUL reduction of a vector, returning the result as a scalar. The return type matches the element-type of the vector input.

Arguments:

The argument to this intrinsic must be a vector of integer values.

`'llvm.experimental.vector.reduce.fmul.*'` Intrinsic**Syntax:**

```
declare float @llvm.experimental.vector.reduce.fmul.f32.v4f32(float %acc, <4 x float>
↳ %a)
declare double @llvm.experimental.vector.reduce.fmul.f64.v2f64(double %acc, <2 x
↳ double> %a)
```

Overview:

The `'llvm.experimental.vector.reduce.fmul.*'` intrinsics do a floating point MUL reduction of a vector, returning the result as a scalar. The return type matches the element-type of the vector input.

If the intrinsic call has fast-math flags, then the reduction will not preserve the associativity of an equivalent scalarized counterpart. If it does not have fast-math flags, then the reduction will be *ordered*, implying that the operation respects the associativity of a scalarized reduction.

Arguments:

The first argument to this intrinsic is a scalar accumulator value, which is only used when there are no fast-math flags attached. This argument may be undef when fast-math flags are used.

The second argument must be a vector of floating point values.

Examples:

```
%fast = call fast float @llvm.experimental.vector.reduce.fmul.f32.v4f32(float undef,
↳ <4 x float> %input) ; fast reduction
%ord = call float @llvm.experimental.vector.reduce.fmul.f32.v4f32(float %acc, <4 x
↳ float> %input) ; ordered reduction
```

`'llvm.experimental.vector.reduce.and.*'` Intrinsic**Syntax:**

```
declare i32 @llvm.experimental.vector.reduce.and.i32.v4i32(<4 x i32> %a)
```

Overview:

The `'llvm.experimental.vector.reduce.and.*'` intrinsics do a bitwise AND reduction of a vector, returning the result as a scalar. The return type matches the element-type of the vector input.

Arguments:

The argument to this intrinsic must be a vector of integer values.

`'llvm.experimental.vector.reduce.or.*'` Intrinsic**Syntax:**

```
declare i32 @llvm.experimental.vector.reduce.or.i32.v4i32(<4 x i32> %a)
```

Overview:

The `'llvm.experimental.vector.reduce.or.*'` intrinsics do a bitwise OR reduction of a vector, returning the result as a scalar. The return type matches the element-type of the vector input.

Arguments:

The argument to this intrinsic must be a vector of integer values.

`'llvm.experimental.vector.reduce.xor.*'` Intrinsic**Syntax:**

```
declare i32 @llvm.experimental.vector.reduce.xor.i32.v4i32(<4 x i32> %a)
```

Overview:

The `'llvm.experimental.vector.reduce.xor.*'` intrinsics do a bitwise XOR reduction of a vector, returning the result as a scalar. The return type matches the element-type of the vector input.

Arguments:

The argument to this intrinsic must be a vector of integer values.

'llvm.experimental.vector.reduce.smax.*' Intrinsic**Syntax:**

```
declare i32 @llvm.experimental.vector.reduce.smax.i32.v4i32(<4 x i32> %a)
```

Overview:

The 'llvm.experimental.vector.reduce.smax.*' intrinsics do a signed integer MAX reduction of a vector, returning the result as a scalar. The return type matches the element-type of the vector input.

Arguments:

The argument to this intrinsic must be a vector of integer values.

'llvm.experimental.vector.reduce.smin.*' Intrinsic**Syntax:**

```
declare i32 @llvm.experimental.vector.reduce.smin.i32.v4i32(<4 x i32> %a)
```

Overview:

The 'llvm.experimental.vector.reduce.smin.*' intrinsics do a signed integer MIN reduction of a vector, returning the result as a scalar. The return type matches the element-type of the vector input.

Arguments:

The argument to this intrinsic must be a vector of integer values.

'llvm.experimental.vector.reduce.umax.*' Intrinsic**Syntax:**

```
declare i32 @llvm.experimental.vector.reduce.umax.i32.v4i32(<4 x i32> %a)
```

Overview:

The 'llvm.experimental.vector.reduce.umax.*' intrinsics do an unsigned integer MAX reduction of a vector, returning the result as a scalar. The return type matches the element-type of the vector input.

Arguments:

The argument to this intrinsic must be a vector of integer values.

'llvm.experimental.vector.reduce.umin.*' Intrinsic**Syntax:**

```
declare i32 @llvm.experimental.vector.reduce.umin.i32.v4i32(<4 x i32> %a)
```

Overview:

The 'llvm.experimental.vector.reduce.umin.*' intrinsics do an unsigned integer MIN reduction of a vector, returning the result as a scalar. The return type matches the element-type of the vector input.

Arguments:

The argument to this intrinsic must be a vector of integer values.

'llvm.experimental.vector.reduce.fmax.*' Intrinsic**Syntax:**

```
declare float @llvm.experimental.vector.reduce.fmax.f32.v4f32(<4 x float> %a)
declare double @llvm.experimental.vector.reduce.fmax.f64.v2f64(<2 x double> %a)
```

Overview:

The 'llvm.experimental.vector.reduce.fmax.*' intrinsics do a floating point MAX reduction of a vector, returning the result as a scalar. The return type matches the element-type of the vector input.

If the intrinsic call has the `nnan` fast-math flag then the operation can assume that NaNs are not present in the input vector.

Arguments:

The argument to this intrinsic must be a vector of floating point values.

'llvm.experimental.vector.reduce.fmin.*' Intrinsic**Syntax:**

```
declare float @llvm.experimental.vector.reduce.fmin.f32.v4f32(<4 x float> %a)
declare double @llvm.experimental.vector.reduce.fmin.f64.v2f64(<2 x double> %a)
```

Overview:

The `'llvm.experimental.vector.reduce.fmin.*'` intrinsics do a floating point MIN reduction of a vector, returning the result as a scalar. The return type matches the element-type of the vector input.

If the intrinsic call has the `nnan` fast-math flag then the operation can assume that NaNs are not present in the input vector.

Arguments:

The argument to this intrinsic must be a vector of floating point values.

Half Precision Floating Point Intrinsics

For most target platforms, half precision floating point is a storage-only format. This means that it is a dense encoding (in memory) but does not support computation in the format.

This means that code must first load the half-precision floating point value as an `i16`, then convert it to float with `llvm.convert.from.fp16`. Computation can then be performed on the float value (including extending to double etc). To store the value back to memory, it is first converted to float if needed, then converted to `i16` with `llvm.convert.to.fp16`, then storing as an `i16` value.

'llvm.convert.to.fp16' Intrinsic

Syntax:

```
declare i16 @llvm.convert.to.fp16.f32(float %a)
declare i16 @llvm.convert.to.fp16.f64(double %a)
```

Overview:

The `'llvm.convert.to.fp16'` intrinsic function performs a conversion from a conventional floating point type to half precision floating point format.

Arguments:

The intrinsic function contains single argument - the value to be converted.

Semantics:

The `'llvm.convert.to.fp16'` intrinsic function performs a conversion from a conventional floating point format to half precision floating point format. The return value is an `i16` which contains the converted number.

Examples:

```
%res = call i16 @llvm.convert.to.fp16.f32(float %a)
store i16 %res, i16* @x, align 2
```

‘`llvm.convert.from.fp16`’ Intrinsic

Syntax:

```
declare float @llvm.convert.from.fp16.f32(i16 %a)
declare double @llvm.convert.from.fp16.f64(i16 %a)
```

Overview:

The ‘`llvm.convert.from.fp16`’ intrinsic function performs a conversion from half precision floating point format to single precision floating point format.

Arguments:

The intrinsic function contains single argument - the value to be converted.

Semantics:

The ‘`llvm.convert.from.fp16`’ intrinsic function performs a conversion from half single precision floating point format to single precision floating point format. The input half-float value is represented by an `i16` value.

Examples:

```
%a = load i16, i16* @x, align 2
%res = call float @llvm.convert.from.fp16(i16 %a)
```

Debugger Intrinsics

The LLVM debugger intrinsics (which all start with `llvm.dbg.` prefix), are described in the [LLVM Source Level Debugging](#) document.

Exception Handling Intrinsics

The LLVM exception handling intrinsics (which all start with `llvm.eh.` prefix), are described in the [LLVM Exception Handling](#) document.

Trampoline Intrinsics

These intrinsics make it possible to excise one parameter, marked with the *nest* attribute, from a function. The result is a callable function pointer lacking the *nest* parameter - the caller does not need to provide a value for it. Instead, the value to use is stored in advance in a “trampoline”, a block of memory usually allocated on the stack, which also contains code to splice the *nest* value into the argument list. This is used to implement the GCC nested function address extension.

For example, if the function is `i32 f(i8* nest %c, i32 %x, i32 %y)` then the resulting function pointer has signature `i32 (i32, i32)*`. It can be created as follows:

```
%tramp = alloca [10 x i8], align 4 ; size and alignment only correct for X86
%tramp1 = getelementptr [10 x i8], [10 x i8]* %tramp, i32 0, i32 0
call i8* @llvm.init.trampoline(i8* %tramp1, i8* bitcast (i32 (i8*, i32, i32)* @f to_
↳ i8*), i8* %nval)
%p = call i8* @llvm.adjust.trampoline(i8* %tramp1)
%fp = bitcast i8* %p to i32 (i32, i32)*
```

The call `%val = call i32 %fp(i32 %x, i32 %y)` is then equivalent to `%val = call i32 %f(i8* %nval, i32 %x, i32 %y)`.

‘`llvm.init.trampoline`’ Intrinsic

Syntax:

```
declare void @llvm.init.trampoline(i8* <tramp>, i8* <func>, i8* <nval>)
```

Overview:

This fills the memory pointed to by `tramp` with executable code, turning it into a trampoline.

Arguments:

The `llvm.init.trampoline` intrinsic takes three arguments, all pointers. The `tramp` argument must point to a sufficiently large and sufficiently aligned block of memory; this memory is written to by the intrinsic. Note that the size and the alignment are target-specific - LLVM currently provides no portable way of determining them, so a front-end that generates this intrinsic needs to have some target-specific knowledge. The `func` argument must hold a function bitcast to an `i8*`.

Semantics:

The block of memory pointed to by `tramp` is filled with target dependent code, turning it into a function. Then `tramp` needs to be passed to `llvm.adjust.trampoline` to get a pointer which can be *bitcast (to a new function) and called*. The new function’s signature is the same as that of `func` with any arguments marked with the `nest` attribute removed. At most one such `nest` argument is allowed, and it must be of pointer type. Calling the new function is equivalent to calling `func` with the same argument list, but with `nval` used for the missing `nest` argument. If, after calling `llvm.init.trampoline`, the memory pointed to by `tramp` is modified, then the effect of any later call to the returned function pointer is undefined.

‘`llvm.adjust.trampoline`’ Intrinsic

Syntax:

```
declare i8* @llvm.adjust.trampoline(i8* <tramp>)
```

Overview:

This performs any required machine-specific adjustment to the address of a trampoline (passed as `tramp`).

Arguments:

`tramp` must point to a block of memory which already has trampoline code filled in by a previous call to *llvm.init.trampoline*.

Semantics:

On some architectures the address of the code to be executed needs to be different than the address where the trampoline is actually stored. This intrinsic returns the executable address corresponding to `tramp` after performing the required machine specific adjustments. The pointer returned can then be *bitcast and executed*.

Masked Vector Load and Store Intrinsics

LLVM provides intrinsics for predicated vector load and store operations. The predicate is specified by a mask operand, which holds one bit per vector element, switching the associated vector lane on or off. The memory addresses corresponding to the “off” lanes are not accessed. When all bits of the mask are on, the intrinsic is identical to a regular vector load or store. When all bits are off, no memory is accessed.

‘`llvm.masked.load.*`’ Intrinsics**Syntax:**

This is an overloaded intrinsic. The loaded data is a vector of any integer, floating point or pointer data type.

```
declare <16 x float> @llvm.masked.load.v16f32.p0v16f32 (<16 x float>* <ptr>, i32
↳<alignment>, <16 x i1> <mask>, <16 x float> <passthru>)
declare <2 x double> @llvm.masked.load.v2f64.p0v2f64 (<2 x double>* <ptr>, i32
↳<alignment>, <2 x i1> <mask>, <2 x double> <passthru>)
;; The data is a vector of pointers to double
declare <8 x double*> @llvm.masked.load.v8p0f64.p0v8p0f64 (<8 x double*>* <ptr>,
↳i32 <alignment>, <8 x i1> <mask>, <8 x double*> <passthru>)
;; The data is a vector of function pointers
declare <8 x i32 ()*> @llvm.masked.load.v8p0f_i32f.p0v8p0f_i32f (<8 x i32 ()*>* <ptr>,
↳ i32 <alignment>, <8 x i1> <mask>, <8 x i32 ()*> <passthru>)
```

Overview:

Reads a vector from memory according to the provided mask. The mask holds a bit for each vector lane, and is used to prevent memory accesses to the masked-off lanes. The masked-off lanes in the result vector are taken from the corresponding lanes of the ‘passthru’ operand.

Arguments:

The first operand is the base pointer for the load. The second operand is the alignment of the source location. It must be a constant integer value. The third operand, mask, is a vector of boolean values with the same number of elements as the return type. The fourth is a pass-through value that is used to fill the masked-off lanes of the result. The return type, underlying type of the base pointer and the type of the ‘passthru’ operand are the same vector types.

Semantics:

The ‘llvm.masked.load’ intrinsic is designed for conditional reading of selected vector elements in a single IR operation. It is useful for targets that support vector masked loads and allows vectorizing predicated basic blocks on these targets. Other targets may support this intrinsic differently, for example by lowering it into a sequence of branches that guard scalar load operations. The result of this operation is equivalent to a regular vector load instruction followed by a ‘select’ between the loaded and the passthru values, predicated on the same mask. However, using this intrinsic prevents exceptions on memory access to masked-off lanes.

```
%res = call <16 x float> @llvm.masked.load.v16f32.p0v16f32 (<16 x float>* %ptr, i32 4,
↳ <16 x i1>%mask, <16 x float> %passthru)

;; The result of the two following instructions is identical aside from potential_
↳ memory access exception
%loadl1 = load <16 x float>, <16 x float>* %ptr, align 4
%res = select <16 x i1> %mask, <16 x float> %loadl1, <16 x float> %passthru
```

‘llvm.masked.store.*’ Intrinsics

Syntax:

This is an overloaded intrinsic. The data stored in memory is a vector of any integer, floating point or pointer data type.

```
declare void @llvm.masked.store.v8i32.p0v8i32 (<8 x i32> <value>, <8 x i32>*
↳ <ptr>, i32 <alignment>, <8 x i1> <mask>)
declare void @llvm.masked.store.v16f32.p0v16f32 (<16 x float> <value>, <16 x float>*
↳ <ptr>, i32 <alignment>, <16 x i1> <mask>)
;; The data is a vector of pointers to double
declare void @llvm.masked.store.v8p0f64.p0v8p0f64 (<8 x double*> <value>, <8 x_
↳ double*>* <ptr>, i32 <alignment>, <8 x i1> <mask>)
;; The data is a vector of function pointers
declare void @llvm.masked.store.v4p0f_i32f.p0v4p0f_i32f (<4 x i32 ()*> <value>, <4 x_
↳ i32 ()*>* <ptr>, i32 <alignment>, <4 x i1> <mask>)
```

Overview:

Writes a vector to memory according to the provided mask. The mask holds a bit for each vector lane, and is used to prevent memory accesses to the masked-off lanes.

Arguments:

The first operand is the vector value to be written to memory. The second operand is the base pointer for the store, it has the same underlying type as the value operand. The third operand is the alignment of the destination location. The fourth operand, mask, is a vector of boolean values. The types of the mask and the value operand must have the same number of vector elements.

Semantics:

The ‘`llvm.masked.store`’ intrinsic is designed for conditional writing of selected vector elements in a single IR operation. It is useful for targets that support vector masked store and allows vectorizing predicated basic blocks on these targets. Other targets may support this intrinsic differently, for example by lowering it into a sequence of branches that guard scalar store operations. The result of this operation is equivalent to a load-modify-store sequence. However, using this intrinsic prevents exceptions and data races on memory access to masked-off lanes.

```
call void @llvm.masked.store.v16f32.p0v16f32(<16 x float> %value, <16 x float>* %ptr,
→i32 4, <16 x i1> %mask)

;; The result of the following instructions is identical aside from potential data_
→races and memory access exceptions
%oldval = load <16 x float>, <16 x float>* %ptr, align 4
%res = select <16 x i1> %mask, <16 x float> %value, <16 x float> %oldval
store <16 x float> %res, <16 x float>* %ptr, align 4
```

Masked Vector Gather and Scatter Intrinsics

LLVM provides intrinsics for vector gather and scatter operations. They are similar to *Masked Vector Load and Store*, except they are designed for arbitrary memory accesses, rather than sequential memory accesses. Gather and scatter also employ a mask operand, which holds one bit per vector element, switching the associated vector lane on or off. The memory addresses corresponding to the “off” lanes are not accessed. When all bits are off, no memory is accessed.

‘`llvm.masked.gather.*`’ Intrinsics

Syntax:

This is an overloaded intrinsic. The loaded data are multiple scalar values of any integer, floating point or pointer data type gathered together into one vector.

```
declare <16 x float> @llvm.masked.gather.v16f32.v16p0f32 (<16 x float*> <ptrs>, i32
→<alignment>, <16 x i1> <mask>, <16 x float> <passthru>)
declare <2 x double> @llvm.masked.gather.v2f64.v2p1f64 (<2 x double addrspace(1)*>
→<ptrs>, i32 <alignment>, <2 x i1> <mask>, <2 x double> <passthru>)
declare <8 x float*> @llvm.masked.gather.v8p0f32.v8p0p0f32 (<8 x float**> <ptrs>, i32
→<alignment>, <8 x i1> <mask>, <8 x float*> <passthru>)
```

Overview:

Reads scalar values from arbitrary memory locations and gathers them into one vector. The memory locations are provided in the vector of pointers ‘`ptrs`’. The memory is accessed according to the provided mask. The mask holds

a bit for each vector lane, and is used to prevent memory accesses to the masked-off lanes. The masked-off lanes in the result vector are taken from the corresponding lanes of the ‘passthru’ operand.

Arguments:

The first operand is a vector of pointers which holds all memory addresses to read. The second operand is an alignment of the source addresses. It must be a constant integer value. The third operand, mask, is a vector of boolean values with the same number of elements as the return type. The fourth is a pass-through value that is used to fill the masked-off lanes of the result. The return type, underlying type of the vector of pointers and the type of the ‘passthru’ operand are the same vector types.

Semantics:

The ‘llvm.masked.gather’ intrinsic is designed for conditional reading of multiple scalar values from arbitrary memory locations in a single IR operation. It is useful for targets that support vector masked gathers and allows vectorizing basic blocks with data and control divergence. Other targets may support this intrinsic differently, for example by lowering it into a sequence of scalar load operations. The semantics of this operation are equivalent to a sequence of conditional scalar loads with subsequent gathering all loaded values into a single vector. The mask restricts memory access to certain lanes and facilitates vectorization of predicated basic blocks.

```
%res = call <4 x double> @llvm.masked.gather.v4f64.v4p0f64 (<4 x double*> %ptrs, i32_
→8, <4 x i1> <i1 true, i1 true, i1 true, i1 true>, <4 x double> undef)

;; The gather with all-true mask is equivalent to the following instruction sequence
%ptr0 = extractelement <4 x double*> %ptrs, i32 0
%ptr1 = extractelement <4 x double*> %ptrs, i32 1
%ptr2 = extractelement <4 x double*> %ptrs, i32 2
%ptr3 = extractelement <4 x double*> %ptrs, i32 3

%val0 = load double, double* %ptr0, align 8
%val1 = load double, double* %ptr1, align 8
%val2 = load double, double* %ptr2, align 8
%val3 = load double, double* %ptr3, align 8

%vec0   = insertelement <4 x double>undef, %val0, 0
%vec01  = insertelement <4 x double>%vec0, %val1, 1
%vec012 = insertelement <4 x double>%vec01, %val2, 2
%vec0123 = insertelement <4 x double>%vec012, %val3, 3
```

‘llvm.masked.scatter.*’ Intrinsics

Syntax:

This is an overloaded intrinsic. The data stored in memory is a vector of any integer, floating point or pointer data type. Each vector element is stored in an arbitrary memory address. Scatter with overlapping addresses is guaranteed to be ordered from least-significant to most-significant element.

```
declare void @llvm.masked.scatter.v8i32.v8p0i32 (<8 x i32> <value>, <8 x i32*>
→ <ptrs>, i32 <alignment>, <8 x i1> <mask>)
declare void @llvm.masked.scatter.v16f32.v16p1f32 (<16 x float> <value>, <16 x_
→float> addrspace(1)*> <ptrs>, i32 <alignment>, <16 x i1> <mask>)
declare void @llvm.masked.scatter.v4p0f64.v4p0p0f64 (<4 x double*> <value>, <4 x_
→double*> <ptrs>, i32 <alignment>, <4 x i1> <mask>)
```

Overview:

Writes each element from the value vector to the corresponding memory address. The memory addresses are represented as a vector of pointers. Writing is done according to the provided mask. The mask holds a bit for each vector lane, and is used to prevent memory accesses to the masked-off lanes.

Arguments:

The first operand is a vector value to be written to memory. The second operand is a vector of pointers, pointing to where the value elements should be stored. It has the same underlying type as the value operand. The third operand is an alignment of the destination addresses. The fourth operand, mask, is a vector of boolean values. The types of the mask and the value operand must have the same number of vector elements.

Semantics:

The ‘`llvm.masked.scatter`’ intrinsic is designed for writing selected vector elements to arbitrary memory addresses in a single IR operation. The operation may be conditional, when not all bits in the mask are switched on. It is useful for targets that support vector masked scatter and allows vectorizing basic blocks with data and control divergence. Other targets may support this intrinsic differently, for example by lowering it into a sequence of branches that guard scalar store operations.

```
;; This instruction unconditionally stores data vector in multiple addresses
call @llvm.masked.scatter.v8i32.v8p0i32 (<8 x i32> %value, <8 x i32*> %ptrs, i32 4,
    ↪<8 x i1> <true, true, .. true>)

;; It is equivalent to a list of scalar stores
%val0 = extractelement <8 x i32> %value, i32 0
%val1 = extractelement <8 x i32> %value, i32 1
..
%val7 = extractelement <8 x i32> %value, i32 7
%ptr0 = extractelement <8 x i32*> %ptrs, i32 0
%ptr1 = extractelement <8 x i32*> %ptrs, i32 1
..
%ptr7 = extractelement <8 x i32*> %ptrs, i32 7
;; Note: the order of the following stores is important when they overlap:
store i32 %val0, i32* %ptr0, align 4
store i32 %val1, i32* %ptr1, align 4
..
store i32 %val7, i32* %ptr7, align 4
```

Memory Use Markers

This class of intrinsics provides information about the lifetime of memory objects and ranges where variables are immutable.

`'llvm.lifetime.start'` Intrinsic

Syntax:

```
declare void @llvm.lifetime.start(i64 <size>, i8* nocapture <ptr>)
```

Overview:

The `'llvm.lifetime.start'` intrinsic specifies the start of a memory object's lifetime.

Arguments:

The first argument is a constant integer representing the size of the object, or -1 if it is variable sized. The second argument is a pointer to the object.

Semantics:

This intrinsic indicates that before this point in the code, the value of the memory pointed to by `ptr` is dead. This means that it is known to never be used and has an undefined value. A load from the pointer that precedes this intrinsic can be replaced with `'undef'`.

`'llvm.lifetime.end'` Intrinsic

Syntax:

```
declare void @llvm.lifetime.end(i64 <size>, i8* nocapture <ptr>)
```

Overview:

The `'llvm.lifetime.end'` intrinsic specifies the end of a memory object's lifetime.

Arguments:

The first argument is a constant integer representing the size of the object, or -1 if it is variable sized. The second argument is a pointer to the object.

Semantics:

This intrinsic indicates that after this point in the code, the value of the memory pointed to by `ptr` is dead. This means that it is known to never be used and has an undefined value. Any stores into the memory object following this intrinsic may be removed as dead.

'llvm.invariant.start' Intrinsic

Syntax:

This is an overloaded intrinsic. The memory object can belong to any address space.

```
declare {}* @llvm.invariant.start.p0i8(i64 <size>, i8* nocapture <ptr>)
```

Overview:

The 'llvm.invariant.start' intrinsic specifies that the contents of a memory object will not change.

Arguments:

The first argument is a constant integer representing the size of the object, or -1 if it is variable sized. The second argument is a pointer to the object.

Semantics:

This intrinsic indicates that until an `llvm.invariant.end` that uses the return value, the referenced memory location is constant and unchanging.

'llvm.invariant.end' Intrinsic

Syntax:

This is an overloaded intrinsic. The memory object can belong to any address space.

```
declare void @llvm.invariant.end.p0i8({}* <start>, i64 <size>, i8* nocapture <ptr>)
```

Overview:

The 'llvm.invariant.end' intrinsic specifies that the contents of a memory object are mutable.

Arguments:

The first argument is the matching `llvm.invariant.start` intrinsic. The second argument is a constant integer representing the size of the object, or -1 if it is variable sized and the third argument is a pointer to the object.

Semantics:

This intrinsic indicates that the memory is mutable again.

'llvm.invariant.group.barrier' Intrinsic

Syntax:

```
declare i8* @llvm.invariant.group.barrier(i8* <ptr>)
```

Overview:

The 'llvm.invariant.group.barrier' intrinsic can be used when an invariant established by invariant.group metadata no longer holds, to obtain a new pointer value that does not carry the invariant information.

Arguments:

The llvm.invariant.group.barrier takes only one argument, which is the pointer to the memory for which the invariant.group no longer holds.

Semantics:

Returns another pointer that aliases its argument but which is considered different for the purposes of load/store invariant.group metadata.

Constrained Floating Point Intrinsics

These intrinsics are used to provide special handling of floating point operations when specific rounding mode or floating point exception behavior is required. By default, LLVM optimization passes assume that the rounding mode is round-to-nearest and that floating point exceptions will not be monitored. Constrained FP intrinsics are used to support non-default rounding modes and accurately preserve exception behavior without compromising LLVM's ability to optimize FP code when the default behavior is used.

Each of these intrinsics corresponds to a normal floating point operation. The first two arguments and the return value are the same as the corresponding FP operation.

The third argument is a metadata argument specifying the rounding mode to be assumed. This argument must be one of the following strings:

```
"round.dynamic"
"round.tonearest"
"round.downward"
"round.upward"
"round.towardzero"
```

If this argument is "round.dynamic" optimization passes must assume that the rounding mode is unknown and may change at runtime. No transformations that depend on rounding mode may be performed in this case.

The other possible values for the rounding mode argument correspond to the similarly named IEEE rounding modes. If the argument is any of these values optimization passes may perform transformations as long as they are consistent with the specified rounding mode.

For example, 'x-0'-'>'x' is not a valid transformation if the rounding mode is "round.downward" or "round.dynamic" because if the value of 'x' is +0 then 'x-0' should evaluate to '-0' when rounding downward. However, this transformation is legal for all other rounding modes.

For values other than “round.dynamic” optimization passes may assume that the actual runtime rounding mode (as defined in a target-specific manner) matches the specified rounding mode, but this is not guaranteed. Using a specific non-dynamic rounding mode which does not match the actual rounding mode at runtime results in undefined behavior.

The fourth argument to the constrained floating point intrinsics specifies the required exception behavior. This argument must be one of the following strings:

```
"fpexcept.ignore"  
"fpexcept.maytrap"  
"fpexcept.strict"
```

If this argument is “fpexcept.ignore” optimization passes may assume that the exception status flags will not be read and that floating point exceptions will be masked. This allows transformations to be performed that may change the exception semantics of the original code. For example, FP operations may be speculatively executed in this case whereas they must not be for either of the other possible values of this argument.

If the exception behavior argument is “fpexcept.maytrap” optimization passes must avoid transformations that may raise exceptions that would not have been raised by the original code (such as speculatively executing FP operations), but passes are not required to preserve all exceptions that are implied by the original code. For example, exceptions may be potentially hidden by constant folding.

If the exception behavior argument is “fpexcept.strict” all transformations must strictly preserve the floating point exception semantics of the original code. Any FP exception that would have been raised by the original code must be raised by the transformed code, and the transformed code must not raise any FP exceptions that would not have been raised by the original code. This is the exception behavior argument that will be used if the code being compiled reads the FP exception status flags, but this mode can also be used with code that unmask FP exceptions.

The number and order of floating point exceptions is NOT guaranteed. For example, a series of FP operations that each may raise exceptions may be vectorized into a single instruction that raises each unique exception a single time.

‘llvm.experimental.constrained.fadd’ Intrinsic

Syntax:

```
declare <type>  
@llvm.experimental.constrained.fadd(<type> <op1>, <type> <op2>,  
                                     metadata <rounding mode>,  
                                     metadata <exception behavior>)
```

Overview:

The ‘llvm.experimental.constrained.fadd’ intrinsic returns the sum of its two operands.

Arguments:

The first two arguments to the ‘llvm.experimental.constrained.fadd’ intrinsic must be *floating point* or *vector* of floating point values. Both arguments must have identical types.

The third and fourth arguments specify the rounding mode and exception behavior as described above.

Semantics:

The value produced is the floating point sum of the two value operands and has the same type as the operands.

'llvm.experimental.constrained.fsub' Intrinsic**Syntax:**

```
declare <type>
@llvm.experimental.constrained.fsub(<type> <op1>, <type> <op2>,
                                     metadata <rounding mode>,
                                     metadata <exception behavior>)
```

Overview:

The 'llvm.experimental.constrained.fsub' intrinsic returns the difference of its two operands.

Arguments:

The first two arguments to the 'llvm.experimental.constrained.fsub' intrinsic must be *floating point* or *vector* of floating point values. Both arguments must have identical types.

The third and fourth arguments specify the rounding mode and exception behavior as described above.

Semantics:

The value produced is the floating point difference of the two value operands and has the same type as the operands.

'llvm.experimental.constrained.fmul' Intrinsic**Syntax:**

```
declare <type>
@llvm.experimental.constrained.fmul(<type> <op1>, <type> <op2>,
                                     metadata <rounding mode>,
                                     metadata <exception behavior>)
```

Overview:

The 'llvm.experimental.constrained.fmul' intrinsic returns the product of its two operands.

Arguments:

The first two arguments to the 'llvm.experimental.constrained.fmul' intrinsic must be *floating point* or *vector* of floating point values. Both arguments must have identical types.

The third and fourth arguments specify the rounding mode and exception behavior as described above.

Semantics:

The value produced is the floating point product of the two value operands and has the same type as the operands.

'llvm.experimental.constrained.fdiv' Intrinsic**Syntax:**

```
declare <type>
@llvm.experimental.constrained.fdiv(<type> <op1>, <type> <op2>,
                                     metadata <rounding mode>,
                                     metadata <exception behavior>)
```

Overview:

The 'llvm.experimental.constrained.fdiv' intrinsic returns the quotient of its two operands.

Arguments:

The first two arguments to the 'llvm.experimental.constrained.fdiv' intrinsic must be *floating point* or *vector* of floating point values. Both arguments must have identical types.

The third and fourth arguments specify the rounding mode and exception behavior as described above.

Semantics:

The value produced is the floating point quotient of the two value operands and has the same type as the operands.

'llvm.experimental.constrained.frem' Intrinsic**Syntax:**

```
declare <type>
@llvm.experimental.constrained.frem(<type> <op1>, <type> <op2>,
                                     metadata <rounding mode>,
                                     metadata <exception behavior>)
```

Overview:

The 'llvm.experimental.constrained.frem' intrinsic returns the remainder from the division of its two operands.

Arguments:

The first two arguments to the `'llvm.experimental.constrained.frem'` intrinsic must be *floating point* or *vector* of floating point values. Both arguments must have identical types.

The third and fourth arguments specify the rounding mode and exception behavior as described above. The rounding mode argument has no effect, since the result of `frem` is never rounded, but the argument is included for consistency with the other constrained floating point intrinsics.

Semantics:

The value produced is the floating point remainder from the division of the two value operands and has the same type as the operands. The remainder has the same sign as the dividend.

Constrained libm-equivalent Intrinsics

In addition to the basic floating point operations for which constrained intrinsics are described above, there are constrained versions of various operations which provide equivalent behavior to a corresponding libm function. These intrinsics allow the precise behavior of these operations with respect to rounding mode and exception behavior to be controlled.

As with the basic constrained floating point intrinsics, the rounding mode and exception behavior arguments only control the behavior of the optimizer. They do not change the runtime floating point environment.

'llvm.experimental.constrained.sqrt' Intrinsic

Syntax:

```
declare <type>
@llvm.experimental.constrained.sqrt(<type> <op1>,
                                     metadata <rounding mode>,
                                     metadata <exception behavior>)
```

Overview:

The `'llvm.experimental.constrained.sqrt'` intrinsic returns the square root of the specified value, returning the same value as the libm `'sqrt'` functions would, but without setting `errno`.

Arguments:

The first argument and the return type are floating point numbers of the same type.

The second and third arguments specify the rounding mode and exception behavior as described above.

Semantics:

This function returns the nonnegative square root of the specified value. If the value is less than negative zero, a floating point exception occurs and the return value is architecture specific.

`'llvm.experimental.constrained.pow'` Intrinsic

Syntax:

```
declare <type>
@llvm.experimental.constrained.pow(<type> <op1>, <type> <op2>,
                                   metadata <rounding mode>,
                                   metadata <exception behavior>)
```

Overview:

The `'llvm.experimental.constrained.pow'` intrinsic returns the first operand raised to the (positive or negative) power specified by the second operand.

Arguments:

The first two arguments and the return value are floating point numbers of the same type. The second argument specifies the power to which the first argument should be raised.

The third and fourth arguments specify the rounding mode and exception behavior as described above.

Semantics:

This function returns the first value raised to the second power, returning the same values as the `libm pow` functions would, and handles error conditions in the same way.

`'llvm.experimental.constrained.powi'` Intrinsic

Syntax:

```
declare <type>
@llvm.experimental.constrained.powi(<type> <op1>, i32 <op2>,
                                   metadata <rounding mode>,
                                   metadata <exception behavior>)
```

Overview:

The `'llvm.experimental.constrained.powi'` intrinsic returns the first operand raised to the (positive or negative) power specified by the second operand. The order of evaluation of multiplications is not defined. When a vector of floating point type is used, the second argument remains a scalar integer value.

Arguments:

The first argument and the return value are floating point numbers of the same type. The second argument is a 32-bit signed integer specifying the power to which the first argument should be raised.

The third and fourth arguments specify the rounding mode and exception behavior as described above.

Semantics:

This function returns the first value raised to the second power with an unspecified sequence of rounding operations.

‘llvm.experimental.constrained.sin’ Intrinsic**Syntax:**

```
declare <type>
@llvm.experimental.constrained.sin(<type> <op1>,
                                   metadata <rounding mode>,
                                   metadata <exception behavior>)
```

Overview:

The ‘llvm.experimental.constrained.sin’ intrinsic returns the sine of the first operand.

Arguments:

The first argument and the return type are floating point numbers of the same type.

The second and third arguments specify the rounding mode and exception behavior as described above.

Semantics:

This function returns the sine of the specified operand, returning the same values as the libm `sin` functions would, and handles error conditions in the same way.

‘llvm.experimental.constrained.cos’ Intrinsic**Syntax:**

```
declare <type>
@llvm.experimental.constrained.cos(<type> <op1>,
                                   metadata <rounding mode>,
                                   metadata <exception behavior>)
```

Overview:

The ‘llvm.experimental.constrained.cos’ intrinsic returns the cosine of the first operand.

Arguments:

The first argument and the return type are floating point numbers of the same type.

The second and third arguments specify the rounding mode and exception behavior as described above.

Semantics:

This function returns the cosine of the specified operand, returning the same values as the libm `cos` functions would, and handles error conditions in the same way.

‘llvm.experimental.constrained.exp’ Intrinsic**Syntax:**

```
declare <type>
@llvm.experimental.constrained.exp(<type> <op1>,
                                   metadata <rounding mode>,
                                   metadata <exception behavior>)
```

Overview:

The ‘llvm.experimental.constrained.exp’ intrinsic computes the base-e exponential of the specified value.

Arguments:

The first argument and the return value are floating point numbers of the same type.

The second and third arguments specify the rounding mode and exception behavior as described above.

Semantics:

This function returns the same values as the libm `exp` functions would, and handles error conditions in the same way.

‘llvm.experimental.constrained.exp2’ Intrinsic**Syntax:**

```
declare <type>
@llvm.experimental.constrained.exp2(<type> <op1>,
                                   metadata <rounding mode>,
                                   metadata <exception behavior>)
```

Overview:

The ‘llvm.experimental.constrained.exp2’ intrinsic computes the base-2 exponential of the specified value.

Arguments:

The first argument and the return value are floating point numbers of the same type.

The second and third arguments specify the rounding mode and exception behavior as described above.

Semantics:

This function returns the same values as the `libm exp2` functions would, and handles error conditions in the same way.

`'llvm.experimental.constrained.log'` Intrinsic**Syntax:**

```
declare <type>
@llvm.experimental.constrained.log(<type> <op1>,
                                   metadata <rounding mode>,
                                   metadata <exception behavior>)
```

Overview:

The `'llvm.experimental.constrained.log'` intrinsic computes the base-e logarithm of the specified value.

Arguments:

The first argument and the return value are floating point numbers of the same type.

The second and third arguments specify the rounding mode and exception behavior as described above.

Semantics:

This function returns the same values as the `libm log` functions would, and handles error conditions in the same way.

`'llvm.experimental.constrained.log10'` Intrinsic**Syntax:**

```
declare <type>
@llvm.experimental.constrained.log10(<type> <op1>,
                                     metadata <rounding mode>,
                                     metadata <exception behavior>)
```

Overview:

The `'llvm.experimental.constrained.log10'` intrinsic computes the base-10 logarithm of the specified value.

Arguments:

The first argument and the return value are floating point numbers of the same type.

The second and third arguments specify the rounding mode and exception behavior as described above.

Semantics:

This function returns the same values as the `libm log10` functions would, and handles error conditions in the same way.

`'llvm.experimental.constrained.log2'` Intrinsic**Syntax:**

```
declare <type>
@llvm.experimental.constrained.log2(<type> <op1>,
                                     metadata <rounding mode>,
                                     metadata <exception behavior>)
```

Overview:

The `'llvm.experimental.constrained.log2'` intrinsic computes the base-2 logarithm of the specified value.

Arguments:

The first argument and the return value are floating point numbers of the same type.

The second and third arguments specify the rounding mode and exception behavior as described above.

Semantics:

This function returns the same values as the `libm log2` functions would, and handles error conditions in the same way.

`'llvm.experimental.constrained rint'` Intrinsic**Syntax:**

```
declare <type>
@llvm.experimental.constrained.rint(<type> <op1>,
                                     metadata <rounding mode>,
                                     metadata <exception behavior>)
```

Overview:

The `'llvm.experimental.constrained rint'` intrinsic returns the first operand rounded to the nearest integer. It may raise an inexact floating point exception if the operand is not an integer.

Arguments:

The first argument and the return value are floating point numbers of the same type.

The second and third arguments specify the rounding mode and exception behavior as described above.

Semantics:

This function returns the same values as the `libm rint` functions would, and handles error conditions in the same way. The rounding mode is described, not determined, by the rounding mode argument. The actual rounding mode is determined by the runtime floating point environment. The rounding mode argument is only intended as information to the compiler.

'llvm.experimental.constrained.nearbyint' Intrinsic**Syntax:**

```
declare <type>
@llvm.experimental.constrained.nearbyint(<type> <op1>,
                                         metadata <rounding mode>,
                                         metadata <exception behavior>)
```

Overview:

The `'llvm.experimental.constrained.nearbyint'` intrinsic returns the first operand rounded to the nearest integer. It will not raise an inexact floating point exception if the operand is not an integer.

Arguments:

The first argument and the return value are floating point numbers of the same type.

The second and third arguments specify the rounding mode and exception behavior as described above.

Semantics:

This function returns the same values as the `libm nearbyint` functions would, and handles error conditions in the same way. The rounding mode is described, not determined, by the rounding mode argument. The actual rounding mode is determined by the runtime floating point environment. The rounding mode argument is only intended as information to the compiler.

General Intrinsics

This class of intrinsics is designed to be generic and has no specific purpose.

'llvm.var.annotation' Intrinsic

Syntax:

```
declare void @llvm.var.annotation(i8* <val>, i8* <str>, i8* <str>, i32 <int>)
```

Overview:

The 'llvm.var.annotation' intrinsic.

Arguments:

The first argument is a pointer to a value, the second is a pointer to a global string, the third is a pointer to a global string which is the source file name, and the last argument is the line number.

Semantics:

This intrinsic allows annotation of local variables with arbitrary strings. This can be useful for special purpose optimizations that want to look for these annotations. These have no other defined use; they are ignored by code generation and optimization.

'llvm.ptr.annotation.*' Intrinsic

Syntax:

This is an overloaded intrinsic. You can use 'llvm.ptr.annotation' on a pointer to an integer of any width. *NOTE* you must specify an address space for the pointer. The identifier for the default address space is the integer '0'.

```
declare i8* @llvm.ptr.annotation.p<address space>i8(i8* <val>, i8* <str>, i8* <str>,
↪ i32 <int>)
declare i16* @llvm.ptr.annotation.p<address space>i16(i16* <val>, i8* <str>, i8*
↪ <str>, i32 <int>)
declare i32* @llvm.ptr.annotation.p<address space>i32(i32* <val>, i8* <str>, i8*
↪ <str>, i32 <int>)
declare i64* @llvm.ptr.annotation.p<address space>i64(i64* <val>, i8* <str>, i8*
↪ <str>, i32 <int>)
declare i256* @llvm.ptr.annotation.p<address space>i256(i256* <val>, i8* <str>, i8*
↪ <str>, i32 <int>)
```

Overview:

The 'llvm.ptr.annotation' intrinsic.

Arguments:

The first argument is a pointer to an integer value of arbitrary bitwidth (result of some expression), the second is a pointer to a global string, the third is a pointer to a global string which is the source file name, and the last argument is the line number. It returns the value of the first argument.

Semantics:

This intrinsic allows annotation of a pointer to an integer with arbitrary strings. This can be useful for special purpose optimizations that want to look for these annotations. These have no other defined use; they are ignored by code generation and optimization.

'llvm.annotation.*' Intrinsic**Syntax:**

This is an overloaded intrinsic. You can use 'llvm.annotation' on any integer bit width.

```
declare i8 @llvm.annotation.i8(i8 <val>, i8* <str>, i8* <str>, i32 <int>)
declare i16 @llvm.annotation.i16(i16 <val>, i8* <str>, i8* <str>, i32 <int>)
declare i32 @llvm.annotation.i32(i32 <val>, i8* <str>, i8* <str>, i32 <int>)
declare i64 @llvm.annotation.i64(i64 <val>, i8* <str>, i8* <str>, i32 <int>)
declare i256 @llvm.annotation.i256(i256 <val>, i8* <str>, i8* <str>, i32 <int>)
```

Overview:

The 'llvm.annotation' intrinsic.

Arguments:

The first argument is an integer value (result of some expression), the second is a pointer to a global string, the third is a pointer to a global string which is the source file name, and the last argument is the line number. It returns the value of the first argument.

Semantics:

This intrinsic allows annotations to be put on arbitrary expressions with arbitrary strings. This can be useful for special purpose optimizations that want to look for these annotations. These have no other defined use; they are ignored by code generation and optimization.

'llvm.trap' Intrinsic**Syntax:**

```
declare void @llvm.trap() noreturn nounwind
```

Overview:

The `'llvm.trap'` intrinsic.

Arguments:

None.

Semantics:

This intrinsic is lowered to the target dependent trap instruction. If the target does not have a trap instruction, this intrinsic will be lowered to a call of the `abort()` function.

'llvm.debugtrap' Intrinsic**Syntax:**

```
declare void @llvm.debugtrap() nounwind
```

Overview:

The `'llvm.debugtrap'` intrinsic.

Arguments:

None.

Semantics:

This intrinsic is lowered to code which is intended to cause an execution trap with the intention of requesting the attention of a debugger.

'llvm.stackprotector' Intrinsic**Syntax:**

```
declare void @llvm.stackprotector(i8* <guard>, i8** <slot>)
```

Overview:

The `llvm.stackprotector` intrinsic takes the `guard` and stores it onto the stack at `slot`. The stack slot is adjusted to ensure that it is placed on the stack before local variables.

Arguments:

The `llvm.stackprotector` intrinsic requires two pointer arguments. The first argument is the value loaded from the stack guard `@__stack_chk_guard`. The second variable is an `alloca` that has enough space to hold the value of the guard.

Semantics:

This intrinsic causes the prologue/epilogue inserter to force the position of the `AllocaInst` stack slot to be before local variables on the stack. This is to ensure that if a local variable on the stack is overwritten, it will destroy the value of the guard. When the function exits, the guard on the stack is checked against the original guard by `llvm.stackprotectorcheck`. If they are different, then `llvm.stackprotectorcheck` causes the program to abort by calling the `__stack_chk_fail()` function.

'llvm.stackguard' Intrinsic**Syntax:**

```
declare i8* @llvm.stackguard()
```

Overview:

The `llvm.stackguard` intrinsic returns the system stack guard value.

It should not be generated by frontends, since it is only for internal usage. The reason why we create this intrinsic is that we still support IR form Stack Protector in FastISel.

Arguments:

None.

Semantics:

On some platforms, the value returned by this intrinsic remains unchanged between loads in the same thread. On other platforms, it returns the same global variable value, if any, e.g. `@__stack_chk_guard`.

Currently some platforms have IR-level customized stack guard loading (e.g. X86 Linux) that is not handled by `llvm.stackguard()`, while they should be in the future.

'llvm.objectsize' Intrinsic**Syntax:**

```
declare i32 @llvm.objectsize.i32(i8* <object>, i1 <min>, i1 <nullunknown>)
declare i64 @llvm.objectsize.i64(i8* <object>, i1 <min>, i1 <nullunknown>)
```

Overview:

The `llvm.objectsize` intrinsic is designed to provide information to the optimizers to determine at compile time whether a) an operation (like `memcpy`) will overflow a buffer that corresponds to an object, or b) that a runtime check for overflow isn't necessary. An object in this context means an allocation of a specific class, structure, array, or other object.

Arguments:

The `llvm.objectsize` intrinsic takes three arguments. The first argument is a pointer to or into the object. The second argument determines whether `llvm.objectsize` returns 0 (if true) or -1 (if false) when the object size is unknown. The third argument controls how `llvm.objectsize` acts when `null` is used as its pointer argument. If it's true and the pointer is in address space 0, `null` is treated as an opaque value with an unknown number of bytes. Otherwise, `llvm.objectsize` reports 0 bytes available when given `null`.

The second and third arguments only accept constants.

Semantics:

The `llvm.objectsize` intrinsic is lowered to a constant representing the size of the object concerned. If the size cannot be determined at compile time, `llvm.objectsize` returns `i32/i64 -1` or `0` (depending on the `min` argument).

'llvm.expect' Intrinsic

Syntax:

This is an overloaded intrinsic. You can use `llvm.expect` on any integer bit width.

```
declare i1 @llvm.expect.i1(i1 <val>, i1 <expected_val>)\ndeclare i32 @llvm.expect.i32(i32 <val>, i32 <expected_val>)\ndeclare i64 @llvm.expect.i64(i64 <val>, i64 <expected_val>)
```

Overview:

The `llvm.expect` intrinsic provides information about expected (the most probable) value of `val`, which can be used by optimizers.

Arguments:

The `llvm.expect` intrinsic takes two arguments. The first argument is a value. The second argument is an expected value, this needs to be a constant value, variables are not allowed.

Semantics:

This intrinsic is lowered to the `val`.

'llvm.assume' Intrinsic

Syntax:

```
declare void @llvm.assume(i1 %cond)
```

Overview:

The `llvm.assume` allows the optimizer to assume that the provided condition is true. This information can then be used in simplifying other parts of the code.

Arguments:

The condition which the optimizer may assume is always true.

Semantics:

The intrinsic allows the optimizer to assume that the provided condition is always true whenever the control flow reaches the intrinsic call. No code is generated for this intrinsic, and instructions that contribute only to the provided condition are not used for code generation. If the condition is violated during execution, the behavior is undefined.

Note that the optimizer might limit the transformations performed on values used by the `llvm.assume` intrinsic in order to preserve the instructions only used to form the intrinsic's input argument. This might prove undesirable if the extra information provided by the `llvm.assume` intrinsic does not cause sufficient overall improvement in code quality. For this reason, `llvm.assume` should not be used to document basic mathematical invariants that the optimizer can otherwise deduce or facts that are of little use to the optimizer.

'llvm.ssa_copy' Intrinsic

Syntax:

```
declare type @llvm.ssa_copy(type %operand) returned(1) readonly
```

Arguments:

The first argument is an operand which is used as the returned value.

Overview:

The `llvm.ssa_copy` intrinsic can be used to attach information to operations by copying them and giving them new names. For example, the `PredicateInfo` utility uses it to build Extended SSA form, and attach various forms of information to operands that dominate specific uses. It is not meant for general use, only for building temporary renaming forms that require value splits at certain points.

`'llvm.type.test'` Intrinsic

Syntax:

```
declare i1 @llvm.type.test(i8* %ptr, metadata %type) nounwind readnone
```

Arguments:

The first argument is a pointer to be tested. The second argument is a metadata object representing a *type identifier*.

Overview:

The `llvm.type.test` intrinsic tests whether the given pointer is associated with the given type identifier.

`'llvm.type.checked.load'` Intrinsic

Syntax:

```
declare {i8*, i1} @llvm.type.checked.load(i8* %ptr, i32 %offset, metadata %type) ↪ argmemonly nounwind readonly
```

Arguments:

The first argument is a pointer from which to load a function pointer. The second argument is the byte offset from which to load the function pointer. The third argument is a metadata object representing a *type identifier*.

Overview:

The `llvm.type.checked.load` intrinsic safely loads a function pointer from a virtual table pointer using type metadata. This intrinsic is used to implement control flow integrity in conjunction with virtual call optimization. The virtual call optimization pass will optimize away `llvm.type.checked.load` intrinsics associated with devirtualized calls, thereby removing the type check in cases where it is not needed to enforce the control flow integrity constraint.

If the given pointer is associated with a type metadata identifier, this function returns true as the second element of its return value. (Note that the function may also return true if the given pointer is not associated with a type metadata identifier.) If the function's return value's second element is true, the following rules apply to the first element:

- If the given pointer is associated with the given type metadata identifier, it is the function pointer loaded from the given byte offset from the given pointer.
- If the given pointer is not associated with the given type metadata identifier, it is one of the following (the choice of which is unspecified):
 1. The function pointer that would have been loaded from an arbitrarily chosen (through an unspecified mechanism) pointer associated with the type metadata.
 2. If the function has a non-void return type, a pointer to a function that returns an unspecified value without causing side effects.

If the function’s return value’s second element is false, the value of the first element is undefined.

‘`llvm.donothing`’ Intrinsic

Syntax:

```
declare void @llvm.donothing() nounwind readnone
```

Overview:

The `llvm.donothing` intrinsic doesn’t perform any operation. It’s one of only three intrinsics (besides `llvm.experimental.patchpoint` and `llvm.experimental.gc.statepoint`) that can be called with an `invoke` instruction.

Arguments:

None.

Semantics:

This intrinsic does nothing, and it’s removed by optimizers and ignored by codegen.

‘`llvm.experimental.deoptimize`’ Intrinsic

Syntax:

```
declare type @llvm.experimental.deoptimize(...) [ "deopt"(...) ]
```

Overview:

This intrinsic, together with *deoptimization operand bundles*, allow frontends to express transfer of control and frame-local state from the currently executing (typically more specialized, hence faster) version of a function into another (typically more generic, hence slower) version.

In languages with a fully integrated managed runtime like Java and JavaScript this intrinsic can be used to implement “uncommon trap” or “side exit” like functionality. In unmanaged languages like C and C++, this intrinsic can be used to represent the slow paths of specialized functions.

Arguments:

The intrinsic takes an arbitrary number of arguments, whose meaning is decided by the *lowering strategy*.

Semantics:

The `@llvm.experimental.deoptimize` intrinsic executes an attached deoptimization continuation (denoted using a *deoptimization operand bundle*) and returns the value returned by the deoptimization continuation. Defining the semantic properties of the continuation itself is out of scope of the language reference – as far as LLVM is concerned, the deoptimization continuation can invoke arbitrary side effects, including reading from and writing to the entire heap.

Deoptimization continuations expressed using "deopt" operand bundles always continue execution to the end of the physical frame containing them, so all calls to `@llvm.experimental.deoptimize` must be in “tail position”:

- `@llvm.experimental.deoptimize` cannot be invoked.
- The call must immediately precede a *ret* instruction.
- The *ret* instruction must return the value produced by the `@llvm.experimental.deoptimize` call if there is one, or void.

Note that the above restrictions imply that the return type for a call to `@llvm.experimental.deoptimize` will match the return type of its immediate caller.

The inliner composes the "deopt" continuations of the caller into the "deopt" continuations present in the inlined, and also updates calls to this intrinsic to return directly from the frame of the function it inlined into.

All declarations of `@llvm.experimental.deoptimize` must share the same calling convention.

Lowering:

Calls to `@llvm.experimental.deoptimize` are lowered to calls to the symbol `__llvm_deoptimize` (it is the frontend’s responsibility to ensure that this symbol is defined). The call arguments to `@llvm.experimental.deoptimize` are lowered as if they were formal arguments of the specified types, and not as varargs.

‘`llvm.experimental.guard`’ Intrinsic

Syntax:

```
declare void @llvm.experimental.guard(i1, ...) [ "deopt"(...) ]
```

Overview:

This intrinsic, together with *deoptimization operand bundles*, allows frontends to express guards or checks on optimistic assumptions made during compilation. The semantics of `@llvm.experimental.guard` is defined in terms of `@llvm.experimental.deoptimize` – its body is defined to be equivalent to:

```
define void @llvm.experimental.guard(i1 %pred, <args...>) {
  %realPred = and i1 %pred, undef
  br i1 %realPred, label %continue, label %leave [, !make.implicit !{}]

leave:
  call void @llvm.experimental.deoptimize(<args...>) [ "deopt"() ]
  ret void

continue:
  ret void
}
```


with the optional `[, !make.implicit !{}]` present if and only if it is present on the call site. For more details on `!make.implicit`, see [FaultMaps and implicit checks](#).

In words, `@llvm.experimental.guard` executes the attached "deopt" continuation if (but **not** only if) its first argument is false. Since the optimizer is allowed to replace the `undef` with an arbitrary value, it can optimize `guard` to fail “spuriously”, i.e. without the original condition being false (hence the “not only if”); and this allows for “check widening” type optimizations.

`@llvm.experimental.guard` cannot be invoked.

‘`llvm.load.relative`’ Intrinsic

Syntax:

```
declare i8* @llvm.load.relative.iN(i8* %ptr, iN %offset) argmemonly nounwind readonly
```

Overview:

This intrinsic loads a 32-bit value from the address `%ptr + %offset`, adds `%ptr` to that value and returns it. The constant folder specifically recognizes the form of this intrinsic and the constant initializers it may load from; if a loaded constant initializer is known to have the form `i32 trunc(x - %ptr)`, the intrinsic call is folded to `x`.

LLVM provides that the calculation of such a constant initializer will not overflow at link time under the medium code model if `x` is an `unnamed_addr` function. However, it does not provide this guarantee for a constant initializer folded into a function body. This intrinsic can be used to avoid the possibility of overflows when loading from such a constant.

Stack Map Intrinsics

LLVM provides experimental intrinsics to support runtime patching mechanisms commonly desired in dynamic language JITs. These intrinsics are described in [Stack maps and patch points in LLVM](#).

Element Wise Atomic Memory Intrinsics

These intrinsics are similar to the standard library memory intrinsics except that they perform memory transfer as a sequence of atomic memory accesses.

‘`llvm.memcpy.element.unordered.atomic`’ Intrinsic

Syntax:

This is an overloaded intrinsic. You can use `llvm.memcpy.element.unordered.atomic` on any integer bit width and for different address spaces. Not all targets support all bit widths however.

```
declare void @llvm.memcpy.element.unordered.atomic.p0i8.p0i8.i32(i8* <dest>,
                                                                i8* <src>,
                                                                i32 <len>,
                                                                i32 <element_size>)
declare void @llvm.memcpy.element.unordered.atomic.p0i8.p0i8.i64(i8* <dest>,
                                                                i8* <src>,
```

```
i64 <len>,  
i32 <element_size>)
```

Overview:

The `'llvm.memcpy.element.unordered.atomic.*'` intrinsic is a specialization of the `'llvm.memcpy.*'` intrinsic. It differs in that the `dest` and `src` are treated as arrays with elements that are exactly `element_size` bytes, and the copy between buffers uses a sequence of *unordered atomic* load/store operations that are a positive integer multiple of the `element_size` in size.

Arguments:

The first three arguments are the same as they are in the `@llvm.memcpy` intrinsic, with the added constraint that `len` is required to be a positive integer multiple of the `element_size`. If `len` is not a positive integer multiple of `element_size`, then the behaviour of the intrinsic is undefined.

`element_size` must be a compile-time constant positive power of two no greater than target-specific atomic access size limit.

For each of the input pointers `align` parameter attribute must be specified. It must be a power of two no less than the `element_size`. Caller guarantees that both the source and destination pointers are aligned to that boundary.

Semantics:

The `'llvm.memcpy.element.unordered.atomic.*'` intrinsic copies `len` bytes of memory from the source location to the destination location. These locations are not allowed to overlap. The memory copy is performed as a sequence of load/store operations where each access is guaranteed to be a multiple of `element_size` bytes wide and aligned at an `element_size` boundary.

The order of the copy is unspecified. The same value may be read from the source buffer many times, but only one write is issued to the destination buffer per element. It is well defined to have concurrent reads and writes to both source and destination provided those reads and writes are unordered atomic when specified.

This intrinsic does not provide any additional ordering guarantees over those provided by a set of unordered loads from the source location and stores to the destination.

Lowering:

In the most general case call to the `'llvm.memcpy.element.unordered.atomic.*'` is lowered to a call to the symbol `__llvm_memcpy_element_unordered_atomic_*`. Where `'*'` is replaced with an actual element size.

Optimizer is allowed to inline memory copy when it's profitable to do so.

'llvm.memmove.element.unordered.atomic' Intrinsic

Syntax:

This is an overloaded intrinsic. You can use `llvm.memmove.element.unordered.atomic` on any integer bit width and for different address spaces. Not all targets support all bit widths however.

```
declare void @llvm.memmove.element.unordered.atomic.p0i8.p0i8.i32(i8* <dest>,
                                                                    i8* <src>,
                                                                    i32 <len>,
                                                                    i32 <element_size>)
declare void @llvm.memmove.element.unordered.atomic.p0i8.p0i8.i64(i8* <dest>,
                                                                    i8* <src>,
                                                                    i64 <len>,
                                                                    i32 <element_size>)
```

Overview:

The ‘`llvm.memmove.element.unordered.atomic.*`’ intrinsic is a specialization of the ‘`llvm.memmove.*`’ intrinsic. It differs in that the `dest` and `src` are treated as arrays with elements that are exactly `element_size` bytes, and the copy between buffers uses a sequence of *unordered atomic* load/store operations that are a positive integer multiple of the `element_size` in size.

Arguments:

The first three arguments are the same as they are in the *@llvm.memmove* intrinsic, with the added constraint that `len` is required to be a positive integer multiple of the `element_size`. If `len` is not a positive integer multiple of `element_size`, then the behaviour of the intrinsic is undefined.

`element_size` must be a compile-time constant positive power of two no greater than a target-specific atomic access size limit.

For each of the input pointers the `align` parameter attribute must be specified. It must be a power of two no less than the `element_size`. Caller guarantees that both the source and destination pointers are aligned to that boundary.

Semantics:

The ‘`llvm.memmove.element.unordered.atomic.*`’ intrinsic copies `len` bytes of memory from the source location to the destination location. These locations are allowed to overlap. The memory copy is performed as a sequence of load/store operations where each access is guaranteed to be a multiple of `element_size` bytes wide and aligned at an `element_size` boundary.

The order of the copy is unspecified. The same value may be read from the source buffer many times, but only one write is issued to the destination buffer per element. It is well defined to have concurrent reads and writes to both source and destination provided those reads and writes are unordered atomic when specified.

This intrinsic does not provide any additional ordering guarantees over those provided by a set of unordered loads from the source location and stores to the destination.

Lowering:

In the most general case call to the ‘`llvm.memmove.element.unordered.atomic.*`’ is lowered to a call to the symbol `__llvm_memmove_element_unordered_atomic_*`. Where ‘`*`’ is replaced with an actual element size.

The optimizer is allowed to inline the memory copy when it’s profitable to do so.

`'llvm.memset.element.unordered.atomic'` Intrinsic

Syntax:

This is an overloaded intrinsic. You can use `llvm.memset.element.unordered.atomic` on any integer bit width and for different address spaces. Not all targets support all bit widths however.

```
declare void @llvm.memset.element.unordered.atomic.p0i8.i32(i8* <dest>,
                                                             i8 <value>,
                                                             i32 <len>,
                                                             i32 <element_size>)
declare void @llvm.memset.element.unordered.atomic.p0i8.i64(i8* <dest>,
                                                             i8 <value>,
                                                             i64 <len>,
                                                             i32 <element_size>)
```

Overview:

The `'llvm.memset.element.unordered.atomic.*'` intrinsic is a specialization of the `'llvm.memset.*'` intrinsic. It differs in that the `dest` is treated as an array with elements that are exactly `element_size` bytes, and the assignment to that array uses a sequence of *unordered atomic* store operations that are a positive integer multiple of the `element_size` in size.

Arguments:

The first three arguments are the same as they are in the `@llvm.memset` intrinsic, with the added constraint that `len` is required to be a positive integer multiple of the `element_size`. If `len` is not a positive integer multiple of `element_size`, then the behaviour of the intrinsic is undefined.

`element_size` must be a compile-time constant positive power of two no greater than target-specific atomic access size limit.

The `dest` input pointer must have the `align` parameter attribute specified. It must be a power of two no less than the `element_size`. Caller guarantees that the destination pointer is aligned to that boundary.

Semantics:

The `'llvm.memset.element.unordered.atomic.*'` intrinsic sets the `len` bytes of memory starting at the destination location to the given value. The memory is set with a sequence of store operations where each access is guaranteed to be a multiple of `element_size` bytes wide and aligned at an `element_size` boundary.

The order of the assignment is unspecified. Only one write is issued to the destination buffer per element. It is well defined to have concurrent reads and writes to the destination provided those reads and writes are unordered atomic when specified.

This intrinsic does not provide any additional ordering guarantees over those provided by a set of unordered stores to the destination.

Lowering:

In the most general case call to the `'llvm.memset.element.unordered.atomic.*'` is lowered to a call to the symbol `__llvm_memset_element_unordered_atomic_*`. Where `'*'` is replaced with an actual element

size.

The optimizer is allowed to inline the memory assignment when it's profitable to do so.

LLVM Language Reference Manual Defines the LLVM intermediate representation.

Introduction to the LLVM Compiler Presentation providing a users introduction to LLVM.

Intro to LLVM Book chapter providing a compiler hacker's introduction to LLVM.

LLVM: A Compilation Framework for Lifelong Program Analysis & Transformation Design overview.

LLVM: An Infrastructure for Multi-Stage Optimization More details (quite old now).

Publications mentioning LLVM

For those new to the LLVM system.

NOTE: If you are a user who is only interested in using LLVM-based compilers, you should look into [Clang](#) or [DragonEgg](#) instead. The documentation here is intended for users who have a need to work with the intermediate LLVM representation.

2.1 Building LLVM with CMake

- *Introduction*
- *Quick start*
- *Basic CMake usage*
- *Options and variables*
 - *Frequently-used CMake variables*
 - *LLVM-specific variables*
- *CMake Caches*
- *Executing the test suite*
- *Cross compiling*
- *Embedding LLVM in your project*
 - *Developing LLVM passes out of source*
- *Compiler/Platform-specific topics*
 - *Microsoft Visual C++*

2.1.1 Introduction

CMake is a cross-platform build-generator tool. CMake does not build the project, it generates the files needed by your build tool (GNU make, Visual Studio, etc.) for building LLVM.

If **you are a new contributor**, please start with the [Getting Started with the LLVM System](#) page. This page is geared for existing contributors moving from the legacy configure/make system.

If you are really anxious about getting a functional LLVM build, go to the [Quick start](#) section. If you are a CMake novice, start with [Basic CMake usage](#) and then go back to the [Quick start](#) section once you know what you are doing. The [Options and variables](#) section is a reference for customizing your build. If you already have experience with CMake, this is the recommended starting point.

This page is geared towards users of the LLVM CMake build. If you're looking for information about modifying the LLVM CMake build system you may want to see the [CMake Primer](#) page. It has a basic overview of the CMake language.

2.1.2 Quick start

We use here the command-line, non-interactive CMake interface.

1. [Download](#) and install CMake. Version 3.4.3 is the minimum required.
2. Open a shell. Your development tools must be reachable from this shell through the PATH environment variable.
3. Create a build directory. Building LLVM in the source directory is not supported. cd to this directory:

```
$ mkdir mybuilddir
$ cd mybuilddir
```

4. Execute this command in the shell replacing *path/to/llvm/source/root* with the path to the root of your LLVM source tree:

```
$ cmake path/to/llvm/source/root
```

CMake will detect your development environment, perform a series of tests, and generate the files required for building LLVM. CMake will use default values for all build parameters. See the [Options and variables](#) section for a list of build parameters that you can modify.

This can fail if CMake can't detect your toolset, or if it thinks that the environment is not sane enough. In this case, make sure that the toolset that you intend to use is the only one reachable from the shell, and that the shell itself is the correct one for your development environment. CMake will refuse to build MinGW makefiles if you have a POSIX shell reachable through the PATH environment variable, for instance. You can force CMake to use a given build tool; for instructions, see the [Usage](#) section, below.

5. After CMake has finished running, proceed to use IDE project files, or start the build from the build directory:

```
$ cmake --build .
```

The `--build` option tells `cmake` to invoke the underlying build tool (`make`, `ninja`, `xcodebuild`, `msbuild`, etc.)

The underlying build tool can be invoked directly, of course, but the `--build` option is portable.

6. After LLVM has finished building, install it from the build directory:

```
$ cmake --build . --target install
```


The `--target` option with `install` parameter in addition to the `--build` option tells `cmake` to build the `install` target.

It is possible to set a different install prefix at installation time by invoking the `cmake_install.cmake` script generated in the build directory:

```
$ cmake -DCMAKE_INSTALL_PREFIX=/tmp/llvm -P cmake_install.cmake
```

2.1.3 Basic CMake usage

This section explains basic aspects of CMake which you may need in your day-to-day usage.

CMake comes with extensive documentation, in the form of html files, and as online help accessible via the `cmake` executable itself. Execute `cmake --help` for further help options.

CMake allows you to specify a build tool (e.g., GNU make, Visual Studio, or Xcode). If not specified on the command line, CMake tries to guess which build tool to use, based on your environment. Once it has identified your build tool, CMake uses the corresponding *Generator* to create files for your build tool (e.g., Makefiles or Visual Studio or Xcode project files). You can explicitly specify the generator with the command line option `-G "Name of the generator"`. To see a list of the available generators on your system, execute

```
$ cmake --help
```

This will list the generator names at the end of the help text.

Generators' names are case-sensitive, and may contain spaces. For this reason, you should enter them exactly as they are listed in the `cmake --help` output, in quotes. For example, to generate project files specifically for Visual Studio 12, you can execute:

```
$ cmake -G "Visual Studio 12" path/to/llvm/source/root
```

For a given development platform there can be more than one adequate generator. If you use Visual Studio, "NMake Makefiles" is a generator you can use for building with NMake. By default, CMake chooses the most specific generator supported by your development environment. If you want an alternative generator, you must tell this to CMake with the `-G` option.

2.1.4 Options and variables

Variables customize how the build will be generated. Options are boolean variables, with possible values ON/OFF. Options and variables are defined on the CMake command line like this:

```
$ cmake -DVARIABLE=value path/to/llvm/source
```

You can set a variable after the initial CMake invocation to change its value. You can also undefine a variable:

```
$ cmake -UVARIABLE path/to/llvm/source
```

Variables are stored in the CMake cache. This is a file named `CMakeCache.txt` stored at the root of your build directory that is generated by `cmake`. Editing it yourself is not recommended.

Variables are listed in the CMake cache and later in this document with the variable name and type separated by a colon. You can also specify the variable and type on the CMake command line:

```
$ cmake -DVARIABLE:TYPE=value path/to/llvm/source
```

Frequently-used CMake variables

Here are some of the CMake variables that are used often, along with a brief explanation and LLVM-specific notes. For full documentation, consult the CMake manual, or execute `cmake --help-variable VARIABLE_NAME`.

CMAKE_BUILD_TYPE:STRING Sets the build type for make-based generators. Possible values are Release, Debug, RelWithDebInfo and MinSizeRel. If you are using an IDE such as Visual Studio, you should use the IDE settings to set the build type. Be aware that Release and RelWithDebInfo use different optimization levels on most platforms.

CMAKE_INSTALL_PREFIX:PATH Path where LLVM will be installed if “make install” is invoked or the “install” target is built.

LLVM_LIBDIR_SUFFIX:STRING Extra suffix to append to the directory where libraries are to be installed. On a 64-bit architecture, one could use `-DLLVM_LIBDIR_SUFFIX=64` to install libraries to `/usr/lib64`.

CMAKE_C_FLAGS:STRING Extra flags to use when compiling C source files.

CMAKE_CXX_FLAGS:STRING Extra flags to use when compiling C++ source files.

LLVM-specific variables

LLVM_TARGETS_TO_BUILD:STRING Semicolon-separated list of targets to build, or *all* for building all targets. Case-sensitive. Defaults to *all*. Example: `-DLLVM_TARGETS_TO_BUILD="X86;PowerPC"`.

LLVM_BUILD_TOOLS:BOOL Build LLVM tools. Defaults to ON. Targets for building each tool are generated in any case. You can build a tool separately by invoking its target. For example, you can build *llvm-as* with a Makefile-based system by executing *make llvm-as* at the root of your build directory.

LLVM_INCLUDE_TOOLS:BOOL Generate build targets for the LLVM tools. Defaults to ON. You can use this option to disable the generation of build targets for the LLVM tools.

LLVM_BUILD_EXAMPLES:BOOL Build LLVM examples. Defaults to OFF. Targets for building each example are generated in any case. See documentation for *LLVM_BUILD_TOOLS* above for more details.

LLVM_INCLUDE_EXAMPLES:BOOL Generate build targets for the LLVM examples. Defaults to ON. You can use this option to disable the generation of build targets for the LLVM examples.

LLVM_BUILD_TESTS:BOOL Build LLVM unit tests. Defaults to OFF. Targets for building each unit test are generated in any case. You can build a specific unit test using the targets defined under *unittests*, such as *ADTTests*, *IRTests*, *SupportTests*, etc. (Search for `add_llvm_unittest` in the subdirectories of *unittests* for a complete list of unit tests.) It is possible to build all unit tests with the target *UnitTests*.

LLVM_INCLUDE_TESTS:BOOL Generate build targets for the LLVM unit tests. Defaults to ON. You can use this option to disable the generation of build targets for the LLVM unit tests.

LLVM_APPEND_VC_REV:BOOL Embed version control revision info (svn revision number or Git revision id). The version info is provided by the `LLVM_REVISION` macro in `llvm/include/llvm/Support/VCSRevision.h`. Developers using git who don’t need revision info can disable this option to avoid re-linking most binaries after a branch switch. Defaults to ON.

LLVM_ENABLE_THREADS:BOOL Build with threads support, if available. Defaults to ON.

LLVM_ENABLE_CXX1Y:BOOL Build in C++1y mode, if available. Defaults to OFF.

LLVM_ENABLE_ASSERTIONS:BOOL Enables code assertions. Defaults to ON if and only if `CMAKE_BUILD_TYPE` is *Debug*.

LLVM_ENABLE_EH:BOOL Build LLVM with exception-handling support. This is necessary if you wish to link against LLVM libraries and make use of C++ exceptions in your own code that need to propagate through LLVM code. Defaults to OFF.

- LLVM_ENABLE_EXPENSIVE_CHECKS:BOOL** Enable additional time/memory expensive checking. Defaults to OFF.
- LLVM_ENABLE_PIC:BOOL** Add the `-fPIC` flag to the compiler command-line, if the compiler supports this flag. Some systems, like Windows, do not need this flag. Defaults to ON.
- LLVM_ENABLE_RTTI:BOOL** Build LLVM with run-time type information. Defaults to OFF.
- LLVM_ENABLE_WARNINGS:BOOL** Enable all compiler warnings. Defaults to ON.
- LLVM_ENABLE_PEDANTIC:BOOL** Enable pedantic mode. This disables compiler-specific extensions, if possible. Defaults to ON.
- LLVM_ENABLE_WERROR:BOOL** Stop and fail the build, if a compiler warning is triggered. Defaults to OFF.
- LLVM_ABI_BREAKING_CHECKS:STRING** Used to decide if LLVM should be built with ABI breaking checks or not. Allowed values are *WITH_ASSERTS* (default), *FORCE_ON* and *FORCE_OFF*. *WITH_ASSERTS* turns on ABI breaking checks in an assertion enabled build. *FORCE_ON* (*FORCE_OFF*) turns them on (off) irrespective of whether normal (*NDEBUG*-based) assertions are enabled or not. A version of LLVM built with ABI breaking checks is not ABI compatible with a version built without it.
- LLVM_BUILD_32_BITS:BOOL** Build 32-bit executables and libraries on 64-bit systems. This option is available only on some 64-bit Unix systems. Defaults to OFF.
- LLVM_TARGET_ARCH:STRING** LLVM target to use for native code generation. This is required for JIT generation. It defaults to “host”, meaning that it shall pick the architecture of the machine where LLVM is being built. If you are cross-compiling, set it to the target architecture name.
- LLVM_TABLEGEN:STRING** Full path to a native TableGen executable (usually named `llvm-tblgen`). This is intended for cross-compiling: if the user sets this variable, no native TableGen will be created.
- LLVM_LIT_ARGS:STRING** Arguments given to `lit`. `make check` and `make clang-test` are affected. By default, `'-sv --no-progress-bar'` on Visual C++ and Xcode, `'-sv'` on others.
- LLVM_LIT_TOOLS_DIR:PATH** The path to GnuWin32 tools for tests. Valid on Windows host. Defaults to the empty string, in which case `lit` will look for tools needed for tests (e.g. `grep`, `sort`, etc.) in your `%PATH%`. If GnuWin32 is not in your `%PATH%`, then you can set this variable to the GnuWin32 directory so that `lit` can find tools needed for tests in that directory.
- LLVM_ENABLE_FFI:BOOL** Indicates whether the LLVM Interpreter will be linked with the Foreign Function Interface library (`libffi`) in order to enable calling external functions. If the library or its headers are installed in a custom location, you can also set the variables `FFI_INCLUDE_DIR` and `FFI_LIBRARY_DIR` to the directories where `ffi.h` and `libffi.so` can be found, respectively. Defaults to OFF.
- LLVM_EXTERNAL_{CLANG,LLD,POLLY}_SOURCE_DIR:PATH** These variables specify the path to the source directory for the external LLVM projects Clang, lld, and Polly, respectively, relative to the top-level source directory. If the in-tree subdirectory for an external project exists (e.g., `llvm/tools/clang` for Clang), then the corresponding variable will not be used. If the variable for an external project does not point to a valid path, then that project will not be built.
- LLVM_ENABLE_PROJECTS:STRING** Semicolon-separated list of projects to build, or *all* for building all (clang, libcxx, libcxxabi, lldb, compiler-rt, lld, polly) projects. This flag assumes that projects are checked out side-by-side and not nested, i.e. clang needs to be in parallel of llvm instead of nested in `llvm/tools`. This feature allows to have one build for only LLVM and another for clang+llvm using the same source checkout.
- LLVM_EXTERNAL_PROJECTS:STRING** Semicolon-separated list of additional external projects to build as part of llvm. For each project `LLVM_EXTERNAL_<NAME>_SOURCE_DIR` have to be specified with the path for the source code of the project. Example: `-DLLVM_EXTERNAL_PROJECTS="Foo;Bar" -DLLVM_EXTERNAL_FOO_SOURCE_DIR=/src/foo -DLLVM_EXTERNAL_BAR_SOURCE_DIR=/src/bar`.
- LLVM_USE_OPROFILE:BOOL** Enable building OProfile JIT support. Defaults to OFF.

LLVM_PROFDATA_FILE:PATH Path to a profdata file to pass into clang's `-fprofile-instr-use` flag. This can only be specified if you're building with clang.

LLVM_USE_INTEL_JITEVENTS:BOOL Enable building support for Intel JIT Events API. Defaults to OFF.

LLVM_ENABLE_ZLIB:BOOL Enable building with zlib to support compression/uncompression in LLVM tools. Defaults to ON.

LLVM_ENABLE_DIA_SDK:BOOL Enable building with MSVC DIA SDK for PDB debugging support. Available only with MSVC. Defaults to ON.

LLVM_USE_SANITIZER:STRING Define the sanitizer used to build LLVM binaries and tests. Possible values are Address, Memory, MemoryWithOrigins, Undefined, Thread, and Address;Undefined. Defaults to empty string.

LLVM_ENABLE_LTO:STRING Add `-flto` or `-flto=` flags to the compile and link command lines, enabling link-time optimization. Possible values are Off, On, Thin and Full. Defaults to OFF.

LLVM_USE_LINKER:STRING Add `-fuse-ld={name}` to the link invocation. The possible value depend on your compiler, for clang the value can be an absolute path to your custom linker, otherwise clang will prefix the name with `ld.` and apply its usual search. For example to link LLVM with the Gold linker, cmake can be invoked with `-DLLVM_USE_LINKER=gold`.

LLVM_ENABLE_LLD:BOOL This option is equivalent to `-DLLVM_USE_LINKER=lld`, except during a 2-stage build where a dependency is added from the first stage to the second ensuring that lld is built before stage2 begins.

LLVM_PARALLEL_COMPILE_JOBS:STRING Define the maximum number of concurrent compilation jobs.

LLVM_PARALLEL_LINK_JOBS:STRING Define the maximum number of concurrent link jobs.

LLVM_BUILD_DOCS:BOOL Adds all *enabled* documentation targets (i.e. Doxygen and Sphinx targets) as dependencies of the default build targets. This results in all of the (enabled) documentation targets being as part of a normal build. If the `install` target is run then this also enables all built documentation targets to be installed. Defaults to OFF. To enable a particular documentation target, see `LLVM_ENABLE_SPHINX` and `LLVM_ENABLE_DOXYGEN`.

LLVM_ENABLE_DOXYGEN:BOOL Enables the generation of browsable HTML documentation using doxygen. Defaults to OFF.

LLVM_ENABLE_DOXYGEN_QT_HELP:BOOL Enables the generation of a Qt Compressed Help file. Defaults to OFF. This affects the make target `doxygen-llvm`. When enabled, apart from the normal HTML output generated by doxygen, this will produce a QCH file named `org.llvm.qch`. You can then load this file into Qt Creator. This option is only useful in combination with `-DLLVM_ENABLE_DOXYGEN=ON`; otherwise this has no effect.

LLVM_DOXYGEN_QCH_FILENAME:STRING The filename of the Qt Compressed Help file that will be generated when `-DLLVM_ENABLE_DOXYGEN=ON` and `-DLLVM_ENABLE_DOXYGEN_QT_HELP=ON` are given. Defaults to `org.llvm.qch`. This option is only useful in combination with `-DLLVM_ENABLE_DOXYGEN_QT_HELP=ON`; otherwise it has no effect.

LLVM_DOXYGEN_QHP_NAMESPACE:STRING Namespace under which the intermediate Qt Help Project file lives. See [Qt Help Project](#) for more information. Defaults to `"org.llvm"`. This option is only useful in combination with `-DLLVM_ENABLE_DOXYGEN_QT_HELP=ON`; otherwise it has no effect.

LLVM_DOXYGEN_QHP_CUST_FILTER_NAME:STRING See [Qt Help Project](#) for more information. Defaults to the CMake variable `${PACKAGE_STRING}` which is a combination of the package name and version string. This filter can then be used in Qt Creator to select only documentation from LLVM when browsing through all the help files that you might have loaded. This option is only useful in combination with `-DLLVM_ENABLE_DOXYGEN_QT_HELP=ON`; otherwise it has no effect.

LLVM_DOXYGEN_QHELPGENERATOR_PATH:STRING The path to the `qhelpgenerator` executable. Defaults to whatever CMake's `find_program()` can find. This option is only useful in combination with `-DLLVM_ENABLE_DOXYGEN_QT_HELP=ON`; otherwise it has no effect.

LLVM_DOXYGEN_SVG:BOOL Uses `.svg` files instead of `.png` files for graphs in the Doxygen output. Defaults to OFF.

LLVM_INSTALL_DOXYGEN_HTML_DIR:STRING The path to install Doxygen-generated HTML documentation to. This path can either be absolute or relative to the `CMAKE_INSTALL_PREFIX`. Defaults to `share/doc/llvm/doxygen-html`.

LLVM_ENABLE_SPHINX:BOOL If specified, CMake will search for the `sphinx-build` executable and will make the `SPHINX_OUTPUT_HTML` and `SPHINX_OUTPUT_MAN` CMake options available. Defaults to OFF.

SPHINX_EXECUTABLE:STRING The path to the `sphinx-build` executable detected by CMake. For installation instructions, see <http://www.sphinx-doc.org/en/latest/install.html>

SPHINX_OUTPUT_HTML:BOOL If enabled (and `LLVM_ENABLE_SPHINX` is enabled) then the targets for building the documentation as html are added (but not built by default unless `LLVM_BUILD_DOCS` is enabled). There is a target for each project in the source tree that uses sphinx (e.g. `docs-llvm-html`, `docs-clang-html` and `docs-lld-html`). Defaults to ON.

SPHINX_OUTPUT_MAN:BOOL If enabled (and `LLVM_ENABLE_SPHINX` is enabled) the targets for building the man pages are added (but not built by default unless `LLVM_BUILD_DOCS` is enabled). Currently the only target added is `docs-llvm-man`. Defaults to ON.

SPHINX_WARNINGS_AS_ERRORS:BOOL If enabled then sphinx documentation warnings will be treated as errors. Defaults to ON.

LLVM_INSTALL_SPHINX_HTML_DIR:STRING The path to install Sphinx-generated HTML documentation to. This path can either be absolute or relative to the `CMAKE_INSTALL_PREFIX`. Defaults to `share/doc/llvm/html`.

LLVM_INSTALL_OCAMLDOC_HTML_DIR:STRING The path to install OCamlDoc-generated HTML documentation to. This path can either be absolute or relative to the `CMAKE_INSTALL_PREFIX`. Defaults to `share/doc/llvm/ocaml-html`.

LLVM_CREATE_XCODE_TOOLCHAIN:BOOL OS X Only: If enabled CMake will generate a target named 'install-xcode-toolchain'. This target will create a directory at `$CMAKE_INSTALL_PREFIX/Toolchains` containing an `xctoolchain` directory which can be used to override the default system tools.

LLVM_BUILD_LLVM_DYLIB:BOOL If enabled, the target for building the libLLVM shared library is added. This library contains all of LLVM's components in a single shared library. Defaults to OFF. This cannot be used in conjunction with `BUILD_SHARED_LIBS`. Tools will only be linked to the libLLVM shared library if `LLVM_LINK_LLVM_DYLIB` is also ON. The components in the library can be customised by setting `LLVM_DYLIB_COMPONENTS` to a list of the desired components.

LLVM_LINK_LLVM_DYLIB:BOOL If enabled, tools will be linked with the libLLVM shared library. Defaults to OFF. Setting `LLVM_LINK_LLVM_DYLIB` to ON also sets `LLVM_BUILD_LLVM_DYLIB` to ON.

BUILD_SHARED_LIBS:BOOL Flag indicating if each LLVM component (e.g. Support) is built as a shared library (ON) or as a static library (OFF). Its default value is OFF. On Windows, shared libraries may be used when building with MinGW, including mingw-w64, but not when building with the Microsoft toolchain.

Note: `BUILD_SHARED_LIBS` is only recommended for use by LLVM developers. If you want to build LLVM as a shared library, you should use the `LLVM_BUILD_LLVM_DYLIB` option.

LLVM_OPTIMIZED_TABLEGEN:BOOL If enabled and building a debug or asserts build the CMake build system will generate a Release build tree to build a fully optimized tablegen for use during the build. Enabling this

option can significantly speed up build times especially when building LLVM in Debug configurations.

LLVM_REVERSE_ITERATION:BOOL If enabled, all supported unordered llvm containers would be iterated in reverse order. This is useful for uncovering non-determinism caused by iteration of unordered containers.

2.1.5 CMake Caches

Recently LLVM and Clang have been adding some more complicated build system features. Utilizing these new features often involves a complicated chain of CMake variables passed on the command line. Clang provides a collection of CMake cache scripts to make these features more approachable.

CMake cache files are utilized using CMake's `-C` flag:

```
$ cmake -C <path to cache file> <path to sources>
```

CMake cache scripts are processed in an isolated scope, only cached variables remain set when the main configuration runs. CMake cached variables do not reset variables that are already set unless the `FORCE` option is specified.

A few notes about CMake Caches:

- Order of command line arguments is important
 - `-D` arguments specified before `-C` are set before the cache is processed and can be read inside the cache file
 - `-D` arguments specified after `-C` are set after the cache is processed and are unset inside the cache file
- All `-D` arguments will override cache file settings
- `CMAKE_TOOLCHAIN_FILE` is evaluated after both the cache file and the command line arguments
- It is recommended that all `-D` options should be specified *before* `-C`

For more information about some of the advanced build configurations supported via Cache files see [Advanced Build Configurations](#).

2.1.6 Executing the test suite

Testing is performed when the *check-all* target is built. For instance, if you are using Makefiles, execute this command in the root of your build directory:

```
$ make check-all
```

On Visual Studio, you may run tests by building the project “check-all”. For more information about testing, see the [LLVM Testing Infrastructure Guide](#).

2.1.7 Cross compiling

See [this wiki page](#) for generic instructions on how to cross-compile with CMake. It goes into detailed explanations and may seem daunting, but it is not. On the wiki page there are several examples including toolchain files. Go directly to [this section](#) for a quick solution.

Also see the [LLVM-specific variables](#) section for variables used when cross-compiling.

2.1.8 Embedding LLVM in your project

From LLVM 3.5 onwards both the CMake and autoconf/Makefile build systems export LLVM libraries as importable CMake targets. This means that clients of LLVM can now reliably use CMake to develop their own LLVM-based projects against an installed version of LLVM regardless of how it was built.

Here is a simple example of a CMakeLists.txt file that imports the LLVM libraries and uses them to build a simple application simple-tool.

```
cmake_minimum_required(VERSION 3.4.3)
project(SimpleProject)

find_package(LLVM REQUIRED CONFIG)

message(STATUS "Found LLVM ${LLVM_PACKAGE_VERSION}")
message(STATUS "Using LLVMConfig.cmake in: ${LLVM_DIR}")

# Set your project compile flags.
# E.g. if using the C++ header files
# you will need to enable C++11 support
# for your compiler.

include_directories(${LLVM_INCLUDE_DIRS})
add_definitions(${LLVM_DEFINITIONS})

# Now build our tools
add_executable(simple-tool tool.cpp)

# Find the libraries that correspond to the LLVM components
# that we wish to use
llvm_map_components_to_libnames(llvm_libs support core irreader)

# Link against LLVM libraries
target_link_libraries(simple-tool ${llvm_libs})
```

The `find_package(...)` directive when used in CONFIG mode (as in the above example) will look for the `LLVMConfig.cmake` file in various locations (see `cmake` manual for details). It creates a `LLVM_DIR` cache entry to save the directory where `LLVMConfig.cmake` is found or allows the user to specify the directory (e.g. by passing `-DLLVM_DIR=/usr/lib/cmake/llvm` to the `cmake` command or by setting it directly in `ccmake` or `cmake-gui`).

This file is available in two different locations.

- `<INSTALL_PREFIX>/lib/cmake/llvm/LLVMConfig.cmake` where `<INSTALL_PREFIX>` is the install prefix of an installed version of LLVM. On Linux typically this is `/usr/lib/cmake/llvm/LLVMConfig.cmake`.
- `<LLVM_BUILD_ROOT>/lib/cmake/llvm/LLVMConfig.cmake` where `<LLVM_BUILD_ROOT>` is the root of the LLVM build tree. **Note: this is only available when building LLVM with CMake.**

If LLVM is installed in your operating system's normal installation prefix (e.g. on Linux this is usually `/usr/`) `find_package(LLVM ...)` will automatically find LLVM if it is installed correctly. If LLVM is not installed or you wish to build directly against the LLVM build tree you can use `LLVM_DIR` as previously mentioned.

The `LLVMConfig.cmake` file sets various useful variables. Notable variables include

LLVM_CMAKE_DIR The path to the LLVM CMake directory (i.e. the directory containing `LLVMConfig.cmake`).

LLVM_DEFINITIONS A list of preprocessor defines that should be used when building against LLVM.

LLVM_ENABLE_ASSERTIONS This is set to ON if LLVM was built with assertions, otherwise OFF.

LLVM_ENABLE_EH This is set to ON if LLVM was built with exception handling (EH) enabled, otherwise OFF.

LLVM_ENABLE_RTTI This is set to ON if LLVM was built with run time type information (RTTI), otherwise OFF.

LLVM_INCLUDE_DIRS A list of include paths to directories containing LLVM header files.

LLVM_PACKAGE_VERSION The LLVM version. This string can be used with CMake conditionals, e.g., if `{LLVM_PACKAGE_VERSION} VERSION_LESS "3.5"`.

LLVM_TOOLS_BINARY_DIR The path to the directory containing the LLVM tools (e.g. `llvm-as`).

Notice that in the above example we link `simple-tool` against several LLVM libraries. The list of libraries is determined by using the `llvm_map_components_to_libnames()` CMake function. For a list of available components look at the output of running `llvm-config --components`.

Note that for LLVM < 3.5 `llvm_map_components_to_libraries()` was used instead of `llvm_map_components_to_libnames()`. This is now deprecated and will be removed in a future version of LLVM.

Developing LLVM passes out of source

It is possible to develop LLVM passes out of LLVM's source tree (i.e. against an installed or built LLVM). An example of a project layout is provided below.

```
<project dir>/
|
|  CMakeLists.txt
|  <pass name>/
|  |
|  |  CMakeLists.txt
|  |  Pass.cpp
|  |  ...
```

Contents of `<project dir>/CMakeLists.txt`:

```
find_package(LLVM REQUIRED CONFIG)

add_definitions(${LLVM_DEFINITIONS})
include_directories(${LLVM_INCLUDE_DIRS})

add_subdirectory(<pass name>)
```

Contents of `<project dir>/<pass name>/CMakeLists.txt`:

```
add_library(LLVMPassname MODULE Pass.cpp)
```

Note if you intend for this pass to be merged into the LLVM source tree at some point in the future it might make more sense to use LLVM's internal `add_llvm_loadable_module` function instead by...

Adding the following to `<project dir>/CMakeLists.txt` (after `find_package(LLVM ...)`)

```
list(APPEND CMAKE_MODULE_PATH "${LLVM_CMAKE_DIR}")
include(AddLLVM)
```

And then changing `<project dir>/<pass name>/CMakeLists.txt` to

```
add_llvm_loadable_module(LLVMPassname
    Pass.cpp
)
```


When you are done developing your pass, you may wish to integrate it into the LLVM source tree. You can achieve it in two easy steps:

1. Copying `<pass name>` folder into `<LLVM root>/lib/Transform` directory.
2. Adding `add_subdirectory(<pass name>)` line into `<LLVM root>/lib/Transform/CMakeLists.txt`.

2.1.9 Compiler/Platform-specific topics

Notes for specific compilers and/or platforms.

Microsoft Visual C++

LLVM_COMPILER_JOBS:STRING Specifies the maximum number of parallel compiler jobs to use per project when building with msbuild or Visual Studio. Only supported for the Visual Studio 2010 CMake generator. 0 means use all processors. Default is 0.

2.2 CMake Primer

- *Introduction*
- *10,000 ft View*
- *Scripting Overview*
- *Variables, Types, and Scope*
 - *Dereferencing*
 - *Lists*
 - *Lists of Lists*
 - *Other Types*
 - *Scope*
- *Control Flow*
 - *If, ElseIf, Else*
 - *Loops*
- *Modules, Functions and Macros*
 - *Modules*
 - *Argument Handling*
 - *Functions Vs Macros*
- *LLVM Project Wrappers*
- *Useful Built-in Commands*

Warning: Disclaimer: This documentation is written by LLVM project contributors *not* anyone affiliated with the CMake project. This document may contain inaccurate terminology, phrasing, or technical details. It is provided with the best intentions.

2.2.1 Introduction

The LLVM project and many of the core projects built on LLVM build using CMake. This document aims to provide a brief overview of CMake for developers modifying LLVM projects or building their own projects on top of LLVM.

The official CMake language references is available in the [cmake-language manpage](#) and [cmake-language online documentation](#).

2.2.2 10,000 ft View

CMake is a tool that reads script files in its own language that describe how a software project builds. As CMake evaluates the scripts it constructs an internal representation of the software project. Once the scripts have been fully processed, if there are no errors, CMake will generate build files to actually build the project. CMake supports generating build files for a variety of command line build tools as well as for popular IDEs.

When a user runs CMake it performs a variety of checks similar to how autoconf worked historically. During the checks and the evaluation of the build description scripts CMake caches values into the CMakeCache. This is useful because it allows the build system to skip long-running checks during incremental development. CMake caching also has some drawbacks, but that will be discussed later.

2.2.3 Scripting Overview

CMake's scripting language has a very simple grammar. Every language construct is a command that matches the pattern `_name_(args_)`. Commands come in three primary types: language-defined (commands implemented in C++ in CMake), defined functions, and defined macros. The CMake distribution also contains a suite of CMake modules that contain definitions for useful functionality.

The example below is the full CMake build for building a C++ “Hello World” program. The example uses only CMake language-defined functions.

```
cmake_minimum_required(VERSION 3.2)
project(HelloWorld)
add_executable(HelloWorld HelloWorld.cpp)
```

The CMake language provides control flow constructs in the form of foreach loops and if blocks. To make the example above more complicated you could add an if block to define “APPLE” when targeting Apple platforms:

```
cmake_minimum_required(VERSION 3.2)
project(HelloWorld)
add_executable(HelloWorld HelloWorld.cpp)
if(APPLE)
    target_compile_definitions(HelloWorld PUBLIC APPLE)
endif()
```

2.2.4 Variables, Types, and Scope

Dereferencing

In CMake variables are “stringly” typed. All variables are represented as strings throughout evaluation. Wrapping a variable in `${ }` dereferences it and results in a literal substitution of the name for the value. CMake refers to this as “variable evaluation” in their documentation. Dereferences are performed *before* the command being called receives the arguments. This means dereferencing a list results in multiple separate arguments being passed to the command.

Variable dereferences can be nested and be used to model complex data. For example:

```
set(var_name var1)
set(${var_name} foo) # same as "set(var1 foo)"
set(${${var_name}}_var bar) # same as "set(foo_var bar)"
```

Dereferencing an unset variable results in an empty expansion. It is a common pattern in CMake to conditionally set variables knowing that it will be used in code paths that the variable isn’t set. There are examples of this throughout the LLVM CMake build system.

An example of variable empty expansion is:

```
if(APPLE)
    set(extra_sources Apple.cpp)
endif()
add_executable(HelloWorld HelloWorld.cpp ${extra_sources})
```

In this example the `extra_sources` variable is only defined if you’re targeting an Apple platform. For all other targets the `extra_sources` will be evaluated as empty before `add_executable` is given its arguments.

Lists

In CMake lists are semi-colon delimited strings, and it is strongly advised that you avoid using semi-colons in lists; it doesn’t go smoothly. A few examples of defining lists:

```
# Creates a list with members a, b, c, and d
set(my_list a b c d)
set(my_list "a;b;c;d")

# Creates a string "a b c d"
set(my_string "a b c d")
```

Lists of Lists

One of the more complicated patterns in CMake is lists of lists. Because a list cannot contain an element with a semi-colon to construct a list of lists you make a list of variable names that refer to other lists. For example:

```
set(list_of_lists a b c)
set(a 1 2 3)
set(b 4 5 6)
set(c 7 8 9)
```

With this layout you can iterate through the list of lists printing each value with the following code:

```
foreach(list_name IN LISTS list_of_lists)
  foreach(value IN LISTS ${list_name})
    message(${value})
  endforeach()
endforeach()
```

You'll notice that the inner `foreach` loop's list is doubly dereferenced. This is because the first dereference turns `list_name` into the name of the sub-list (a, b, or c in the example), then the second dereference is to get the value of the list.

This pattern is used throughout CMake, the most common example is the compiler flags options, which CMake refers to using the following variable expansions: `CMAKE_${LANGUAGE}_FLAGS` and `CMAKE_${LANGUAGE}_FLAGS_${CMAKE_BUILD_TYPE}`.

Other Types

Variables that are cached or specified on the command line can have types associated with them. The variable's type is used by CMake's UI tool to display the right input field. The variable's type generally doesn't impact evaluation. One of the few examples is `PATH` variables, which CMake does have some special handling for. You can read more about the special handling in [CMake's set documentation](#).

Scope

CMake inherently has a directory-based scoping. Setting a variable in a CMakeLists file, will set the variable for that file, and all subdirectories. Variables set in a CMake module that is included in a CMakeLists file will be set in the scope they are included from, and all subdirectories.

When a variable that is already set is set again in a subdirectory it overrides the value in that scope and any deeper subdirectories.

The CMake `set` command provides two scope-related options. `PARENT_SCOPE` sets a variable into the parent scope, and not the current scope. The `CACHE` option sets the variable in the CMakeCache, which results in it being set in all scopes. The `CACHE` option will not set a variable that already exists in the CACHE unless the `FORCE` option is specified.

In addition to directory-based scope, CMake functions also have their own scope. This means variables set inside functions do not bleed into the parent scope. This is not true of macros, and it is for this reason LLVM prefers functions over macros whenever reasonable.

Note: Unlike C-based languages, CMake's loop and control flow blocks do not have their own scopes.

2.2.5 Control Flow

CMake features the same basic control flow constructs you would expect in any scripting language, but there are a few quarks because, as with everything in CMake, control flow constructs are commands.

If, Elself, Else

Note: For the full documentation on the CMake `if` command go [here](#). That resource is far more complete.

In general CMake if blocks work the way you'd expect:

```
if(<condition>)
  message("do stuff")
elseif(<condition>)
  message("do other stuff")
else()
  message("do other other stuff")
endif()
```

The single most important thing to know about CMake's if blocks coming from a C background is that they do not have their own scope. Variables set inside conditional blocks persist after the `endif()`.

Loops

The most common form of the CMake `foreach` block is:

```
foreach(var ...)
  message("do stuff")
endforeach()
```

The variable argument portion of the `foreach` block can contain dereferenced lists, values to iterate, or a mix of both:

```
foreach(var foo bar baz)
  message(${var})
endforeach()
# prints:
# foo
# bar
# baz

set(my_list 1 2 3)
foreach(var ${my_list})
  message(${var})
endforeach()
# prints:
# 1
# 2
# 3

foreach(var ${my_list} out_of_bounds)
  message(${var})
endforeach()
# prints:
# 1
# 2
# 3
# out_of_bounds
```

There is also a more modern CMake `foreach` syntax. The code below is equivalent to the code above:

```
foreach(var IN ITEMS foo bar baz)
  message(${var})
endforeach()
# prints:
# foo
```

```
# bar
# baz

set(my_list 1 2 3)
foreach(var IN LISTS my_list)
  message(${var})
endforeach()
# prints:
# 1
# 2
# 3

foreach(var IN LISTS my_list ITEMS out_of_bounds)
  message(${var})
endforeach()
# prints:
# 1
# 2
# 3
# out_of_bounds
```

Similar to the conditional statements, these generally behave how you would expect, and they do not have their own scope.

CMake also supports `while` loops, although they are not widely used in LLVM.

2.2.6 Modules, Functions and Macros

Modules

Modules are CMake's vehicle for enabling code reuse. CMake modules are just CMake script files. They can contain code to execute on include as well as definitions for commands.

In CMake macros and functions are universally referred to as commands, and they are the primary method of defining code that can be called multiple times.

In LLVM we have several CMake modules that are included as part of our distribution for developers who don't build our project from source. Those modules are the fundamental pieces needed to build LLVM-based projects with CMake. We also rely on modules as a way of organizing the build system's functionality for maintainability and re-use within LLVM projects.

Argument Handling

When defining a CMake command handling arguments is very useful. The examples in this section will all use the CMake `function` block, but this all applies to the `macro` block as well.

CMake commands can have named arguments, but all commands are implicitly variable argument. If the command has named arguments they are required and must be specified at every call site. Below is a trivial example of providing a wrapper function for CMake's built in function `add_dependencies`.

```
function(add_deps target)
  add_dependencies(${target} ${ARGV})
endfunction()
```

This example defines a new macro named `add_deps` which takes a required first argument, and just calls another function passing through the first argument and all trailing arguments. When variable arguments are present CMake defines them in a list named `ARGV`, and the count of the arguments is defined in `ARGN`.

CMake provides a module `CMakeParseArguments` which provides an implementation of advanced argument parsing. We use this all over LLVM, and it is recommended for any function that has complex argument-based behaviors or optional arguments. CMake's official documentation for the module is in the `cmake-modules` manpage, and is also available at the [cmake-modules online documentation](#).

Note: As of CMake 3.5 the `cmake_parse_arguments` command has become a native command and the `CMakeParseArguments` module is empty and only left around for compatibility.

Functions Vs Macros

Functions and Macros look very similar in how they are used, but there is one fundamental difference between the two. Functions have their own scope, and macros don't. This means variables set in macros will bleed out into the calling scope. That makes macros suitable for defining very small bits of functionality only.

The other difference between CMake functions and macros is how arguments are passed. Arguments to macros are not set as variables, instead dereferences to the parameters are resolved across the macro before executing it. This can result in some unexpected behavior if using unreferenced variables. For example:

```
macro(print_list my_list)
  foreach(var IN LISTS my_list)
    message("${var}")
  endforeach()
endmacro()

set(my_list a b c d)
set(my_list_of_numbers 1 2 3 4)
print_list(my_list_of_numbers)
# prints:
# a
# b
# c
# d
```

Generally speaking this issue is uncommon because it requires using non-dereferenced variables with names that overlap in the parent scope, but it is important to be aware of because it can lead to subtle bugs.

2.2.7 LLVM Project Wrappers

LLVM projects provide lots of wrappers around critical CMake built-in commands. We use these wrappers to provide consistent behaviors across LLVM components and to reduce code duplication.

We generally (but not always) follow the convention that commands prefaced with `llvm_` are intended to be used only as building blocks for other commands. Wrapper commands that are intended for direct use are generally named following with the project in the middle of the command name (i.e. `add_llvm_executable` is the wrapper for `add_executable`). The LLVM `add_*` wrapper functions are all defined in `AddLLVM.cmake` which is installed as part of the LLVM distribution. It can be included and used by any LLVM sub-project that requires LLVM.

Note: Not all LLVM projects require LLVM for all use cases. For example `compiler-rt` can be built without LLVM,

and the compiler-rt sanitizer libraries are used with GCC.

2.2.8 Useful Built-in Commands

CMake has a bunch of useful built-in commands. This document isn't going to go into details about them because The CMake project has excellent documentation. To highlight a few useful functions see:

- `add_custom_command`
- `add_custom_target`
- `file`
- `list`
- `math`
- `string`

The full documentation for CMake commands is in the `cmake-commands` manpage and available on [CMake's website](#)

2.3 Advanced Build Configurations

- *Introduction*
- *Bootstrap Builds*
- *Apple Clang Builds (A More Complex Bootstrap)*
- *Multi-stage PGO*
- *3-Stage Non-Determinism*

2.3.1 Introduction

CMake is a cross-platform build-generator tool. CMake does not build the project, it generates the files needed by your build tool (GNU make, Visual Studio, etc.) for building LLVM.

If **you are a new contributor**, please start with the *Getting Started with the LLVM System* or *Building LLVM with CMake* pages. This page is intended for users doing more complex builds.

Many of the examples below are written assuming specific CMake Generators. Unless otherwise explicitly called out these commands should work with any CMake generator.

2.3.2 Bootstrap Builds

The Clang CMake build system supports bootstrap (aka multi-stage) builds. At a high level a multi-stage build is a chain of builds that pass data from one stage into the next. The most common and simple version of this is a traditional bootstrap build.

In a simple two-stage bootstrap build, we build clang using the system compiler, then use that just-built clang to build clang again. In CMake this simplest form of a bootstrap build can be configured with a single option, `CLANG_ENABLE_BOOTSTRAP`.

```
$ cmake -G Ninja -DCLANG_ENABLE_BOOTSTRAP=On <path to source>
$ ninja stage2
```

This command itself isn't terribly useful because it assumes default configurations for each stage. The next series of examples utilize CMake cache scripts to provide more complex options.

The clang build system refers to builds as stages. A stage1 build is a standard build using the compiler installed on the host, and a stage2 build is built using the stage1 compiler. This nomenclature holds up to more stages too. In general a stage*n* build is built using the output from stage*n-1*.

2.3.3 Apple Clang Builds (A More Complex Bootstrap)

Apple's Clang builds are a slightly more complicated example of the simple bootstrapping scenario. Apple Clang is built using a 2-stage build.

The stage1 compiler is a host-only compiler with some options set. The stage1 compiler is a balance of optimization vs build time because it is a throwaway. The stage2 compiler is the fully optimized compiler intended to ship to users.

Setting up these compilers requires a lot of options. To simplify the configuration the Apple Clang build settings are contained in CMake Cache files. You can build an Apple Clang compiler using the following commands:

```
$ cmake -G Ninja -C <path to clang>/cmake/caches/Apple-stage1.cmake <path to source>
$ ninja stage2-distribution
```

This CMake invocation configures the stage1 host compiler, and sets `CLANG_BOOTSTRAP_CMAKE_ARGS` to pass the Apple-stage2.cmake cache script to the stage2 configuration step.

When you build the stage2-distribution target it builds the minimal stage1 compiler and required tools, then configures and builds the stage2 compiler based on the settings in Apple-stage2.cmake.

This pattern of using cache scripts to set complex settings, and specifically to make later stage builds include cache scripts is common in our more advanced build configurations.

2.3.4 Multi-stage PGO

Profile-Guided Optimizations (PGO) is a really great way to optimize the code clang generates. Our multi-stage PGO builds are a workflow for generating PGO profiles that can be used to optimize clang.

At a high level, the way PGO works is that you build an instrumented compiler, then you run the instrumented compiler against sample source files. While the instrumented compiler runs it will output a bunch of files containing performance counters (.profrac files). After generating all the profraw files you use `llvm-profdata` to merge the files into a single profdata file that you can feed into the `LLVM_PROFDATA_FILE` option.

Our PGO.cmake cache script automates that whole process. You can use it by running:

```
$ cmake -G Ninja -C <path_to_clang>/cmake/caches/PGO.cmake <source dir>
$ ninja stage2-instrumented-generate-profdata
```

If you let that run for a few hours or so, it will place a profdata file in your build directory. This takes a really long time because it builds clang twice, and you *must* have compiler-rt in your build tree.

This process uses any source files under the perf-training directory as training data as long as the source files are marked up with LIT-style RUN lines.

After it finishes you can use “find . -name clang.profdata” to find it, but it should be at a path something like:

```
<build_dir>/tools/clang/stage2-instrumented-bins/utils/perf-training/clang.profdata
```

You can feed that file into the LLVM_PROFDATA_FILE option when you build your optimized compiler.

The PGO came cache has a slightly different stage naming scheme than other multi-stage builds. It generates three stages; stage1, stage2-instrumented, and stage2. Both of the stage2 builds are built using the stage1 compiler.

The PGO came cache generates the following additional targets:

stage2-instrumented Builds a stage1 x86 compiler, runtime, and required tools (llvm-config, llvm-profdata) then uses that compiler to build an instrumented stage2 compiler.

stage2-instrumented-generate-profdata Depends on “stage2-instrumented” and will use the instrumented compiler to generate profdata based on the training files in <clang>/utils/perf-training

stage2 Depends of “stage2-instrumented-generate-profdata” and will use the stage1 compiler with the stage2 profdata to build a PGO-optimized compiler.

stage2-check-llvm Depends on stage2 and runs check-llvm using the stage2 compiler.

stage2-check-clang Depends on stage2 and runs check-clang using the stage2 compiler.

stage2-check-all Depends on stage2 and runs check-all using the stage2 compiler.

stage2-test-suite Depends on stage2 and runs the test-suite using the stage3 compiler (requires in-tree test-suite).

2.3.5 3-Stage Non-Determinism

In the ancient lore of compilers non-determinism is like the multi-headed hydra. Whenever it’s head pops up, terror and chaos ensue.

Historically one of the tests to verify that a compiler was deterministic would be a three stage build. The idea of a three stage build is you take your sources and build a compiler (stage1), then use that compiler to rebuild the sources (stage2), then you use that compiler to rebuild the sources a third time (stage3) with an identical configuration to the stage2 build. At the end of this, you have a stage2 and stage3 compiler that should be bit-for-bit identical.

You can perform one of these 3-stage builds with LLVM & clang using the following commands:

```
$ cmake -G Ninja -C <path_to_clang>/cmake/caches/3-stage.cmake <source_dir>
$ ninja stage3
```

After the build you can compare the stage2 & stage3 compilers. We have a bot setup [here](#) that runs this build and compare configuration.

2.4 How To Build On ARM

2.4.1 Introduction

This document contains information about building/testing LLVM and Clang on an ARM machine.

This document is *NOT* tailored to help you cross-compile LLVM/Clang to ARM on another architecture, for example an x86_64 machine. To find out more about cross-compiling, please check [How To Cross-Compile Clang/LLVM using Clang/LLVM](#).

2.4.2 Notes On Building LLVM/Clang on ARM

Here are some notes on building/testing LLVM/Clang on ARM. Note that ARM encompasses a wide variety of CPUs; this advice is primarily based on the ARMv6 and ARMv7 architectures and may be inapplicable to older chips.

1. The most popular Linaro/Ubuntu OS's for ARM boards, e.g., the Pandaboard, have become hard-float platforms. There are a number of choices when using CMake. Autoconf usage is deprecated as of 3.8.

Building LLVM/Clang in Release mode is preferred since it consumes a lot less memory. Otherwise, the building process will very likely fail due to insufficient memory. It's also a lot quicker to only build the relevant back-ends (ARM and AArch64), since it's very unlikely that you'll use an ARM board to cross-compile to other arches. If you're running Compiler-RT tests, also include the x86 back-end, or some tests will fail.

```
cmake $LLVM_SRC_DIR -DCMAKE_BUILD_TYPE=Release \
      -DLLVM_TARGETS_TO_BUILD="ARM;X86;AArch64"
```

Other options you can use are:

```
Use Ninja instead of Make: "-G Ninja"
Build with assertions on: "-DLLVM_ENABLE_ASSERTIONS=True"
Force Python2: "-DPYTHON_EXECUTABLE=/usr/bin/python2"
Local (non-sudo) install path: "-DCMAKE_INSTALL_PREFIX=$HOME/llvm/instal"
CPU flags: "DCMAKE_C_FLAGS=-mcpu=cortex-a15" (same for CXX_FLAGS)
```

After that, just typing `make -jN` or `ninja` will build everything. `make -jN check-all` or `ninja check-all` will run all compiler tests. For running the test suite, please refer to *LLVM Testing Infrastructure Guide*.

2. If you are building LLVM/Clang on an ARM board with 1G of memory or less, please use `gold` rather than `GNU ld`. In any case it is probably a good idea to set up a swap partition, too.

```
$ sudo ln -sf /usr/bin/ld /usr/bin/ld.gold
```

3. ARM development boards can be unstable and you may experience that cores are disappearing, caches being flushed on every big.LITTLE switch, and other similar issues. To help ease the effect of this, set the Linux scheduler to “performance” on **all** cores using this little script:

```
# The code below requires the package 'cpufrequtils' to be installed.
for ((cpu=0; cpu<`grep -c proc /proc/cpuinfo`; cpu++)); do
    sudo cpufreq-set -c $cpu -g performance
done
```

Remember to turn that off after the build, or you may risk burning your CPU. Most modern kernels don't need that, so only use it if you have problems.

4. Running the build on SD cards is ok, but they are more prone to failures than good quality USB sticks, and those are more prone to failures than external hard-drives (those are also a lot faster). So, at least, you should consider to buy a fast USB stick. On systems with a fast eMMC, that's a good option too.
5. Make sure you have a decent power supply (dozens of dollars worth) that can provide *at least* 4 amperes, this is especially important if you use USB devices with your board. Externally powered USB/SATA harddrives are even better than having a good power supply.

2.5 How To Cross-Compile Clang/LLVM using Clang/LLVM

2.5.1 Introduction

This document contains information about building LLVM and Clang on host machine, targeting another platform.

For more information on how to use Clang as a cross-compiler, please check <http://clang.llvm.org/docs/CrossCompilation.html>.

TODO: Add MIPS and other platforms to this document.

2.5.2 Cross-Compiling from x86_64 to ARM

In this use case, we'll be using CMake and Ninja, on a Debian-based Linux system, cross-compiling from an x86_64 host (most Intel and AMD chips nowadays) to a hard-float ARM target (most ARM targets nowadays).

The packages you'll need are:

- `cmake`
- `ninja-build` (from backports in Ubuntu)
- `gcc-4.7-arm-linux-gnueabi`
- `gcc-4.7-multilib-arm-linux-gnueabi`
- `binutils-arm-linux-gnueabi`
- `libgcc1-armhf-cross`
- `libsfgcc1-armhf-cross`
- `libstdc++6-armhf-cross`
- `libstdc++6-4.7-dev-armhf-cross`

Configuring CMake

For more information on how to configure CMake for LLVM/Clang, see *Building LLVM with CMake*.

The CMake options you need to add are:

- `-DCMAKE_CROSSCOMPILING=True`
- `-DCMAKE_INSTALL_PREFIX=<install-dir>`
- `-DLLVM_TABLEGEN=<path-to-host-bin>/llvm-tblgen`
- `-DCLANG_TABLEGEN=<path-to-host-bin>/clang-tblgen`
- `-DLLVM_DEFAULT_TARGET_TRIPLE=arm-linux-gnueabi`
- `-DLLVM_TARGET_ARCH=ARM`
- `-DLLVM_TARGETS_TO_BUILD=ARM`

If you're compiling with GCC, you can use architecture options for your target, and the compiler driver will detect everything that it needs:

- `-DCMAKE_CXX_FLAGS='-march=armv7-a -mcpu=cortex-a9 -mfloat-abi=hard'`

However, if you're using Clang, the driver might not be up-to-date with your specific Linux distribution, version or GCC layout, so you'll need to fudge.

In addition to the ones above, you'll also need:

- `'-target arm-linux-gnueabi'` or whatever is the triple of your cross GCC.
- `'--sysroot=/usr/arm-linux-gnueabi'`, `'--sysroot=/opt/gcc/arm-linux-gnueabi'` or whatever is the location of your GCC's sysroot (where `/lib`, `/bin` etc are).
- Appropriate use of `-I` and `-L`, depending on how the cross GCC is installed, and where are the libraries and headers.

The TableGen options are required to compile it with the host compiler, so you'll need to compile LLVM (or at least `llvm-tblgen`) to your host platform before you start. The CXX flags define the target, `cpu` (which in this case defaults to `fpv=VFP3` with NEON), and forcing the hard-float ABI. If you're using Clang as a cross-compiler, you will *also* have to set `--sysroot` to make sure it picks the correct linker.

When using Clang, it's important that you choose the triple to be *identical* to the GCC triple and the sysroot. This will make it easier for Clang to find the correct tools and include headers. But that won't mean all headers and libraries will be found. You'll still need to use `-I` and `-L` to locate those extra ones, depending on your distribution.

Most of the time, what you want is to have a native compiler to the platform itself, but not others. So there's rarely a point in compiling all back-ends. For that reason, you should also set the `TARGETS_TO_BUILD` to only build the back-end you're targeting to.

You must set the `CMAKE_INSTALL_PREFIX`, otherwise a `ninja install` will copy ARM binaries to your root filesystem, which is not what you want.

Hacks

There are some bugs in current LLVM, which require some fiddling before running CMake:

1. If you're using Clang as the cross-compiler, there is a problem in the LLVM ARM back-end that is producing absolute relocations on position-independent code (`R_ARM_THM_MOVW_ABS_NC`), so for now, you should disable PIC:

```
-DLLVM_ENABLE_PIC=False
```

This is not a problem, since Clang/LLVM libraries are statically linked anyway, it shouldn't affect much.

2. The ARM libraries won't be installed in your system. But the CMake prepare step, which checks for dependencies, will check the *host* libraries, not the *target* ones. Below there's a list of some dependencies, but your project could have more, or this document could be outdated. You'll see the errors while linking as an indication of that.

Debian based distros have a way to add `multiarch`, which adds a new architecture and allows you to install packages for those systems. See <https://wiki.debian.org/Multiarch/HOWTO> for more info.

But not all distros will have that, and possibly not an easy way to install them in any way, so you'll have to build/download them separately.

A quick way of getting the libraries is to download them from a distribution repository, like Debian (<http://packages.debian.org/jessie/>), and download the missing libraries. Note that the `libXXX` will have the shared objects (`.so`) and the `libXXX-dev` will give you the headers and the static (`.a`) library. Just in case, download both.

The ones you need for ARM are: `libtinfo`, `zlib1g`, `libxml2` and `liblzma`. In the Debian repository you'll find downloads for all architectures.

After you download and unpack all `.deb` packages, copy all `.so` and `.a` to a directory, make the appropriate symbolic links (if necessary), and add the relevant `-L` and `-I` paths to `-DCMAKE_CXX_FLAGS` above.

Running CMake and Building

Finally, if you're using your platform compiler, run:

```
$ cmake -G Ninja <source-dir> <options above>
```

If you're using Clang as the cross-compiler, run:

```
$ CC='clang' CXX='clang++' cmake -G Ninja <source-dir> <options above>
```

If you have `clang/clang++` on the path, it should just work, and special Ninja files will be created in the build directory. I strongly suggest you to run `cmake` on a separate build directory, *not* inside the source tree.

To build, simply type:

```
$ ninja
```

It should automatically find out how many cores you have, what are the rules that needs building and will build the whole thing.

You can't run `ninja check-all` on this tree because the created binaries are targeted to ARM, not `x86_64`.

Installing and Using

After the LLVM/Clang has built successfully, you should install it via:

```
$ ninja install
```

which will create a sysroot on the `install-dir`. You can then tar that directory into a binary with the full triple name (for easy identification), like:

```
$ ln -sf <install-dir> arm-linux-gnueabi-hf-clang
$ tar zchf arm-linux-gnueabi-hf-clang.tar.gz arm-linux-gnueabi-hf-clang
```

If you copy that tarball to your target board, you'll be able to use it for running the test-suite, for example. Follow the guidelines at <http://llvm.org/docs/ln/quickstart.html>, unpack the tarball in the test directory, and use options:

```
$ ./sandbox/bin/python sandbox/bin/ln runtest nt \
--sandbox sandbox \
--test-suite `pwd`/test-suite \
--cc `pwd`/arm-linux-gnueabi-hf-clang/bin/clang \
--cxx `pwd`/arm-linux-gnueabi-hf-clang/bin/clang++
```

Remember to add the `-jN` options to `ln` to the number of CPUs on your board. Also, the path to your clang has to be absolute, so you'll need the `pwd` trick above.

2.6 LLVM Command Guide

The following documents are command descriptions for all of the LLVM tools. These pages describe how to use the LLVM commands and what their options are. Note that these pages do not describe all of the options available for all tools. To get a complete listing, pass the `--help` (general options) or `--help-hidden` (general and debugging options) arguments to the tool you are interested in.

2.6.1 Basic Commands

llvm-as - LLVM assembler

SYNOPSIS

llvm-as [*options*] [*filename*]

DESCRIPTION

llvm-as is the LLVM assembler. It reads a file containing human-readable LLVM assembly language, translates it to LLVM bitcode, and writes the result into a file or to standard output.

If *filename* is omitted or is `-`, then **llvm-as** reads its input from standard input.

If an output file is not specified with the **-o** option, then **llvm-as** sends its output to a file or standard output by following these rules:

- If the input is standard input, then the output is standard output.
- If the input is a file that ends with `.ll`, then the output file is of the same name, except that the suffix is changed to `.bc`.
- If the input is a file that does not end with the `.ll` suffix, then the output file has the same name as the input file, except that the `.bc` suffix is appended.

OPTIONS

-f Enable binary output on terminals. Normally, **llvm-as** will refuse to write raw bitcode output if the output stream is a terminal. With this option, **llvm-as** will write raw bitcode regardless of the output device.

-help Print a summary of command line options.

-o filename Specify the output file name. If *filename* is `-`, then **llvm-as** sends its output to standard output.

EXIT STATUS

If **llvm-as** succeeds, it will exit with 0. Otherwise, if an error occurs, it will exit with a non-zero value.

SEE ALSO

llvm-dis|llvm-dis, gccas|gccas

llvm-dis - LLVM disassembler

SYNOPSIS

llvm-dis [*options*] [*filename*]

DESCRIPTION

The **llvm-dis** command is the LLVM disassembler. It takes an LLVM bitcode file and converts it into human-readable LLVM assembly language.

If filename is omitted or specified as `-`, **llvm-dis** reads its input from standard input.

If the input is being read from standard input, then **llvm-dis** will send its output to standard output by default. Otherwise, the output will be written to a file named after the input file, with a `.ll` suffix added (any existing `.bc` suffix will first be removed). You can override the choice of output file using the **-o** option.

OPTIONS

-f

Enable binary output on terminals. Normally, **llvm-dis** will refuse to write raw bitcode output if the output stream is a terminal. With this option, **llvm-dis** will write raw bitcode regardless of the output device.

-help

Print a summary of command line options.

-o filename

Specify the output file name. If *filename* is `-`, then the output is sent to standard output.

EXIT STATUS

If **llvm-dis** succeeds, it will exit with 0. Otherwise, if an error occurs, it will exit with a non-zero value.

SEE ALSO

llvm-as|llvm-as

opt - LLVM optimizer

SYNOPSIS

opt [*options*] [*filename*]

DESCRIPTION

The **opt** command is the modular LLVM optimizer and analyzer. It takes LLVM source files as input, runs the specified optimizations or analyses on it, and then outputs the optimized file or the analysis results. The function of **opt** depends on whether the *-analyze* option is given.

When *-analyze* is specified, **opt** performs various analyses of the input source. It will usually print the results on standard output, but in a few cases, it will print output to standard error or generate a file with the analysis output, which is usually done when the output is meant for another program.

While *-analyze* is *not* given, **opt** attempts to produce an optimized output file. The optimizations available via **opt** depend upon what libraries were linked into it as well as any additional libraries that have been loaded with the *-load* option. Use the *-help* option to determine what optimizations you can use.

If `filename` is omitted from the command line or is “-“, **opt** reads its input from standard input. Inputs can be in either the LLVM assembly language format (`.ll`) or the LLVM bitcode format (`.bc`).

If an output filename is not specified with the `-o` option, **opt** writes its output to the standard output.

OPTIONS

-f

Enable binary output on terminals. Normally, **opt** will refuse to write raw bitcode output if the output stream is a terminal. With this option, **opt** will write raw bitcode regardless of the output device.

-help

Print a summary of command line options.

-o <filename>

Specify the output filename.

-S

Write output in LLVM intermediate language (instead of bitcode).

-{passname}

opt provides the ability to run any of LLVM’s optimization or analysis passes in any order. The `-help` option lists all the passes available. The order in which the options occur on the command line are the order in which they are executed (within pass constraints).

-disable-inlining

This option simply removes the inlining pass from the standard list.

-disable-opt

This option is only meaningful when `-std-link-opts` is given. It disables most passes.

-strip-debug

This option causes **opt** to strip debug information from the module before applying other optimizations. It is essentially the same as `-strip` but it ensures that stripping of debug information is done first.

-verify-each

This option causes **opt** to add a verify pass after every pass otherwise specified on the command line (including `-verify`). This is useful for cases where it is suspected that a pass is creating an invalid module but it is not clear which pass is doing it.

-stats

Print statistics.

-time-passes

Record the amount of time needed for each pass and print it to standard error.

-debug

If this is a debug build, this option will enable debug printouts from passes which use the `DEBUG()` macro. See the [LLVM Programmer’s Manual](#), section `#DEBUG` for more information.

-load=<plugin>

Load the dynamic object plugin. This object should register new optimization or analysis passes. Once loaded, the object will add new command line options to enable various optimizations or analyses. To see the new complete list of optimizations, use the `-help` and `-load` options together. For example:

```
opt -load=plugin.so -help
```

-p

Print module after each transformation.

EXIT STATUS

If **opt** succeeds, it will exit with 0. Otherwise, if an error occurs, it will exit with a non-zero value.

llc - LLVM static compiler

SYNOPSIS

llc [*options*] [*filename*]

DESCRIPTION

The **llc** command compiles LLVM source inputs into assembly language for a specified architecture. The assembly language output can then be passed through a native assembler and linker to generate a native executable.

The choice of architecture for the output assembly code is automatically determined from the input file, unless the `-march` option is used to override the default.

OPTIONS

If *filename* is “-” or omitted, **llc** reads from standard input. Otherwise, it will from *filename*. Inputs can be in either the LLVM assembly language format (`.ll`) or the LLVM bitcode format (`.bc`).

If the `-o` option is omitted, then **llc** will send its output to standard output if the input is from standard input. If the `-o` option specifies “-“, then the output will also be sent to standard output.

If no `-o` option is specified and an input file other than “-” is specified, then **llc** creates the output filename by taking the input filename, removing any existing `.bc` extension, and adding a `.s` suffix.

Other **llc** options are described below.

End-user Options

-help

Print a summary of command line options.

-O=uint

Generate code at different optimization levels. These correspond to the `-O0`, `-O1`, `-O2`, and `-O3` optimization levels used by **clang**.

-mtriple=<target triple>

Override the target triple specified in the input file with the specified string.

-march=<arch>

Specify the architecture for which to generate assembly, overriding the target encoded in the input file. See the output of `llc -help` for a list of valid architectures. By default this is inferred from the target triple or autodetected to the current architecture.

-mcpu=<cpuname>

Specify a specific chip in the current architecture to generate code for. By default this is inferred from the target triple and autodetected to the current architecture. For a list of available CPUs, use:

```
llvm-as < /dev/null | llc -march=xyz -mcpu=help
```

-filetype=<output file type>

Specify what kind of output `llc` should generate. Options are: `asm` for textual assembly (`' .s'`), `obj` for native object files (`' .o'`) and `null` for not emitting anything (for performance testing).

Note that not all targets support all options.

-mattr=`a1,+a2,-a3,...`

Override or control specific attributes of the target, such as whether SIMD operations are enabled or not. The default set of attributes is set by the current CPU. For a list of available attributes, use:

```
llvm-as < /dev/null | llc -march=xyz -mattr=help
```

--disable-fp-elim

Disable frame pointer elimination optimization.

--disable-excess-fp-precision

Disable optimizations that may produce excess precision for floating point. Note that this option can dramatically slow down code on some systems (e.g. X86).

--enable-no-infs-fp-math

Enable optimizations that assume no Inf values.

--enable-no-nans-fp-math

Enable optimizations that assume no NAN values.

--enable-unsafe-fp-math

Enable optimizations that make unsafe assumptions about IEEE math (e.g. that addition is associative) or may not work for all input ranges. These optimizations allow the code generator to make use of some instructions which would otherwise not be usable (such as `fsin` on X86).

--stats

Print statistics recorded by code-generation passes.

--time-passes

Record the amount of time needed for each pass and print a report to standard error.

--load=<dso_path>

Dynamically load `dso_path` (a path to a dynamically shared object) that implements an LLVM target. This will permit the target name to be used with the `-march` option so that code can be generated for that target.

-meabi=[`default|gnu|4|5`]

Specify which EABI version should conform to. Valid EABI versions are *gnu*, *4* and *5*. Default value (*default*) depends on the triple.

Tuning/Configuration Options

--print-machineinstrs

Print generated machine code between compilation phases (useful for debugging).

--regalloc=<allocator>

Specify the register allocator to use. Valid register allocators are:

basic

Basic register allocator.

fast

Fast register allocator. It is the default for unoptimized code.

greedy

Greedy register allocator. It is the default for optimized code.

pbqp

Register allocator based on ‘Partitioned Boolean Quadratic Programming’.

--spiller=<spiller>

Specify the spiller to use for register allocators that support it. Currently this option is used only by the linear scan register allocator. The default *spiller* is *local*. Valid spillers are:

simple

Simple spiller

local

Local spiller

Intel IA-32-specific Options

--x86-asm-syntax=[att|intel]

Specify whether to emit assembly code in AT&T syntax (the default) or Intel syntax.

EXIT STATUS

If **lli** succeeds, it will exit with 0. Otherwise, if an error occurs, it will exit with a non-zero value.

SEE ALSO

lli

lli - directly execute programs from LLVM bitcode

SYNOPSIS

lli [*options*] [*filename*] [*program args*]

DESCRIPTION

lli directly executes programs in LLVM bitcode format. It takes a program in LLVM bitcode format and executes it using a just-in-time compiler or an interpreter.

lli is *not* an emulator. It will not execute IR of different architectures and it can only interpret (or JIT-compile) for the host architecture.

The JIT compiler takes the same arguments as other tools, like **llc**, but they don’t necessarily work for the interpreter.

If *filename* is not specified, then **lli** reads the LLVM bitcode for the program from standard input.

The optional *args* specified on the command line are passed to the program as arguments.

GENERAL OPTIONS

-fake-argv0=executable

Override the `argv[0]` value passed into the executing program.

-force-interpreter={false,true}

If set to true, use the interpreter even if a just-in-time compiler is available for this architecture. Defaults to false.

-help

Print a summary of command line options.

-load=pluginfilename

Causes **lli** to load the plugin (shared object) named *pluginfilename* and use it for optimization.

-stats

Print statistics from the code-generation passes. This is only meaningful for the just-in-time compiler, at present.

-time-passes

Record the amount of time needed for each code-generation pass and print it to standard error.

-version

Print out the version of **lli** and exit without doing anything else.

TARGET OPTIONS

-mtriple=target triple

Override the target triple specified in the input bitcode file with the specified string. This may result in a crash if you pick an architecture which is not compatible with the current system.

-march=arch

Specify the architecture for which to generate assembly, overriding the target encoded in the bitcode file. See the output of **llc -help** for a list of valid architectures. By default this is inferred from the target triple or autodetected to the current architecture.

-mcpu=cputype

Specify a specific chip in the current architecture to generate code for. By default this is inferred from the target triple and autodetected to the current architecture. For a list of available CPUs, use: **llvm-as < /dev/null | llc**

-march=xyz -mcpu=help

-mattr=a1,+a2,-a3,...

Override or control specific attributes of the target, such as whether SIMD operations are enabled or not. The default set of attributes is set by the current CPU. For a list of available attributes, use: **llvm-as < /dev/null | llc**

-march=xyz -mattr=help

FLOATING POINT OPTIONS

-disable-excess-fp-precision

Disable optimizations that may increase floating point precision.

-enable-no-infs-fp-math

Enable optimizations that assume no Inf values.

-enable-no-nans-fp-math

Enable optimizations that assume no NAN values.

-enable-unsafe-fp-math

Causes **lli** to enable optimizations that may decrease floating point precision.

-soft-float

Causes `lli` to generate software floating point library calls instead of equivalent hardware instructions.

CODE GENERATION OPTIONS**-code-model=model**

Choose the code model from:

```
default: Target default code model
small: Small code model
kernel: Kernel code model
medium: Medium code model
large: Large code model
```

-disable-post-RA-scheduler

Disable scheduling after register allocation.

-disable-spill-fusing

Disable fusing of spill code into instructions.

-jit-enable-eh

Exception handling should be enabled in the just-in-time compiler.

-join-liveintervals

Coalesce copies (default=true).

-nozero-initialized-in-bss

Don't place zero-initialized symbols into the BSS section.

-pre-RA-sched=scheduler

Instruction schedulers available (before register allocation):

```
=default: Best scheduler for the target
=none: No scheduling: breadth first sequencing
=simple: Simple two pass scheduling: minimize critical path and maximize_
    ↪ processor utilization
=simple-noitin: Simple two pass scheduling: Same as simple except using generic_
    ↪ latency
=list-burr: Bottom-up register reduction list scheduling
=list-tdrr: Top-down register reduction list scheduling
=list-td: Top-down list scheduler -print-machineinstrs - Print generated machine_
    ↪ code
```

-regalloc=allocator

Register allocator to use (default=linearscan)

```
=bigblock: Big-block register allocator
=linearscan: linear scan register allocator =local - local register allocator
=simple: simple register allocator
```

-relocation-model=model

Choose relocation model from:

```
=default: Target default relocation model
=static: Non-relocatable code =pic - Fully relocatable, position independent_
    ↪ code
=dynamic-no-pic: Relocatable external references, non-relocatable code
```

-spiller

Spiller to use (default=local)

```
=simple: simple spiller
=local: local spiller
```

-x86-asm-syntax=syntax

Choose style of code to emit from X86 backend:

```
=att: Emit AT&T-style assembly
=intel: Emit Intel-style assembly
```

EXIT STATUS

If **lli** fails to load the program, it will exit with an exit code of 1. Otherwise, it will return the exit code of the program it executes.

SEE ALSO

llc

llvm-link - LLVM bitcode linker

SYNOPSIS

llvm-link [*options*] *filename* ...

DESCRIPTION

llvm-link takes several LLVM bitcode files and links them together into a single LLVM bitcode file. It writes the output file to standard output, unless the **-o** option is used to specify a filename.

OPTIONS**-f**

Enable binary output on terminals. Normally, **llvm-link** will refuse to write raw bitcode output if the output stream is a terminal. With this option, **llvm-link** will write raw bitcode regardless of the output device.

-o filename

Specify the output file name. If *filename* is “-“, then **llvm-link** will write its output to standard output.

-S

Write output in LLVM intermediate language (instead of bitcode).

-d

If specified, **llvm-link** prints a human-readable version of the output bitcode file to standard error.

-help

Print a summary of command line options.

-v

Verbose mode. Print information about what **llvm-link** is doing. This typically includes a message for each bitcode file linked in and for each library found.

EXIT STATUS

If **llvm-link** succeeds, it will exit with 0. Otherwise, if an error occurs, it will exit with a non-zero value.

llvm-ar - LLVM archiver

SYNOPSIS

llvm-ar [-]{dmpqrtx}[Rabfikou] [relpos] [count] <archive> [files...]

DESCRIPTION

The **llvm-ar** command is similar to the common Unix utility, **ar**. It archives several files together into a single file. The intent for this is to produce archive libraries by LLVM bitcode that can be linked into an LLVM program. However, the archive can contain any kind of file. By default, **llvm-ar** generates a symbol table that makes linking faster because only the symbol table needs to be consulted, not each individual file member of the archive.

The **llvm-ar** command can be used to *read* SVR4, GNU and BSD style archive files. However, right now it can only write in the GNU format. If an SVR4 or BSD style archive is used with the **r** (replace) or **q** (quick update) operations, the archive will be reconstructed in GNU format.

Here's where **llvm-ar** departs from previous **ar** implementations:

Symbol Table

Since **llvm-ar** supports bitcode files. The symbol table it creates is in GNU format and includes both native and bitcode files.

Long Paths

Currently **llvm-ar** can read GNU and BSD long file names, but only writes archives with the GNU format.

OPTIONS

The options to **llvm-ar** are compatible with other **ar** implementations. However, there are a few modifiers (*R*) that are not found in other **ar** implementations. The options to **llvm-ar** specify a single basic operation to perform on the archive, a variety of modifiers for that operation, the name of the archive file, and an optional list of file names. These options are used to determine how **llvm-ar** should process the archive file.

The Operations and Modifiers are explained in the sections below. The minimal set of options is at least one operator and the name of the archive. Typically archive files end with a **.a** suffix, but this is not required. Following the *archive-name* comes a list of *files* that indicate the specific members of the archive to operate on. If the *files* option is not specified, it generally means either “none” or “all” members, depending on the operation.

Operations

d

Delete files from the archive. No modifiers are applicable to this operation. The *files* options specify which members should be removed from the archive. It is not an error if a specified file does not appear in the archive. If no *files* are specified, the archive is not modified.

m[abi]

Move files from one location in the archive to another. The *a*, *b*, and *i* modifiers apply to this operation. The *files* will all be moved to the location given by the modifiers. If no modifiers are used, the files will be moved to the end of the archive. If no *files* are specified, the archive is not modified.

p

Print files to the standard output. This operation simply prints the *files* indicated to the standard output. If no *files* are specified, the entire archive is printed. Printing bitcode files is ill-advised as they might confuse your terminal settings. The *p* operation never modifies the archive.

q

Quickly append files to the end of the archive. This operation quickly adds the *files* to the archive without checking for duplicates that should be removed first. If no *files* are specified, the archive is not modified. Because of the way that **llvm-ar** constructs the archive file, its dubious whether the *q* operation is any faster than the *r* operation.

r[abu]

Replace or insert file members. The *a*, *b*, and *u* modifiers apply to this operation. This operation will replace existing *files* or insert them at the end of the archive if they do not exist. If no *files* are specified, the archive is not modified.

t[v]

Print the table of contents. Without any modifiers, this operation just prints the names of the members to the standard output. With the *v* modifier, **llvm-ar** also prints out the file type (B=bitcode, S=symbol table, blank=regular file), the permission mode, the owner and group, the size, and the date. If any *files* are specified, the listing is only for those files. If no *files* are specified, the table of contents for the whole archive is printed.

x[oP]

Extract archive members back to files. The *o* modifier applies to this operation. This operation retrieves the indicated *files* from the archive and writes them back to the operating system's file system. If no *files* are specified, the entire archive is extract.

Modifiers (operation specific)

The modifiers below are specific to certain operations. See the Operations section (above) to determine which modifiers are applicable to which operations.

[a]

When inserting or moving member files, this option specifies the destination of the new files as being after the *relopos* member. If *relopos* is not found, the files are placed at the end of the archive.

[b]

When inserting or moving member files, this option specifies the destination of the new files as being before the *relopos* member. If *relopos* is not found, the files are placed at the end of the archive. This modifier is identical to the *i* modifier.

[i]

A synonym for the *b* option.

[o]

When extracting files, this option will cause **llvm-ar** to preserve the original modification times of the files it writes.

[u]

When replacing existing files in the archive, only replace those files that have a time stamp than the time stamp of the member in the archive.

Modifiers (generic)

The modifiers below may be applied to any operation.

[c]

For all operations, **llvm-ar** will always create the archive if it doesn't exist. Normally, **llvm-ar** will print a warning message indicating that the archive is being created. Using this modifier turns off that warning.

[s]

This modifier requests that an archive index (or symbol table) be added to the archive. This is the default mode of operation. The symbol table will contain all the externally visible functions and global variables defined by all the bitcode files in the archive.

[S]

This modifier is the opposite of the *s* modifier. It instructs **llvm-ar** to not build the symbol table. If both *s* and *S* are used, the last modifier to occur in the options will prevail.

[v]

This modifier instructs **llvm-ar** to be verbose about what it is doing. Each editing operation taken against the archive will produce a line of output saying what is being done.

STANDARDS

The **llvm-ar** utility is intended to provide a superset of the IEEE Std 1003.2 (POSIX.2) functionality for *ar*. **llvm-ar** can read both SVR4 and BSD4.4 (or Mac OS X) archives. If the *f* modifier is given to the *x* or *r* operations then **llvm-ar** will write SVR4 compatible archives. Without this modifier, **llvm-ar** will write BSD4.4 compatible archives that have long names immediately after the header and indicated using the “#1/ddd” notation for the name in the header.

FILE FORMAT

The file format for LLVM Archive files is similar to that of BSD 4.4 or Mac OSX archive files. In fact, except for the symbol table, the *ar* commands on those operating systems should be able to read LLVM archive files. The details of the file format follow.

Each archive begins with the archive magic number which is the eight printable characters “!<arch>n” where *n* represents the newline character (0x0A). Following the magic number, the file is composed of even length members that begin with an archive header and end with a *n* padding character if necessary (to make the length even). Each file member is composed of a header (defined below), an optional newline-terminated “long file name” and the contents of the file.

The fields of the header are described in the items below. All fields of the header contain only ASCII characters, are left justified and are right padded with space characters.

name - char[16]

This field of the header provides the name of the archive member. If the name is longer than 15 characters or contains a slash (/) character, then this field contains #1/nnn where nnn provides the length of the name and the #1/ is literal. In this case, the actual name of the file is provided in the nnn bytes immediately following the header. If the name is 15 characters or less, it is contained directly in this field and terminated with a slash (/) character.

date - char[12]

This field provides the date of modification of the file in the form of a decimal encoded number that provides the number of seconds since the epoch (since 00:00:00 Jan 1, 1970) per Posix specifications.

uid - char[6]

This field provides the user id of the file encoded as a decimal ASCII string. This field might not make much sense on non-Unix systems. On Unix, it is the same value as the st_uid field of the stat structure returned by the stat(2) operating system call.

gid - char[6]

This field provides the group id of the file encoded as a decimal ASCII string. This field might not make much sense on non-Unix systems. On Unix, it is the same value as the st_gid field of the stat structure returned by the stat(2) operating system call.

mode - char[8]

This field provides the access mode of the file encoded as an octal ASCII string. This field might not make much sense on non-Unix systems. On Unix, it is the same value as the st_mode field of the stat structure returned by the stat(2) operating system call.

size - char[10]

This field provides the size of the file, in bytes, encoded as a decimal ASCII string.

fmag - char[2]

This field is the archive file member magic number. Its content is always the two characters back tick (0x60) and newline (0x0A). This provides some measure utility in identifying archive files that have been corrupted.

offset - vbr encoded 32-bit integer

The offset item provides the offset into the archive file where the bitcode member is stored that is associated with the symbol. The offset value is 0 based at the start of the first “normal” file member. To derive the actual file offset of the member, you must add the number of bytes occupied by the file signature (8 bytes) and the symbol tables. The value of this item is encoded using variable bit rate encoding to reduce the size of the symbol table. Variable bit rate encoding uses the high bit (0x80) of each byte to indicate if there are more bytes to follow. The remaining 7 bits in each byte carry bits from the value. The final byte does not have the high bit set.

length - vbr encoded 32-bit integer

The length item provides the length of the symbol that follows. Like this *offset* item, the length is variable bit rate encoded.

symbol - character array

The symbol item provides the text of the symbol that is associated with the *offset*. The symbol is not terminated by any character. Its length is provided by the *length* field. Note that is allowed (but unwise) to use non-printing characters (even 0x00) in the symbol. This allows for multiple encodings of symbol names.

EXIT STATUS

If **llvm-ar** succeeds, it will exit with 0. A usage error, results in an exit code of 1. A hard (file system typically) error results in an exit code of 2. Miscellaneous or unknown errors result in an exit code of 3.

SEE ALSO

ar(1)

llvm-lib - LLVM lib.exe compatible library tool

SYNOPSIS

llvm-lib [/libpath:<path>] [/out:<output>] [/llvmlibthin] [/ignore] [/machine] [/nologo] [files...]

DESCRIPTION

The **llvm-lib** command is intended to be a `lib.exe` compatible tool. See <https://msdn.microsoft.com/en-us/library/7ykb2k5f> for the general description.

llvm-lib has the following extensions:

- Bitcode files in symbol tables. **llvm-lib** includes symbols from both bitcode files and regular object files in the symbol table.
- Creating thin archives. The `/llvmlibthin` option causes **llvm-lib** to create thin archive that contain only the symbol table and the header for the various members. These files are much smaller, but are not compatible with `link.exe` (`lld` can handle them).

llvm-nm - list LLVM bitcode and object file's symbol table

SYNOPSIS

llvm-nm [options] [filenames...]

DESCRIPTION

The **llvm-nm** utility lists the names of symbols from the LLVM bitcode files, object files, or **ar** archives containing them, named on the command line. Each symbol is listed along with some simple information about its provenance. If no file name is specified, or `-` is used as a file name, **llvm-nm** will process a file on its standard input stream.

llvm-nm's default output format is the traditional BSD **nm** output format. Each such output record consists of an (optional) 8-digit hexadecimal address, followed by a type code character, followed by a name, for each symbol. One record is printed per line; fields are separated by spaces. When the address is omitted, it is replaced by 8 spaces.

Type code characters currently supported, and their meanings, are as follows:

U

Named object is referenced but undefined in this bitcode file

C

	Common (multiple definitions link together into one def)
W	Weak reference (multiple definitions link together into zero or one definitions)
t	Local function (text) object
T	Global function (text) object
d	Local data object
D	Global data object
?	Something unrecognizable

Because LLVM bitcode files typically contain objects that are not considered to have addresses until they are linked into an executable image or dynamically compiled “just-in-time”, **llvm-nm** does not print an address for any symbol in an LLVM bitcode file, even symbols which are defined in the bitcode file.

OPTIONS

- B** (default)
Use BSD output format. Alias for *-format=bsd*.
- P**
Use POSIX.2 output format. Alias for *-format=posix*.
- debug-syms, -a**
Show all symbols, even debugger only.
- defined-only**
Print only symbols defined in this file (as opposed to symbols which may be referenced by objects in this file, but not defined in this file.)
- dynamic, -D**
Display dynamic symbols instead of normal symbols.
- extern-only, -g**
Print only symbols whose definitions are external; that is, accessible from other files.
- format=format, -f format**
Select an output format; *format* may be *sysv*, *posix*, or *bsd*. The default is *bsd*.
- help**
Print a summary of command-line options and their meanings.
- no-sort, -p**
Shows symbols in order encountered.
- numeric-sort, -n, -v**
Sort symbols by address.

--print-file-name, -A, -o

Precede each symbol with the file it came from.

--print-size, -S

Show symbol size instead of address.

--size-sort

Sort symbols by size.

--undefined-only, -u

Print only symbols referenced but not defined in this file.

--radix=RADIX, -t

Specify the radix of the symbol address(es). Values accepted d(decimal), x(hexadecimal) and o(octal).

BUGS

- **llvm-nm** does not support the full set of arguments that GNU **nm** does.

EXIT STATUS

llvm-nm exits with an exit code of zero.

SEE ALSO

llvm-dis, ar(1), nm(1)

llvm-config - Print LLVM compilation options

SYNOPSIS

llvm-config *option* [*components...*]

DESCRIPTION

llvm-config makes it easier to build applications that use LLVM. It can print the compiler flags, linker flags and object libraries needed to link against LLVM.

EXAMPLES

To link against the JIT:

```
g++ `llvm-config --cxxflags` -o HowToUseJIT.o -c HowToUseJIT.cpp
g++ `llvm-config --ldflags` -o HowToUseJIT HowToUseJIT.o \
    `llvm-config --libs engine bcreader scalaropts`
```

OPTIONS

-version

Print the version number of LLVM.

-help

Print a summary of **llvm-config** arguments.

-prefix

Print the installation prefix for LLVM.

-src-root

Print the source root from which LLVM was built.

-obj-root

Print the object root used to build LLVM.

-bindir

Print the installation directory for LLVM binaries.

-includedir

Print the installation directory for LLVM headers.

-libdir

Print the installation directory for LLVM libraries.

-cxxflags

Print the C++ compiler flags needed to use LLVM headers.

-ldflags

Print the flags needed to link against LLVM libraries.

-libs

Print all the libraries needed to link against the specified LLVM *components*, including any dependencies.

-libnames

Similar to **-libs**, but prints the bare filenames of the libraries without **-l** or pathnames. Useful for linking against a not-yet-installed copy of LLVM.

-libfiles

Similar to **-libs**, but print the full path to each library file. This is useful when creating makefile dependencies, to ensure that a tool is relinked if any library it uses changes.

-components

Print all valid component names.

-targets-built

Print the component names for all targets supported by this copy of LLVM.

-build-mode

Print the build mode used when LLVM was built (e.g. Debug or Release)

COMPONENTS

To print a list of all available components, run **llvm-config --components**. In most cases, components correspond directly to LLVM libraries. Useful “virtual” components include:

all

Includes all LLVM libraries. The default if no components are specified.

backend

Includes either a native backend or the C backend.

engine

Includes either a native JIT or the bitcode interpreter.

EXIT STATUS

If **llvm-config** succeeds, it will exit with 0. Otherwise, if an error occurs, it will exit with a non-zero value.

llvm-diff - LLVM structural ‘diff’

SYNOPSIS

llvm-diff [*options*] *module 1 module 2* [*global name ...*]

DESCRIPTION

llvm-diff compares the structure of two LLVM modules, primarily focusing on differences in function definitions. Insignificant differences, such as changes in the ordering of globals or in the names of local values, are ignored.

An input module will be interpreted as an assembly file if its name ends in ‘.ll’; otherwise it will be read in as a bitcode file.

If a list of global names is given, just the values with those names are compared; otherwise, all global values are compared, and diagnostics are produced for globals which only appear in one module or the other.

llvm-diff compares two functions by comparing their basic blocks, beginning with the entry blocks. If the terminators seem to match, then the corresponding successors are compared; otherwise they are ignored. This algorithm is very sensitive to changes in control flow, which tend to stop any downstream changes from being detected.

llvm-diff is intended as a debugging tool for writers of LLVM passes and frontends. It does not have a stable output format.

EXIT STATUS

If **llvm-diff** finds no differences between the modules, it will exit with 0 and produce no output. Otherwise it will exit with a non-zero value.

BUGS

Many important differences, like changes in linkage or function attributes, are not diagnosed.

Changes in memory behavior (for example, coalescing loads) can cause massive detected differences in blocks.

llvm-cov - emit coverage information

SYNOPSIS

llvm-cov *command* [*args...*]

DESCRIPTION

The **llvm-cov** tool shows code coverage information for programs that are instrumented to emit profile data. It can be used to work with gcov-style coverage or with clang's instrumentation based profiling.

If the program is invoked with a base name of gcov, it will behave as if the **llvm-cov gcov** command were called. Otherwise, a command should be provided.

COMMANDS

- *gcov*
- *show*
- *report*
- *export*

GCOV COMMAND

SYNOPSIS

llvm-cov gcov [*options*] *SOURCEFILE*

DESCRIPTION

The **llvm-cov gcov** tool reads code coverage data files and displays the coverage information for a specified source file. It is compatible with the gcov tool from version 4.2 of GCC and may also be compatible with some later versions of gcov.

To use **llvm-cov gcov**, you must first build an instrumented version of your application that collects coverage data as it runs. Compile with the `-fprofile-arcs` and `-ftest-coverage` options to add the instrumentation. (Alternatively, you can use the `--coverage` option, which includes both of those other options.) You should compile with debugging information (`-g`) and without optimization (`-O0`); otherwise, the coverage data cannot be accurately mapped back to the source code.

At the time you compile the instrumented code, a `.gcno` data file will be generated for each object file. These `.gcno` files contain half of the coverage data. The other half of the data comes from `.gda` files that are generated when you run the instrumented program, with a separate `.gda` file for each object file. Each time you run the program,

the execution counts are summed into any existing `.gcda` files, so be sure to remove any old files if you do not want their contents to be included.

By default, the `.gcda` files are written into the same directory as the object files, but you can override that by setting the `GCOV_PREFIX` and `GCOV_PREFIX_STRIP` environment variables. The `GCOV_PREFIX_STRIP` variable specifies a number of directory components to be removed from the start of the absolute path to the object file directory. After stripping those directories, the prefix from the `GCOV_PREFIX` variable is added. These environment variables allow you to run the instrumented program on a machine where the original object file directories are not accessible, but you will then need to copy the `.gcda` files back to the object file directories where `llvm-cov gcov` expects to find them.

Once you have generated the coverage data files, run `llvm-cov gcov` for each main source file where you want to examine the coverage results. This should be run from the same directory where you previously ran the compiler. The results for the specified source file are written to a file named by appending a `.gcov` suffix. A separate output file is also created for each file included by the main source file, also with a `.gcov` suffix added.

The basic content of an `.gcov` output file is a copy of the source file with an execution count and line number prepended to every line. The execution count is shown as `-` if a line does not contain any executable code. If a line contains code but that code was never executed, the count is displayed as `#####`.

OPTIONS

-a, --all-blocks

Display all basic blocks. If there are multiple blocks for a single line of source code, this option causes `llvm-cov` to show the count for each block instead of just one count for the entire line.

-b, --branch-probabilities

Display conditional branch probabilities and a summary of branch information.

-c, --branch-counts

Display branch counts instead of probabilities (requires `-b`).

-f, --function-summaries

Show a summary of coverage for each function instead of just one summary for an entire source file.

--help

Display available options (`-help-hidden` for more).

-l, --long-file-names

For coverage output of files included from the main source file, add the main file name followed by `##` as a prefix to the output file names. This can be combined with the `-preserve-paths` option to use complete paths for both the main file and the included file.

-n, --no-output

Do not output any `.gcov` files. Summary information is still displayed.

-o=<DIR|FILE>, --object-directory=<DIR>, --object-file=<FILE>

Find objects in `DIR` or based on `FILE`'s path. If you specify a particular object file, the coverage data files are expected to have the same base name with `.gcno` and `.gcda` extensions. If you specify a directory, the files are expected in that directory with the same base name as the source file.

-p, --preserve-paths

Preserve path components when naming the coverage output files. In addition to the source file name, include the directories from the path to that file. The directories are separate by `#` characters, with `.` directories removed and `..` directories replaced by `^` characters. When used with the `-long-file-names` option, this applies to both the main file name and the included file name.

-u, --unconditional-branches

Include unconditional branches in the output for the `-branch-probabilities` option.

-version

Display the version of `llvm-cov`.

EXIT STATUS

`llvm-cov gcov` returns 1 if it cannot read input files. Otherwise, it exits with zero.

SHOW COMMAND**SYNOPSIS**

```
llvm-cov show [options] -instr-profile PROFILE BIN [-object BIN,... ] [[-object BIN]] [SOURCES]
```

DESCRIPTION

The `llvm-cov show` command shows line by line coverage of the binaries *BIN*,... using the profile data *PROFILE*. It can optionally be filtered to only show the coverage for the files listed in *SOURCES*.

To use `llvm-cov show`, you need a program that is compiled with instrumentation to emit profile and coverage data. To build such a program with `clang` use the `-fprofile-instr-generate` and `-fcoverage-mapping` flags. If linking with the `clang` driver, pass `-fprofile-instr-generate` to the link stage to make sure the necessary runtime libraries are linked in.

The coverage information is stored in the built executable or library itself, and this is what you should pass to `llvm-cov show` as a *BIN* argument. The profile data is generated by running this instrumented program normally. When the program exits it will write out a raw profile file, typically called `default.profraw`, which can be converted to a format that is suitable for the *PROFILE* argument using the `llvm-profdata merge` tool.

OPTIONS**-show-line-counts**

Show the execution counts for each line. This is enabled by default, unless another `-show` option is used.

-show-expansions

Expand inclusions, such as preprocessor macros or textual inclusions, inline in the display of the source file.

-show-instantiations

For source regions that are instantiated multiple times, such as templates in C++, show each instantiation separately as well as the combined summary.

-show-regions

Show the execution counts for each region by displaying a caret that points to the character where the region starts.

-show-line-counts-or-regions

Show the execution counts for each line if there is only one region on the line, but show the individual regions if there are multiple on the line.

-use-color[=VALUE]

Enable or disable color output. By default this is autodetected.

-arch=<name>

If the covered binary is a universal binary, select the architecture to use. It is an error to specify an architecture that is not included in the universal binary or to use an architecture that does not match a non-universal binary.

-name=<NAME>

Show code coverage only for functions with the given name.

-name-regex=<PATTERN>

Show code coverage only for functions that match the given regular expression.

-format=<FORMAT>

Use the specified output format. The supported formats are: “text”, “html”.

-tab-size=<TABSIZ>

Replace tabs with <TABSIZ> spaces when preparing reports. Currently, this is only supported for the html format.

-output-dir=PATH

Specify a directory to write coverage reports into. If the directory does not exist, it is created. When used in function view mode (i.e when -name or -name-regex are used to select specific functions), the report is written to PATH/functions.EXTENSION. When used in file view mode, a report for each file is written to PATH/REL_PATH_TO_FILE.EXTENSION.

-Xdemangler=<TOOL> | <TOOL-OPTION>

Specify a symbol demangler. This can be used to make reports more human-readable. This option can be specified multiple times to supply arguments to the demangler (e.g *-Xdemangler c++filt -Xdemangler -n* for C++). The demangler is expected to read a newline-separated list of symbols from stdin and write a newline-separated list of the same length to stdout.

-num-threads=N, -j=N

Use N threads to write file reports (only applicable when -output-dir is specified). When N=0, llvm-cov auto-detects an appropriate number of threads to use. This is the default.

-line-coverage-gt=<N>

Show code coverage only for functions with line coverage greater than the given threshold.

-line-coverage-lt=<N>

Show code coverage only for functions with line coverage less than the given threshold.

-region-coverage-gt=<N>

Show code coverage only for functions with region coverage greater than the given threshold.

-region-coverage-lt=<N>

Show code coverage only for functions with region coverage less than the given threshold.

REPORT COMMAND

SYNOPSIS

```
llvm-cov report [options] -instr-profile PROFILE BIN [-object BIN,... ] [[-object BIN]] [SOURCES]
```

DESCRIPTION

The **llvm-cov report** command displays a summary of the coverage of the binaries *BIN*,... using the profile data *PROFILE*. It can optionally be filtered to only show the coverage for the files listed in *SOURCES*.

If no source files are provided, a summary line is printed for each file in the coverage data. If any files are provided, summaries are shown for each function in the listed files instead.

For information on compiling programs for coverage and generating profile data, see [SHOW COMMAND](#).

OPTIONS

-use-color [=VALUE]

Enable or disable color output. By default this is autodetected.

-arch=<name>

If the covered binary is a universal binary, select the architecture to use. It is an error to specify an architecture that is not included in the universal binary or to use an architecture that does not match a non-universal binary.

-show-functions

Show coverage summaries for each function.

EXPORT COMMAND

SYNOPSIS

```
llvm-cov export [options] -instr-profile PROFILE BIN [-object BIN,... ] [[-object BIN]]
```

DESCRIPTION

The **llvm-cov export** command exports regions, functions, expansions, and summaries of the coverage of the binaries *BIN*,... using the profile data *PROFILE* as JSON.

For information on compiling programs for coverage and generating profile data, see [SHOW COMMAND](#).

OPTIONS

-arch=<name>

If the covered binary is a universal binary, select the architecture to use. It is an error to specify an architecture that is not included in the universal binary or to use an architecture that does not match a non-universal binary.

llvm-profdata - Profile data tool

SYNOPSIS

```
llvm-profdata command [args...]
```

DESCRIPTION

The **llvm-profdata** tool is a small utility for working with profile data files.

COMMANDS

- *merge*
- *show*

MERGE

SYNOPSIS

```
llvm-profdata merge [options] [filename...]
```

DESCRIPTION

llvm-profdata merge takes several profile data files generated by PGO instrumentation and merges them together into a single indexed profile data file.

By default profile data is merged without modification. This means that the relative importance of each input file is proportional to the number of samples or counts it contains. In general, the input from a longer training run will be interpreted as relatively more important than a shorter run. Depending on the nature of the training runs it may be useful to adjust the weight given to each input file by using the `-weighted-input` option.

Profiles passed in via `-weighted-input`, `-input-files`, or via positional arguments are processed once for each time they are seen.

OPTIONS

-help

Print a summary of command line options.

-output=output, -o=output

Specify the output file name. *Output* cannot be `-` as the resulting indexed profile data can't be written to standard output.

-weighted-input=weight, filename

Specify an input file name along with a weight. The profile counts of the supplied *filename* will be scaled (multiplied) by the supplied *weight*, where *weight* is a decimal integer ≥ 1 . Input files specified without using this option are assigned a default weight of 1. Examples are shown below.

-input-files=path, -f=path

Specify a file which contains a list of files to merge. The entries in this file are newline-separated. Lines starting with `#` are skipped. Entries may be of the form `<filename>` or `<weight>,<filename>`.

-instr (default)

Specify that the input profile is an instrumentation-based profile.

-sample

Specify that the input profile is a sample-based profile.

The format of the generated file can be generated in one of three ways:

-binary (default)

Emit the profile using a binary encoding. For instrumentation-based profile the output format is the indexed binary format.

-text

Emit the profile in text mode. This option can also be used with both sample-based and instrumentation-based profile. When this option is used the profile will be dumped in the text format that is parsable by the profile reader.

-gcc

Emit the profile using GCC's gcov format (Not yet supported).

-sparse [=true|false]

Do not emit function records with 0 execution count. Can only be used in conjunction with -instr. Defaults to false, since it can inhibit compiler optimization during PGO.

-num-threads=N, **-j**=N

Use N threads to perform profile merging. When N=0, llvm-profdata auto-detects an appropriate number of threads to use. This is the default.

EXAMPLES

Basic Usage

Merge three profiles:

```
llvm-profdata merge foo.profdata bar.profdata baz.profdata -output merged.profdata
```

Weighted Input

The input file *foo.profdata* is especially important, multiply its counts by 10:

```
llvm-profdata merge -weighted-input=10,foo.profdata bar.profdata baz.profdata -output merged.profdata
```

Exactly equivalent to the previous invocation (explicit form; useful for programmatic invocation):

```
llvm-profdata merge -weighted-input=10,foo.profdata -weighted-input=1,bar.profdata -weighted-input=1,baz.profdata -output merged.profdata
```

SHOW

SYNOPSIS

```
llvm-profdata show [options] [filename]
```

DESCRIPTION

llvm-profdata show takes a profile data file and displays the information about the profile counters for this file and for any of the specified function(s).

If *filename* is omitted or is -, then **llvm-profdata show** reads its input from standard input.

OPTIONS

-all-functions

Print details for every function.

-counts

Print the counter values for the displayed functions.

- function=string**
Print details for a function if the function's name contains the given string.
- help**
Print a summary of command line options.
- output=output, -o=output**
Specify the output file name. If *output* is `-` or it isn't specified, then the output is sent to standard output.
- instr** (default)
Specify that the input profile is an instrumentation-based profile.
- text**
Instruct the profile dumper to show profile counts in the text format of the instrumentation-based profile data representation. By default, the profile information is dumped in a more human readable form (also in text) with annotations.
- topn=n**
Instruct the profile dumper to show the top *n* functions with the hottest basic blocks in the summary section. By default, the *topn* functions are not dumped.
- sample**
Specify that the input profile is a sample-based profile.
- memop-sizes**
Show the profiled sizes of the memory intrinsic calls for shown functions.

EXIT STATUS

llvm-profdata returns 1 if the command is omitted or is invalid, if it cannot read input files, or if there is a mismatch between their data.

llvm-stress - generate random .ll files

SYNOPSIS

llvm-stress [-size=filesize] [-seed=initialseed] [-o=outfile]

DESCRIPTION

The **llvm-stress** tool is used to generate random `.ll` files that can be used to test different components of LLVM.

OPTIONS

- o** filename
Specify the output filename.
- size** size
Specify the size of the generated `.ll` file.
- seed** seed
Specify the seed to be used for the randomly generated instructions.

EXIT STATUS

llvm-stress returns 0.

llvm-symbolizer - convert addresses into source code locations

SYNOPSIS

llvm-symbolizer [options]

DESCRIPTION

llvm-symbolizer reads object file names and addresses from standard input and prints corresponding source code locations to standard output. If object file is specified in command line, **llvm-symbolizer** processes only addresses from standard input, the rest is output verbatim. This program uses debug info sections and symbol table in the object files.

EXAMPLE

```

$ cat addr.txt
a.out 0x4004f4
/tmp/b.out 0x400528
/tmp/c.so 0x710
/tmp/mach_universal_binary:i386 0x1f84
/tmp/mach_universal_binary:x86_64 0x100000f24
$ llvm-symbolizer < addr.txt
main
/tmp/a.cc:4

f(int, int)
/tmp/b.cc:11

h_inlined_into_g
/tmp/header.h:2
g_inlined_into_f
/tmp/header.h:7
f_inlined_into_main
/tmp/source.cc:3
main
/tmp/source.cc:8

_main
/tmp/source_i386.cc:8

_main
/tmp/source_x86_64.cc:8
$ cat addr2.txt
0x4004f4
0x401000
$ llvm-symbolizer -obj=a.out < addr2.txt
main
/tmp/a.cc:4

```

```
foo(int)
/tmp/a.cc:12
$cat addr.txt
0x40054d
$llvm-symbolizer -inlining -print-address -pretty-print -obj=addr.exe < addr.txt
0x40054d: inc at /tmp/x.c:3:3
      (inlined by) main at /tmp/x.c:9:0
$llvm-symbolizer -inlining -pretty-print -obj=addr.exe < addr.txt
inc at /tmp/x.c:3:3
      (inlined by) main at /tmp/x.c:9:0
```

OPTIONS

-obj

Path to object file to be symbolized.

-functions=[none|short|linkage]

Specify the way function names are printed (omit function name, print short function name, or print full linkage name, respectively). Defaults to `linkage`.

-use-symbol-table

Prefer function names stored in symbol table to function names in debug info sections. Defaults to `true`.

-demangle

Print demangled function names. Defaults to `true`.

-inlining

If a source code location is in an inlined function, prints all the inlined frames. Defaults to `true`.

-default-arch

If a binary contains object files for multiple architectures (e.g. it is a Mach-O universal binary), symbolize the object file for a given architecture. You can also specify architecture by writing `binary_name:arch_name` in the input (see example above). If architecture is not specified in either way, address will not be symbolized. Defaults to empty string.

-dsym-hint=<path/to/file.dSYM>

(Darwin-only flag). If the debug info for a binary isn't present in the default location, look for the debug info at the `.dSYM` path provided via the `-dsym-hint` flag. This flag can be used multiple times.

-print-address

Print address before the source code location. Defaults to `false`.

-pretty-print

Print human readable output. If `-inlining` is specified, enclosing scope is prefixed by (inlined by). Refer to listed examples.

EXIT STATUS

`llvm-symbolizer` returns 0. Other exit codes imply internal program error.

llvm-dwarfdump - print contents of DWARF sections

SYNOPSIS

```
llvm-dwarfdump [options] [filenames...]
```

DESCRIPTION

llvm-dwarfdump parses DWARF sections in the object files and prints their contents in human-readable form.

OPTIONS

-debug-dump=section

Specify the DWARF section to dump. For example, use `abbrev` to dump the contents of `.debug_abbrev` section, `loc.dwo` to dump the contents of `.debug_loc.dwo` etc. See `llvm-dwarfdump --help` for the complete list of supported sections. Use `all` to dump all DWARF sections. It is the default.

EXIT STATUS

llvm-dwarfdump returns 0 if the input files were parsed and dumped successfully. Otherwise, it returns 1.

2.6.2 Debugging Tools

bugpoint - automatic test case reduction tool

SYNOPSIS

bugpoint [*options*] [*input LLVM ll/bc files*] [*LLVM passes*] **-args** *program arguments*

DESCRIPTION

bugpoint narrows down the source of problems in LLVM tools and passes. It can be used to debug three types of failures: optimizer crashes, miscompilations by optimizers, or bad native code generation (including problems in the static and JIT compilers). It aims to reduce large test cases to small, useful ones. For more information on the design and inner workings of **bugpoint**, as well as advice for using bugpoint, see [LLVM bugpoint tool: design and usage](#) in the LLVM distribution.

OPTIONS

-additional-so *library*

Load the dynamic shared object *library* into the test program whenever it is run. This is useful if you are debugging programs which depend on non-LLVM libraries (such as the X or curses libraries) to run.

-append-exit-code={*true,false*}

Append the test programs exit code to the output file so that a change in exit code is considered a test failure. Defaults to false.

-args *program args*

Pass all arguments specified after **-args** to the test program whenever it runs. Note that if any of the *program args* start with a “-“, you should use:

```
bugpoint [bugpoint args] --args -- [program args]
```

The “--” right after the **-args** option tells **bugpoint** to consider any options starting with “-” to be part of the **-args** option, not as options to **bugpoint** itself.

-tool-args *tool args*

Pass all arguments specified after **-tool-args** to the LLVM tool under test (**llc**, **lli**, etc.) whenever it runs. You should use this option in the following way:

```
bugpoint [bugpoint args] --tool-args -- [tool args]
```

The “--” right after the **-tool-args** option tells **bugpoint** to consider any options starting with “-” to be part of the **-tool-args** option, not as options to **bugpoint** itself. (See **-args**, above.)

-safe-tool-args *tool args*

Pass all arguments specified after **-safe-tool-args** to the “safe” execution tool.

-gcc-tool-args *gcc tool args*

Pass all arguments specified after **-gcc-tool-args** to the invocation of **gcc**.

-opt-args *opt args*

Pass all arguments specified after **-opt-args** to the invocation of **opt**.

-disable-{dce,simplifycfg}

Do not run the specified passes to clean up and reduce the size of the test program. By default, **bugpoint** uses these passes internally when attempting to reduce test programs. If you’re trying to find a bug in one of these passes, **bugpoint** may crash.

-enable-valgrind

Use valgrind to find faults in the optimization phase. This will allow bugpoint to find otherwise asymptomatic problems caused by memory mis-management.

-find-bugs

Continually randomize the specified passes and run them on the test program until a bug is found or the user kills **bugpoint**.

-help

Print a summary of command line options.

-input *filename*

Open *filename* and redirect the standard input of the test program, whenever it runs, to come from that file.

-load *plugin*

Load the dynamic object *plugin* into **bugpoint** itself. This object should register new optimization passes. Once loaded, the object will add new command line options to enable various optimizations. To see the new complete list of optimizations, use the **-help** and **-load** options together; for example:

```
bugpoint --load myNewPass.so -help
```

-mlimit *megabytes*

Specifies an upper limit on memory usage of the optimization and codegen. Set to zero to disable the limit.

-output *filename*

Whenever the test program produces output on its standard output stream, it should match the contents of *filename* (the “reference output”). If you do not use this option, **bugpoint** will attempt to generate a reference output by compiling the program with the “safe” backend and running it.

-run-{int,jit,llc,custom}

Whenever the test program is compiled, **bugpoint** should generate code for it using the specified code generator. These options allow you to choose the interpreter, the JIT compiler, the static native code compiler, or a custom command (see **-exec-command**) respectively.

-safe-{llc,custom}

When debugging a code generator, **bugpoint** should use the specified code generator as the “safe” code generator. This is a known-good code generator used to generate the “reference output” if it has not been provided, and to compile portions of the program that as they are excluded from the testcase. These options allow you to choose the static native code compiler, or a custom command, (see **-exec-command**) respectively. The interpreter and the JIT backends cannot currently be used as the “safe” backends.

-exec-command *command*

This option defines the command to use with the **-run-custom** and **-safe-custom** options to execute the bitcode testcase. This can be useful for cross-compilation.

-compile-command *command*

This option defines the command to use with the **-compile-custom** option to compile the bitcode testcase. The command should exit with a failure exit code if the file is “interesting” and should exit with a success exit code (i.e. 0) otherwise (this is the same as if it crashed on “interesting” inputs).

This can be useful for testing compiler output without running any link or execute stages. To generate a reduced unit test, you may add CHECK directives to the testcase and pass the name of an executable compile-command script in this form:

```
#!/bin/sh
llc "$@"
not FileCheck [bugpoint input file].ll < bugpoint-test-program.s
```

This script will “fail” as long as FileCheck passes. So the result will be the minimum bitcode that passes FileCheck.

-safe-path *path*

This option defines the path to the command to execute with the **-safe-{int,jit,llc,custom}** option.

-verbose-errors={true,false}

The default behavior of bugpoint is to print “<crash>” when it finds a reduced test that crashes compilation. This flag prints the output of the crashing program to stderr. This is useful to make sure it is the same error being tracked down and not a different error that happens to crash the compiler as well. Defaults to false.

EXIT STATUS

If **bugpoint** succeeds in finding a problem, it will exit with 0. Otherwise, if an error occurs, it will exit with a non-zero value.

SEE ALSO

optlopt

llvm-extract - extract a function from an LLVM module

SYNOPSIS

llvm-extract [*options*] **-func** *function-name* [*filename*]

DESCRIPTION

The **llvm-extract** command takes the name of a function and extracts it from the specified LLVM bitcode file. It is primarily used as a debugging tool to reduce test cases from larger programs that are triggering a bug.

In addition to extracting the bitcode of the specified function, **llvm-extract** will also remove unreachable global variables, prototypes, and unused types.

The **llvm-extract** command reads its input from standard input if filename is omitted or if filename is **-**. The output is always written to standard output, unless the **-o** option is specified (see below).

OPTIONS

-f

Enable binary output on terminals. Normally, **llvm-extract** will refuse to write raw bitcode output if the output stream is a terminal. With this option, **llvm-extract** will write raw bitcode regardless of the output device.

-func *function-name*

Extract the function named *function-name* from the LLVM bitcode. May be specified multiple times to extract multiple functions at once.

-rfunc *function-regular-expr*

Extract the function(s) matching *function-regular-expr* from the LLVM bitcode. All functions matching the regular expression will be extracted. May be specified multiple times.

-glob *global-name*

Extract the global variable named *global-name* from the LLVM bitcode. May be specified multiple times to extract multiple global variables at once.

-rglob *glob-regular-expr*

Extract the global variable(s) matching *global-regular-expr* from the LLVM bitcode. All global variables matching the regular expression will be extracted. May be specified multiple times.

-help

Print a summary of command line options.

-o *filename*

Specify the output filename. If filename is **-** (the default), then **llvm-extract** sends its output to standard output.

-S

Write output in LLVM intermediate language (instead of bitcode).

EXIT STATUS

If **llvm-extract** succeeds, it will exit with 0. Otherwise, if an error occurs, it will exit with a non-zero value.

SEE ALSO

bugpoint

llvm-bcanalyzer - LLVM bitcode analyzer

SYNOPSIS

llvm-bcanalyzer [*options*] [*filename*]

DESCRIPTION

The **llvm-bcanalyzer** command is a small utility for analyzing bitcode files. The tool reads a bitcode file (such as generated with the **llvm-as** tool) and produces a statistical report on the contents of the bitcode file. The tool can also dump a low level but human readable version of the bitcode file. This tool is probably not of much interest or utility except for those working directly with the bitcode file format. Most LLVM users can just ignore this tool.

If *filename* is omitted or is **-**, then **llvm-bcanalyzer** reads its input from standard input. This is useful for combining the tool into a pipeline. Output is written to the standard output.

OPTIONS

-nodetails

Causes **llvm-bcanalyzer** to abbreviate its output by writing out only a module level summary. The details for individual functions are not displayed.

-dump

Causes **llvm-bcanalyzer** to dump the bitcode in a human readable format. This format is significantly different from LLVM assembly and provides details about the encoding of the bitcode file.

-verify

Causes **llvm-bcanalyzer** to verify the module produced by reading the bitcode. This ensures that the statistics generated are based on a consistent module.

-help

Print a summary of command line options.

EXIT STATUS

If **llvm-bcanalyzer** succeeds, it will exit with 0. Otherwise, if an error occurs, it will exit with a non-zero value, usually 1.

SUMMARY OUTPUT DEFINITIONS

The following items are always printed by `llvm-bcanalyzer`. They comprise the summary output.

Bitcode Analysis Of Module

This just provides the name of the module for which bitcode analysis is being generated.

Bitcode Version Number

The bitcode version (not LLVM version) of the file read by the analyzer.

File Size

The size, in bytes, of the entire bitcode file.

Module Bytes

The size, in bytes, of the module block. Percentage is relative to File Size.

Function Bytes

The size, in bytes, of all the function blocks. Percentage is relative to File Size.

Global Types Bytes

The size, in bytes, of the Global Types Pool. Percentage is relative to File Size. This is the size of the definitions of all types in the bitcode file.

Constant Pool Bytes

The size, in bytes, of the Constant Pool Blocks Percentage is relative to File Size.

Module Globals Bytes

This size, in bytes, of the Global Variable Definitions and their initializers. Percentage is relative to File Size.

Instruction List Bytes

The size, in bytes, of all the instruction lists in all the functions. Percentage is relative to File Size. Note that this value is also included in the Function Bytes.

Compaction Table Bytes

The size, in bytes, of all the compaction tables in all the functions. Percentage is relative to File Size. Note that this value is also included in the Function Bytes.

Symbol Table Bytes

The size, in bytes, of all the symbol tables in all the functions. Percentage is relative to File Size. Note that this value is also included in the Function Bytes.

Dependent Libraries Bytes

The size, in bytes, of the list of dependent libraries in the module. Percentage is relative to File Size. Note that this value is also included in the Module Global Bytes.

Number Of Bitcode Blocks

The total number of blocks of any kind in the bitcode file.

Number Of Functions

The total number of function definitions in the bitcode file.

Number Of Types

The total number of types defined in the Global Types Pool.

Number Of Constants

The total number of constants (of any type) defined in the Constant Pool.

Number Of Basic Blocks

The total number of basic blocks defined in all functions in the bitcode file.

Number Of Instructions

The total number of instructions defined in all functions in the bitcode file.

Number Of Long Instructions

The total number of long instructions defined in all functions in the bitcode file. Long instructions are those taking greater than 4 bytes. Typically long instructions are GetElementPtr with several indices, PHI nodes, and calls to functions with large numbers of arguments.

Number Of Operands

The total number of operands used in all instructions in the bitcode file.

Number Of Compaction Tables

The total number of compaction tables in all functions in the bitcode file.

Number Of Symbol Tables

The total number of symbol tables in all functions in the bitcode file.

Number Of Dependent Libs

The total number of dependent libraries found in the bitcode file.

Total Instruction Size

The total size of the instructions in all functions in the bitcode file.

Average Instruction Size

The average number of bytes per instruction across all functions in the bitcode file. This value is computed by dividing Total Instruction Size by Number Of Instructions.

Maximum Type Slot Number

The maximum value used for a type's slot number. Larger slot number values take more bytes to encode.

Maximum Value Slot Number

The maximum value used for a value's slot number. Larger slot number values take more bytes to encode.

Bytes Per Value

The average size of a Value definition (of any type). This is computed by dividing File Size by the total number of values of any type.

Bytes Per Global

The average size of a global definition (constants and global variables).

Bytes Per Function

The average number of bytes per function definition. This is computed by dividing Function Bytes by Number Of Functions.

of VBR 32-bit Integers

The total number of 32-bit integers encoded using the Variable Bit Rate encoding scheme.

of VBR 64-bit Integers

The total number of 64-bit integers encoded using the Variable Bit Rate encoding scheme.

of VBR Compressed Bytes

The total number of bytes consumed by the 32-bit and 64-bit integers that use the Variable Bit Rate encoding scheme.

of VBR Expanded Bytes

The total number of bytes that would have been consumed by the 32-bit and 64-bit integers had they not been compressed with the Variable Bit Rate encoding scheme.

Bytes Saved With VBR

The total number of bytes saved by using the Variable Bit Rate encoding scheme. The percentage is relative to # of VBR Expanded Bytes.

DETAILED OUTPUT DEFINITIONS

The following definitions occur only if the `-nodetails` option was not given. The detailed output provides additional information on a per-function basis.

Type

The type signature of the function.

Byte Size

The total number of bytes in the function's block.

Basic Blocks

The number of basic blocks defined by the function.

Instructions

The number of instructions defined by the function.

Long Instructions

The number of instructions using the long instruction format in the function.

Operands

The number of operands used by all instructions in the function.

Instruction Size

The number of bytes consumed by instructions in the function.

Average Instruction Size

The average number of bytes consumed by the instructions in the function. This value is computed by dividing Instruction Size by Instructions.

Bytes Per Instruction

The average number of bytes used by the function per instruction. This value is computed by dividing Byte Size by Instructions. Note that this is not the same as Average Instruction Size. It computes a number relative to the total function size not just the size of the instruction list.

Number of VBR 32-bit Integers

The total number of 32-bit integers found in this function (for any use).

Number of VBR 64-bit Integers

The total number of 64-bit integers found in this function (for any use).

Number of VBR Compressed Bytes

The total number of bytes in this function consumed by the 32-bit and 64-bit integers that use the Variable Bit Rate encoding scheme.

Number of VBR Expanded Bytes

The total number of bytes in this function that would have been consumed by the 32-bit and 64-bit integers had they not been compressed with the Variable Bit Rate encoding scheme.

Bytes Saved With VBR

The total number of bytes saved in this function by using the Variable Bit Rate encoding scheme. The percentage is relative to # of VBR Expanded Bytes.

SEE ALSO

llvm-dis - LLVM disassembler, *LLVM Bitcode File Format*

2.6.3 Developer Tools

FileCheck - Flexible pattern matching file verifier

SYNOPSIS

FileCheck *match-filename* [*--check-prefix=XXX*] [*--strict-whitespace*]

DESCRIPTION

FileCheck reads two files (one from standard input, and one specified on the command line) and uses one to verify the other. This behavior is particularly useful for the testsuite, which wants to verify that the output of some tool (e.g. **l1c**) contains the expected information (for example, a `movsd` from `esp` or whatever is interesting). This is similar to using **grep**, but it is optimized for matching multiple different inputs in one file in a specific order.

The *match-filename* file specifies the file that contains the patterns to match. The file to verify is read from standard input unless the *--input-file* option is used.

OPTIONS

-help

Print a summary of command line options.

--check-prefix *prefix*

FileCheck searches the contents of *match-filename* for patterns to match. By default, these patterns are prefixed with "CHECK:". If you'd like to use a different prefix (e.g. because the same input file is checking multiple different tool or options), the *--check-prefix* argument allows you to specify one or more prefixes to match. Multiple prefixes are useful for tests which might change for different run options, but most lines remain the same.

--check-prefixes *prefix1, prefix2, ...*

An alias of *--check-prefix* that allows multiple prefixes to be specified as a comma separated list.

--input-file filename
File to check (defaults to stdin).

--match-full-lines
By default, FileCheck allows matches of anywhere on a line. This option will require all positive matches to cover an entire line. Leading and trailing whitespace is ignored, unless `--strict-whitespace` is also specified. (Note: negative matches from CHECK-NOT are not affected by this option!)

Passing this option is equivalent to inserting `{{^ *}}` or `{{^}}` before, and `{{ *$}}` or `{{ $}}` after every positive check pattern.

--strict-whitespace
By default, FileCheck canonicalizes input horizontal whitespace (spaces and tabs) which causes it to ignore these differences (a space will match a tab). The `--strict-whitespace` argument disables this behavior. End-of-line sequences are canonicalized to UNIX-style `\n` in all modes.

--implicit-check-not check-pattern
Adds implicit negative checks for the specified patterns between positive checks. The option allows writing stricter tests without stuffing them with CHECK-NOTs.

For example, “`--implicit-check-not warning:`” can be useful when testing diagnostic messages from tools that don’t have an option similar to `clang -verify`. With this option FileCheck will verify that input does not contain warnings not covered by any `CHECK:` patterns.

--enable-var-scope
Enables scope for regex variables.

Variables with names that start with `$` are considered global and remain set throughout the file.

All other variables get undefined after each encountered `CHECK-LABEL`.

-version
Show the version number of this program.

EXIT STATUS

If **FileCheck** verifies that the file matches the expected contents, it exits with 0. Otherwise, if not, or if an error occurs, it will exit with a non-zero value.

TUTORIAL

FileCheck is typically used from LLVM regression tests, being invoked on the RUN line of the test. A simple example of using FileCheck from a RUN line looks like this:

```
; RUN: llvm-as < %s | llc -march=x86-64 | FileCheck %s
```

This syntax says to pipe the current file (“%s”) into `llvm-as`, pipe that into `llc`, then pipe the output of `llc` into `FileCheck`. This means that `FileCheck` will be verifying its standard input (the `llc` output) against the filename argument specified (the original `.ll` file specified by “%s”). To see how this works, let’s look at the rest of the `.ll` file (after the RUN line):

```
define void @sub1(i32* %p, i32 %v) {  
entry:  
; CHECK: sub1:  
; CHECK: sub1  
    %0 = tail call i32 @llvm.atomic.load.sub.i32.p0i32(i32* %p, i32 %v)  
    ret void
```

```

}

define void @inc4(i64* %p) {
entry:
; CHECK: inc4:
; CHECK: incq
    %0 = tail call i64 @llvm.atomic.load.add.i64.p0i64(i64* %p, i64 1)
    ret void
}

```

Here you can see some “CHECK:” lines specified in comments. Now you can see how the file is piped into `llvm-as`, then `llc`, and the machine code output is what we are verifying. FileCheck checks the machine code output to verify that it matches what the “CHECK:” lines specify.

The syntax of the “CHECK:” lines is very simple: they are fixed strings that must occur in order. FileCheck defaults to ignoring horizontal whitespace differences (e.g. a space is allowed to match a tab) but otherwise, the contents of the “CHECK:” line is required to match some thing in the test file exactly.

One nice thing about FileCheck (compared to `grep`) is that it allows merging test cases together into logical groups. For example, because the test above is checking for the “sub1:” and “inc4:” labels, it will not match unless there is a “sub1” in between those labels. If it existed somewhere else in the file, that would not count: “`grep sub1`” matches if “sub1” exists anywhere in the file.

The FileCheck -check-prefix option

The FileCheck `-check-prefix` option allows multiple test configurations to be driven from one `.ll` file. This is useful in many circumstances, for example, testing different architectural variants with `llc`. Here’s a simple example:

```

; RUN: llvm-as < %s | llc -mtriple=i686-apple-darwin9 -mattr=sse41 \
; RUN:                               | FileCheck %s -check-prefix=X32
; RUN: llvm-as < %s | llc -mtriple=x86_64-apple-darwin9 -mattr=sse41 \
; RUN:                               | FileCheck %s -check-prefix=X64

define <4 x i32> @pinsrd_1(i32 %s, <4 x i32> %tmp) nounwind {
    %tmp1 = insertelement <4 x i32>; %tmp, i32 %s, i32 1
    ret <4 x i32> %tmp1
; X32: pinsrd_1:
; X32:     pinsrd $1, 4(%esp), %xmm0

; X64: pinsrd_1:
; X64:     pinsrd $1, %edi, %xmm0
}

```

In this case, we’re testing that we get the expected code generation with both 32-bit and 64-bit code generation.

The “CHECK-NEXT:” directive

Sometimes you want to match lines and would like to verify that matches happen on exactly consecutive lines with no other lines in between them. In this case, you can use “CHECK:” and “CHECK-NEXT:” directives to specify this. If you specified a custom check prefix, just use “<PREFIX>-NEXT:”. For example, something like this works as you’d expect:

```

define void @t2(<2 x double>* %r, <2 x double>* %A, double %B) {
    %tmp3 = load <2 x double>* %A, align 16
    %tmp7 = insertelement <2 x double> undef, double %B, i32 0
}

```

```

    %tmp9 = shufflevector <2 x double> %tmp3,
                        <2 x double> %tmp7,
                        <2 x i32> < i32 0, i32 2 >
    store <2 x double> %tmp9, <2 x double>* %r, align 16
    ret void

; CHECK:          t2:
; CHECK:          movl    8(%esp), %eax
; CHECK-NEXT:     movapd  (%eax), %xmm0
; CHECK-NEXT:     movhpd  12(%esp), %xmm0
; CHECK-NEXT:     movl    4(%esp), %eax
; CHECK-NEXT:     movapd  %xmm0, (%eax)
; CHECK-NEXT:     ret
}

```

“CHECK-NEXT:” directives reject the input unless there is exactly one newline between it and the previous directive. A “CHECK-NEXT:” cannot be the first directive in a file.

The “CHECK-SAME:” directive

Sometimes you want to match lines and would like to verify that matches happen on the same line as the previous match. In this case, you can use “CHECK:” and “CHECK-SAME:” directives to specify this. If you specified a custom check prefix, just use “<PREFIX>-SAME:”.

“CHECK-SAME:” is particularly powerful in conjunction with “CHECK-NOT:” (described below).

For example, the following works like you’d expect:

```

!0 = !DILocation(line: 5, scope: !1, inlinedAt: !2)

; CHECK:          !DILocation(line: 5,
; CHECK-NOT:      column:
; CHECK-SAME:     scope: ![[SCOPE:[0-9]+]]

```

“CHECK-SAME:” directives reject the input if there are any newlines between it and the previous directive. A “CHECK-SAME:” cannot be the first directive in a file.

The “CHECK-NOT:” directive

The “CHECK-NOT:” directive is used to verify that a string doesn’t occur between two matches (or before the first match, or after the last match). For example, to verify that a load is removed by a transformation, a test like this can be used:

```

define i8 @coerce_offset0(i32 %V, i32* %P) {
    store i32 %V, i32* %P

    %P2 = bitcast i32* %P to i8*
    %P3 = getelementptr i8* %P2, i32 2

    %A = load i8* %P3
    ret i8 %A
; CHECK: @coerce_offset0
; CHECK-NOT: load
; CHECK: ret i8
}

```

The “CHECK-DAG:” directive

If it’s necessary to match strings that don’t occur in a strictly sequential order, “CHECK-DAG:” could be used to verify them between two matches (or before the first match, or after the last match). For example, clang emits vtable globals in reverse order. Using CHECK-DAG:, we can keep the checks in the natural order:

```
// RUN: %clang_cc1 %s -emit-llvm -o - | FileCheck %s

struct Foo { virtual void method(); };
Foo f; // emit vtable
// CHECK-DAG: @_ZTV3Foo =

struct Bar { virtual void method(); };
Bar b;
// CHECK-DAG: @_ZTV3Bar =
```

CHECK-NOT: directives could be mixed with CHECK-DAG: directives to exclude strings between the surrounding CHECK-DAG: directives. As a result, the surrounding CHECK-DAG: directives cannot be reordered, i.e. all occurrences matching CHECK-DAG: before CHECK-NOT: must not fall behind occurrences matching CHECK-DAG: after CHECK-NOT:. For example,

```
; CHECK-DAG: BEFORE
; CHECK-NOT: NOT
; CHECK-DAG: AFTER
```

This case will reject input strings where BEFORE occurs after AFTER.

With captured variables, CHECK-DAG: is able to match valid topological orderings of a DAG with edges from the definition of a variable to its use. It’s useful, e.g., when your test cases need to match different output sequences from the instruction scheduler. For example,

```
; CHECK-DAG: add [[REG1:r[0-9]+]], r1, r2
; CHECK-DAG: add [[REG2:r[0-9]+]], r3, r4
; CHECK:      mul r5, [[REG1]], [[REG2]]
```

In this case, any order of that two add instructions will be allowed.

If you are defining *and* using variables in the same CHECK-DAG: block, be aware that the definition rule can match *after* its use.

So, for instance, the code below will pass:

```
; CHECK-DAG: vmov.32 [[REG2:d[0-9]+]] [0]
; CHECK-DAG: vmov.32 [[REG2]] [1]
vmov.32 d0[1]
vmov.32 d0[0]
```

While this other code, will not:

```
; CHECK-DAG: vmov.32 [[REG2:d[0-9]+]] [0]
; CHECK-DAG: vmov.32 [[REG2]] [1]
vmov.32 d1[1]
vmov.32 d0[0]
```

While this can be very useful, it’s also dangerous, because in the case of register sequence, you must have a strong order (read before write, copy before use, etc). If the definition your test is looking for doesn’t match (because of a bug in the compiler), it may match further away from the use, and mask real bugs away.

In those cases, to enforce the order, use a non-DAG directive between DAG-blocks.

The “CHECK-LABEL:” directive

Sometimes in a file containing multiple tests divided into logical blocks, one or more `CHECK:` directives may inadvertently succeed by matching lines in a later block. While an error will usually eventually be generated, the check flagged as causing the error may not actually bear any relationship to the actual source of the problem.

In order to produce better error messages in these cases, the “CHECK-LABEL:” directive can be used. It is treated identically to a normal `CHECK` directive except that FileCheck makes an additional assumption that a line matched by the directive cannot also be matched by any other check present in `match-filename`; this is intended to be used for lines containing labels or other unique identifiers. Conceptually, the presence of `CHECK-LABEL` divides the input stream into separate blocks, each of which is processed independently, preventing a `CHECK:` directive in one block matching a line in another block. If `--enable-var-scope` is in effect, all local variables are cleared at the beginning of the block.

For example,

```
define %struct.C* @C_ctor_base(%struct.C* %this, i32 %x) {
entry:
; CHECK-LABEL: C_ctor_base:
; CHECK: mov [[SAVETHIS:r[0-9]+]], r0
; CHECK: bl A_ctor_base
; CHECK: mov r0, [[SAVETHIS]]
%0 = bitcast %struct.C* %this to %struct.A*
%call = tail call %struct.A* @A_ctor_base(%struct.A* %0)
%1 = bitcast %struct.C* %this to %struct.B*
%call2 = tail call %struct.B* @B_ctor_base(%struct.B* %1, i32 %x)
ret %struct.C* %this
}

define %struct.D* @D_ctor_base(%struct.D* %this, i32 %x) {
entry:
; CHECK-LABEL: D_ctor_base:
```

The use of `CHECK-LABEL:` directives in this case ensures that the three `CHECK:` directives only accept lines corresponding to the body of the `@C_ctor_base` function, even if the patterns match lines found later in the file. Furthermore, if one of these three `CHECK:` directives fail, FileCheck will recover by continuing to the next block, allowing multiple test failures to be detected in a single invocation.

There is no requirement that `CHECK-LABEL:` directives contain strings that correspond to actual syntactic labels in a source or output language: they must simply uniquely match a single line in the file being verified.

`CHECK-LABEL:` directives cannot contain variable definitions or uses.

FileCheck Pattern Matching Syntax

All FileCheck directives take a pattern to match. For most uses of FileCheck, fixed string matching is perfectly sufficient. For some things, a more flexible form of matching is desired. To support this, FileCheck allows you to specify regular expressions in matching strings, surrounded by double braces: `{{yourregex}}`. Because we want to use fixed string matching for a majority of what we do, FileCheck has been designed to support mixing and matching fixed string matching with regular expressions. This allows you to write things like this:

```
; CHECK: movhpd      {{[0-9]+}}(%esp), {{%xmm[0-7]}}
```

In this case, any offset from the ESP register will be allowed, and any xmm register will be allowed.

Because regular expressions are enclosed with double braces, they are visually distinct, and you don’t need to use escape characters within the double braces like you would in C. In the rare case that you want to match double braces

explicitly from the input, you can use something ugly like `{ { [{ } [{ }] }` as your pattern.

FileCheck Variables

It is often useful to match a pattern and then verify that it occurs again later in the file. For codegen tests, this can be useful to allow any register, but verify that that register is used consistently later. To do this, **FileCheck** allows named variables to be defined and substituted into patterns. Here is a simple example:

```
; CHECK: test5:
; CHECK:      notw      [[REGISTER:%[a-z]+]]
; CHECK:      andw      {{.*}} [[REGISTER]]
```

The first check line matches a regex `%[a-z]+` and captures it into the variable `REGISTER`. The second line verifies that whatever is in `REGISTER` occurs later in the file after an “andw”. **FileCheck** variable references are always contained in `[[]]` pairs, and their names can be formed with the regex `[a-zA-Z][a-zA-Z0-9]*`. If a colon follows the name, then it is a definition of the variable; otherwise, it is a use.

FileCheck variables can be defined multiple times, and uses always get the latest value. Variables can also be used later on the same line they were defined on. For example:

```
; CHECK: op [[REG:r[0-9]+]], [[REG]]
```

Can be useful if you want the operands of `op` to be the same register, and don’t care exactly which register it is.

If `--enable-var-scope` is in effect, variables with names that start with `$` are considered to be global. All others variables are local. All local variables get undefined at the beginning of each CHECK-LABEL block. Global variables are not affected by CHECK-LABEL. This makes it easier to ensure that individual tests are not affected by variables set in preceding tests.

FileCheck Expressions

Sometimes there’s a need to verify output which refers line numbers of the match file, e.g. when testing compiler diagnostics. This introduces a certain fragility of the match file structure, as “CHECK:” lines contain absolute line numbers in the same file, which have to be updated whenever line numbers change due to text addition or deletion.

To support this case, FileCheck allows using `[@LINE]`, `[@LINE+<offset>]`, `[@LINE-<offset>]` expressions in patterns. These expressions expand to a number of the line where a pattern is located (with an optional integer offset).

This way match patterns can be put near the relevant test lines and include relative line number references, for example:

```
// CHECK: test.cpp:[[@LINE+4]]:6: error: expected ';' after top level declarator
// CHECK-NEXT: {{^int a}}
// CHECK-NEXT: {{^      \^}}
// CHECK-NEXT: {{^      ;}}
int a
```

Matching Newline Characters

To match newline characters in regular expressions the character class `[[:space:]]` can be used. For example, the following pattern:

```
// CHECK: DW_AT_location [DW_FORM_sec_offset] ([[DLOC:0x[0-9a-f]+]]){{[[:space:]].*}}
↪ "intd"
```

matches output of the form (from `llvm-dwarfdump`):

```
DW_AT_location [DW_FORM_sec_offset] (0x00000233)
DW_AT_name [DW_FORM_strp] (.debug_str[0x000000c9] = "intd")
```

letting us set the **FileCheck** variable `DLOC` to the desired value `0x00000233`, extracted from the line immediately preceding `"intd"`.

tblgen - Target Description To C++ Code Generator

SYNOPSIS

tblgen [*options*] [*filename*]

DESCRIPTION

tblgen translates from target description (`.td`) files into C++ code that can be included in the definition of an LLVM target library. Most users of LLVM will not need to use this program. It is only for assisting with writing an LLVM target backend.

The input and output of **tblgen** is beyond the scope of this short introduction; please see the [introduction to TableGen](#).

The *filename* argument specifies the name of a Target Description (`.td`) file to read as input.

OPTIONS

-help

Print a summary of command line options.

-o filename

Specify the output file name. If *filename* is `-`, then **tblgen** sends its output to standard output.

-I directory

Specify where to find other target description files for inclusion. The *directory* value should be a full or partial path to a directory that contains target description files.

-asmparsernum N

Make `-gen-asm-parser` emit assembly writer number *N*.

-asmwriternum N

Make `-gen-asm-writer` emit assembly writer number *N*.

-class className

Print the enumeration list for this class.

-print-records

Print all records to standard output (default).

-print-enums

Print enumeration values for a class.

-print-sets

Print expanded sets for testing DAG exprs.

-gen-emitter

Generate machine code emitter.

-gen-register-info
Generate registers and register classes info.

-gen-instr-info
Generate instruction descriptions.

-gen-asm-writer
Generate the assembly writer.

-gen-disassembler
Generate disassembler.

-gen-pseudo-lowering
Generate pseudo instruction lowering.

-gen-dag-isel
Generate a DAG (Directed Acycle Graph) instruction selector.

-gen-asm-matcher
Generate assembly instruction matcher.

-gen-dfa-packetizer
Generate DFA Packetizer for VLIW targets.

-gen-fast-isel
Generate a “fast” instruction selector.

-gen-subtarget
Generate subtarget enumerations.

-gen-intrinsic
Generate intrinsic information.

-gen-tgt-intrinsic
Generate target intrinsic information.

-gen-enhanced-disassembly-info
Generate enhanced disassembly info.

-version
Show the version number of this program.

EXIT STATUS

If **tblgen** succeeds, it will exit with 0. Otherwise, if an error occurs, it will exit with a non-zero value.

lit - LLVM Integrated Tester

SYNOPSIS

lit [*options*] [*tests*]

DESCRIPTION

lit is a portable tool for executing LLVM and Clang style test suites, summarizing their results, and providing indication of failures. **lit** is designed to be a lightweight testing tool with as simple a user interface as possible.

lit should be run with one or more *tests* to run specified on the command line. Tests can be either individual test files or directories to search for tests (see [TEST DISCOVERY](#)).

Each specified test will be executed (potentially in parallel) and once all tests have been run **lit** will print summary information on the number of tests which passed or failed (see [TEST STATUS RESULTS](#)). The **lit** program will execute with a non-zero exit code if any tests fail.

By default **lit** will use a succinct progress display and will only print summary information for test failures. See [OUTPUT OPTIONS](#) for options controlling the **lit** progress display and output.

lit also includes a number of options for controlling how tests are executed (specific features may depend on the particular test format). See [EXECUTION OPTIONS](#) for more information.

Finally, **lit** also supports additional options for only running a subset of the options specified on the command line, see [SELECTION OPTIONS](#) for more information.

Users interested in the **lit** architecture or designing a **lit** testing implementation should see [LIT INFRASTRUCTURE](#).

GENERAL OPTIONS

-h, --help

Show the **lit** help message.

-j N, --threads=N

Run *N* tests in parallel. By default, this is automatically chosen to match the number of detected available CPUs.

--config-prefix=NAME

Search for *NAME.cfg* and *NAME.site.cfg* when searching for test suites, instead of *lit.cfg* and *lit.site.cfg*.

-D NAME [=VALUE], --param NAME [=VALUE]

Add a user defined parameter *NAME* with the given *VALUE* (or the empty string if not given). The meaning and use of these parameters is test suite dependent.

OUTPUT OPTIONS

-q, --quiet

Suppress any output except for test failures.

-s, --succinct

Show less output, for example don't show information on tests that pass.

-v, --verbose

Show more information on test failures, for example the entire test output instead of just the test result.

-vv, --echo-all-commands

Echo all commands to stdout, as they are being executed. This can be valuable for debugging test failures, as the last echoed command will be the one which has failed. This option implies **--verbose**.

-a, --show-all

Show more information about all tests, for example the entire test commandline and output.

--no-progress-bar

Do not use curses based progress bar.

--show-unsupported

Show the names of unsupported tests.

--show-xfail

Show the names of tests that were expected to fail.

EXECUTION OPTIONS**--path=PATH**

Specify an additional `PATH` to use when searching for executables in tests.

--vg

Run individual tests under valgrind (using the memcheck tool). The `--error-exitcode` argument for valgrind is used so that valgrind failures will cause the program to exit with a non-zero status.

When this option is enabled, `lit` will also automatically provide a “valgrind” feature that can be used to conditionally disable (or expect failure in) certain tests.

--vg-arg=ARG

When `--vg` is used, specify an additional argument to pass to **valgrind** itself.

--vg-leak

When `--vg` is used, enable memory leak checks. When this option is enabled, `lit` will also automatically provide a “vg_leak” feature that can be used to conditionally disable (or expect failure in) certain tests.

--time-tests

Track the wall time individual tests take to execute and includes the results in the summary output. This is useful for determining which tests in a test suite take the most time to execute. Note that this option is most useful with `-j 1`.

SELECTION OPTIONS**--max-tests=N**

Run at most `N` tests and then terminate.

--max-time=N

Spend at most `N` seconds (approximately) running tests and then terminate.

--shuffle

Run the tests in a random order.

--num-shards=M

Divide the set of selected tests into `M` equal-sized subsets or “shards”, and run only one of them. Must be used with the `--run-shard=N` option, which selects the shard to run. The environment variable `LIT_NUM_SHARDS` can also be used in place of this option. These two options provide a coarse mechanism for partitioning large test suites, for parallel execution on separate machines (say in a large testing farm).

--run-shard=N

Select which shard to run, assuming the `--num-shards=M` option was provided. The two options must be used together, and the value of `N` must be in the range `1..M`. The environment variable `LIT_RUN_SHARD` can also be used in place of this option.

--filter=REGEXP

Run only those tests whose name matches the regular expression specified in `REGEXP`. The environment variable `LIT_FILTER` can be also used in place of this option, which is especially useful in environments where the call to `lit` is issued indirectly.

ADDITIONAL OPTIONS

--debug

Run **lit** in debug mode, for debugging configuration issues and **lit** itself.

--show-suites

List the discovered test suites and exit.

--show-tests

List all of the discovered tests and exit.

EXIT STATUS

lit will exit with an exit code of 1 if there are any FAIL or XPASS results. Otherwise, it will exit with the status 0. Other exit codes are used for non-test related failures (for example a user error or an internal program error).

TEST DISCOVERY

The inputs passed to **lit** can be either individual tests, or entire directories or hierarchies of tests to run. When **lit** starts up, the first thing it does is convert the inputs into a complete list of tests to run as part of *test discovery*.

In the **lit** model, every test must exist inside some *test suite*. **lit** resolves the inputs specified on the command line to test suites by searching upwards from the input path until it finds a `lit.cfg` or `lit.site.cfg` file. These files serve as both a marker of test suites and as configuration files which **lit** loads in order to understand how to find and run the tests inside the test suite.

Once **lit** has mapped the inputs into test suites it traverses the list of inputs adding tests for individual files and recursively searching for tests in directories.

This behavior makes it easy to specify a subset of tests to run, while still allowing the test suite configuration to control exactly how tests are interpreted. In addition, **lit** always identifies tests by the test suite they are in, and their relative path inside the test suite. For appropriately configured projects, this allows **lit** to provide convenient and flexible support for out-of-tree builds.

TEST STATUS RESULTS

Each test ultimately produces one of the following six results:

PASS

The test succeeded.

XFAIL

The test failed, but that is expected. This is used for test formats which allow specifying that a test does not currently work, but wish to leave it in the test suite.

XPASS

The test succeeded, but it was expected to fail. This is used for tests which were specified as expected to fail, but are now succeeding (generally because the feature they test was broken and has been fixed).

FAIL

The test failed.

UNRESOLVED

The test result could not be determined. For example, this occurs when the test could not be run, the test itself is invalid, or the test was interrupted.

UNSUPPORTED

The test is not supported in this environment. This is used by test formats which can report unsupported tests.

Depending on the test format tests may produce additional information about their status (generally only for failures). See the *OUTPUT OPTIONS* section for more information.

LIT INFRASTRUCTURE

This section describes the **lit** testing architecture for users interested in creating a new **lit** testing implementation, or extending an existing one.

lit proper is primarily an infrastructure for discovering and running arbitrary tests, and to expose a single convenient interface to these tests. **lit** itself doesn't know how to run tests, rather this logic is defined by *test suites*.

TEST SUITES

As described in *TEST DISCOVERY*, tests are always located inside a *test suite*. Test suites serve to define the format of the tests they contain, the logic for finding those tests, and any additional information to run the tests.

lit identifies test suites as directories containing `lit.cfg` or `lit.site.cfg` files (see also *--config-prefix*). Test suites are initially discovered by recursively searching up the directory hierarchy for all the input files passed on the command line. You can use *--show-suites* to display the discovered test suites at startup.

Once a test suite is discovered, its config file is loaded. Config files themselves are Python modules which will be executed. When the config file is executed, two important global variables are predefined:

lit_config

The global **lit** configuration object (a *LitConfig* instance), which defines the builtin test formats, global configuration parameters, and other helper routines for implementing test configurations.

config

This is the config object (a *TestingConfig* instance) for the test suite, which the config file is expected to populate. The following variables are also available on the *config* object, some of which must be set by the config and others are optional or predefined:

name [*required*] The name of the test suite, for use in reports and diagnostics.

test_format [*required*] The test format object which will be used to discover and run tests in the test suite. Generally this will be a builtin test format available from the *lit.formats* module.

test_source_root The filesystem path to the test suite root. For out-of-dir builds this is the directory that will be scanned for tests.

test_exec_root For out-of-dir builds, the path to the test suite root inside the object directory. This is where tests will be run and temporary output files placed.

environment A dictionary representing the environment to use when executing tests in the suite.

suffixes For **lit** test formats which scan directories for tests, this variable is a list of suffixes to identify test files. Used by: *ShTest*.

substitutions For **lit** test formats which substitute variables into a test script, the list of substitutions to perform. Used by: *ShTest*.

unsupported Mark an unsupported directory, all tests within it will be reported as unsupported. Used by: *ShTest*.

parent The parent configuration, this is the config object for the directory containing the test suite, or None.

root The root configuration. This is the top-most **lit** configuration in the project.

pipefail Normally a test using a shell pipe fails if any of the commands on the pipe fail. If this is not desired, setting this variable to false makes the test fail only if the last command in the pipe fails.

available_features A set of features that can be used in *XFAIL*, *REQUIRES*, and *UNSUPPORTED* directives.

TEST DISCOVERY

Once test suites are located, **lit** recursively traverses the source directory (following *test_source_root*) looking for tests. When **lit** enters a sub-directory, it first checks to see if a nested test suite is defined in that directory. If so, it loads that test suite recursively, otherwise it instantiates a local test config for the directory (see [LOCAL CONFIGURATION FILES](#)).

Tests are identified by the test suite they are contained within, and the relative path inside that suite. Note that the relative path may not refer to an actual file on disk; some test formats (such as *GoogleTest*) define “virtual tests” which have a path that contains both the path to the actual test file and a subpath to identify the virtual test.

LOCAL CONFIGURATION FILES

When **lit** loads a subdirectory in a test suite, it instantiates a local test configuration by cloning the configuration for the parent directory — the root of this configuration chain will always be a test suite. Once the test configuration is cloned **lit** checks for a *lit.local.cfg* file in the subdirectory. If present, this file will be loaded and can be used to specialize the configuration for each individual directory. This facility can be used to define subdirectories of optional tests, or to change other configuration parameters — for example, to change the test format, or the suffixes which identify test files.

PRE-DEFINED SUBSTITUTIONS

lit provides various patterns that can be used with the RUN command. These are defined in *TestRunner.py*. The base set of substitutions are:

Macro	Substitution
<code>%s</code>	source path (path to the file currently being run)
<code>%S</code>	source dir (directory of the file currently being run)
<code>%p</code>	same as <code>%S</code>
<code>%{pathsep}</code>	path separator
<code>%t</code>	temporary file name unique to the test
<code>%T</code>	temporary directory unique to the test
<code>%%</code>	<code>%</code>

Other substitutions are provided that are variations on this base set and further substitution patterns can be defined by each test module. See the modules [LOCAL CONFIGURATION FILES](#).

More detailed information on substitutions can be found in the *LLVM Testing Infrastructure Guide*.

TEST RUN OUTPUT FORMAT

The **lit** output for a test run conforms to the following schema, in both short and verbose modes (although in short mode no PASS lines will be shown). This schema has been chosen to be relatively easy to reliably parse by a machine (for example in buildbot log scraping), and for other tools to generate.

Each test result is expected to appear on a line that matches:

```
<result code>: <test name> (<progress info>)
```

where `<result-code>` is a standard test result such as PASS, FAIL, XFAIL, XPASS, UNRESOLVED, or UNSUPPORTED. The performance result codes of IMPROVED and REGRESSED are also allowed.

The `<test name>` field can consist of an arbitrary string containing no newline.

The `<progress info>` field can be used to report progress information such as (1/300) or can be empty, but even when empty the parentheses are required.

Each test result may include additional (multiline) log information in the following format:

```
<log delineator> TEST '(<test name>)' <trailing delineator>
... log message ...
<log delineator>
```

where `<test name>` should be the name of a preceding reported test, `<log delineator>` is a string of “*” characters *at least* four characters long (the recommended length is 20), and `<trailing delineator>` is an arbitrary (unparsed) string.

The following is an example of a test run output which consists of four tests A, B, C, and D, and a log message for the failing test C:

```
PASS: A (1 of 4)
PASS: B (2 of 4)
FAIL: C (3 of 4)
***** TEST 'C' FAILED *****
Test 'C' failed as a result of exit code 1.
*****
PASS: D (4 of 4)
```

LIT EXAMPLE TESTS

The **lit** distribution contains several example implementations of test suites in the *ExampleTests* directory.

SEE ALSO

valgrind(1)

llvm-build - LLVM Project Build Utility

SYNOPSIS

llvm-build [*options*]

DESCRIPTION

llvm-build is a tool for working with LLVM projects that use the LLVMBuild system for describing their components.

At heart, **llvm-build** is responsible for loading, verifying, and manipulating the project's component data. The tool is primarily designed for use in implementing build systems and tools which need access to the project structure information.

OPTIONS

-h, -help

Print the builtin program help.

-source-root=PATH

If given, load the project at the given source root path. If this option is not given, the location of the project sources will be inferred from the location of the **llvm-build** script itself.

-print-tree

Print the component tree for the project.

-write-library-table

Write out the C++ fragment which defines the components, library names, and required libraries. This C++ fragment is built into `llvm-config` in order to provide clients with the list of required libraries for arbitrary component combinations.

-write-llvmbuild

Write out new *LLVMBuild.txt* files based on the loaded components. This is useful for auto-upgrading the schema of the files. **llvm-build** will try to a limited extent to preserve the comments which were written in the original source file, although at this time it only preserves block comments that precede the section names in the *LLVMBuild* files.

-write-cmake-fragment

Write out the LLVMBuild in the form of a CMake fragment, so it can easily be consumed by the CMake based build system. The exact contents and format of this file are closely tied to how LLVMBuild is integrated with CMake, see LLVM's top-level `CMakeLists.txt`.

-write-make-fragment

Write out the LLVMBuild in the form of a Makefile fragment, so it can easily be consumed by a Make based build system. The exact contents and format of this file are closely tied to how LLVMBuild is integrated with the Makefiles, see LLVM's `Makefile.rules`.

-llvmbuild-source-root=PATH

If given, expect the *LLVMBuild* files for the project to be rooted at the given path, instead of inside the source tree itself. This option is primarily designed for use in conjunction with **-write-llvmbuild** to test changes to *LLVMBuild* schema.

EXIT STATUS

llvm-build exits with 0 if operation was successful. Otherwise, it will exist with a non-zero value.

llvm-readobj - LLVM Object Reader

SYNOPSIS

```
llvm-readobj [options] [input...]
```

DESCRIPTION

The **llvm-readobj** tool displays low-level format-specific information about one or more object files. The tool and its output is primarily designed for use in FileCheck-based tests.

OPTIONS

If input is “-” or omitted, **llvm-readobj** reads from standard input. Otherwise, it will read from the specified filenames.

-help

Print a summary of command line options.

-version

Display the version of this program

-file-headers, -h

Display file headers.

-sections, -s

Display all sections.

-section-data, -sd

When used with **-sections**, display section data for each section shown.

-section-relocations, -sr

When used with **-sections**, display relocations for each section shown.

-section-symbols, -st

When used with **-sections**, display symbols for each section shown.

-relocations, -r

Display the relocation entries in the file.

-symbols, -t

Display the symbol table.

-dyn-symbols

Display the dynamic symbol table (only for ELF object files).

-unwind, -u

Display unwind information.

-expand-relocs

When used with **-relocations**, display each relocation in an expanded multi-line format.

-dynamic-table

Display the ELF .dynamic section table (only for ELF object files).

-needed-libs

Display the needed libraries (only for ELF object files).

-program-headers

Display the ELF program headers (only for ELF object files).

-elf-section-groups, -g

Display section groups (only for ELF object files).

EXIT STATUS

`llvm-readobj` returns 0.

2.7 Getting Started with the LLVM System

- *Overview*
- *Getting Started Quickly (A Summary)*
- *Requirements*
 - *Hardware*
 - *Software*
 - *Host C++ Toolchain, both Compiler and Standard Library*
 - * *Getting a Modern Host C++ Toolchain*
- *Getting Started with LLVM*
 - *Terminology and Notation*
 - *Unpacking the LLVM Archives*
 - *Checkout LLVM from Subversion*
 - *Git Mirror*
 - * *Sending patches with Git*
 - * *For developers to work with git-svn*
 - * *For developers to work with a git monorepo*
 - *Local LLVM Configuration*
 - *Compiling the LLVM Suite Source Code*
 - *Cross-Compiling LLVM*
 - *The Location of LLVM Object Files*
 - *Optional Configuration Items*
- *Directory Layout*
 - *llvm/examples*
 - *llvm/include*
 - *llvm/lib*
 - *llvm/projects*

- `llvm/test`
- `test-suite`
- `llvm/tools`
- `llvm/utils`
- *An Example Using the LLVM Tool Chain*
 - *Example with clang*
- *Common Problems*
- *Links*

2.7.1 Overview

Welcome to LLVM! In order to get started, you first need to know some basic information.

First, LLVM comes in three pieces. The first piece is the LLVM suite. This contains all of the tools, libraries, and header files needed to use LLVM. It contains an assembler, disassembler, bitcode analyzer and bitcode optimizer. It also contains basic regression tests that can be used to test the LLVM tools and the Clang front end.

The second piece is the [Clang](#) front end. This component compiles C, C++, Objective C, and Objective C++ code into LLVM bitcode. Once compiled into LLVM bitcode, a program can be manipulated with the LLVM tools from the LLVM suite.

There is a third, optional piece called Test Suite. It is a suite of programs with a testing harness that can be used to further test LLVM's functionality and performance.

2.7.2 Getting Started Quickly (A Summary)

The LLVM Getting Started documentation may be out of date. So, the [Clang Getting Started](#) page might also be a good place to start.

Here's the short story for getting up and running quickly with LLVM:

1. Read the documentation.
2. Read the documentation.
3. Remember that you were warned twice about reading the documentation.
 - In particular, the *relative paths specified are important*.
4. Checkout LLVM:
 - `cd where-you-want-llvm-to-live`
 - `svn co http://llvm.org/svn/llvm-project/llvm/trunk llvm`
5. Checkout Clang:
 - `cd where-you-want-llvm-to-live`
 - `cd llvm/tools`
 - `svn co http://llvm.org/svn/llvm-project/cfe/trunk clang`
6. Checkout LLD linker [**Optional**]:
 - `cd where-you-want-llvm-to-live`

- `cd llvm/tools`
- `svn co http://llvm.org/svn/llvm-project/lld/trunk lld`

7. Checkout Polly Loop Optimizer **[Optional]**:

- `cd where-you-want-llvm-to-live`
- `cd llvm/tools`
- `svn co http://llvm.org/svn/llvm-project/polly/trunk polly`

8. Checkout Compiler-RT (required to build the sanitizers) **[Optional]**:

- `cd where-you-want-llvm-to-live`
- `cd llvm/projects`
- `svn co http://llvm.org/svn/llvm-project/compiler-rt/trunk compiler-rt`

9. Checkout Libomp (required for OpenMP support) **[Optional]**:

- `cd where-you-want-llvm-to-live`
- `cd llvm/projects`
- `svn co http://llvm.org/svn/llvm-project/openmp/trunk openmp`

10. Checkout libcxx and libcxxabi **[Optional]**:

- `cd where-you-want-llvm-to-live`
- `cd llvm/projects`
- `svn co http://llvm.org/svn/llvm-project/libcxx/trunk libcxx`
- `svn co http://llvm.org/svn/llvm-project/libcxxabi/trunk libcxxabi`

11. Get the Test Suite Source Code **[Optional]**

- `cd where-you-want-llvm-to-live`
- `cd llvm/projects`
- `svn co http://llvm.org/svn/llvm-project/test-suite/trunk test-suite`

12. Configure and build LLVM and Clang:

Warning: Make sure you've checked out *all of* the source code before trying to configure with `cmake`. `cmake` does not pickup newly added source directories in incremental builds.

The build uses [CMake](#). LLVM requires CMake 3.4.3 to build. It is generally recommended to use a recent CMake, especially if you're generating Ninja build files. This is because the CMake project is constantly improving the quality of the generators, and the Ninja generator gets a lot of attention.

- `cd where you want to build llvm`
- `mkdir build`
- `cd build`
- `cmake -G <generator> [options] <path to llvm sources>`

Some common generators are:

- Unix Makefiles — for generating make-compatible parallel makefiles.
- Ninja — for generating [Ninja](#) build files. Most llvm developers use Ninja.
- Visual Studio — for generating Visual Studio projects and solutions.

- Xcode — for generating Xcode projects.

Some Common options:

- `-DCMAKE_INSTALL_PREFIX=directory` — Specify for *directory* the full pathname of where you want the LLVM tools and libraries to be installed (default `/usr/local`).
- `-DCMAKE_BUILD_TYPE=type` — Valid options for *type* are Debug, Release, RelWithDebInfo, and MinSizeRel. Default is Debug.
- `-DLLVM_ENABLE_ASSERTIONS=On` — Compile with assertion checks enabled (default is Yes for Debug builds, No for all other build types).
- Run your build tool of choice!
 - The default target (i.e. `make`) will build all of LLVM
 - The `check-all` target (i.e. `make check-all`) will run the regression tests to ensure everything is in working order.
 - CMake will generate build targets for each tool and library, and most LLVM sub-projects generate their own `check-<project>` target.
 - Running a serial build will be *slow*. Make sure you run a parallel build; for `make`, use `make -j`.
- For more information see [CMake](#)
- If you get an “internal compiler error (ICE)” or test failures, see [below](#).

Consult the [Getting Started with LLVM](#) section for detailed information on configuring and compiling LLVM. Go to [Directory Layout](#) to learn about the layout of the source code tree.

2.7.3 Requirements

Before you begin to use the LLVM system, review the requirements given below. This may save you some trouble by knowing ahead of time what hardware and software you will need.

Hardware

LLVM is known to work on the following host platforms:

OS	Arch	Compilers
Linux	x86 ¹	GCC, Clang
Linux	amd64	GCC, Clang
Linux	ARM ⁴	GCC, Clang
Linux	PowerPC	GCC, Clang
Solaris	V9 (Ultrasparc)	GCC
FreeBSD	x86 ¹	GCC, Clang
FreeBSD	amd64	GCC, Clang
NetBSD	x86 ¹	GCC, Clang
NetBSD	amd64	GCC, Clang
MacOS X ²	PowerPC	GCC
MacOS X	x86	GCC, Clang
Cygwin/Win32	x86 ^{1,3}	GCC
Windows	x86 ¹	Visual Studio
Windows x64	x86-64	Visual Studio

Note:

1. Code generation supported for Pentium processors and up
 2. Code generation supported for 32-bit ABI only
 3. To use LLVM modules on Win32-based system, you may configure LLVM with `-DBUILD_SHARED_LIBS=On`.
 4. MCJIT not working well pre-v7, old JIT engine not supported any more.
-

Note that Debug builds require a lot of time and disk space. An LLVM-only build will need about 1-3 GB of space. A full build of LLVM and Clang will need around 15-20 GB of disk space. The exact space requirements will vary by system. (It is so large because of all the debugging information and the fact that the libraries are statically linked into multiple tools).

If you are space-constrained, you can build only selected tools or only selected targets. The Release build requires considerably less space.

The LLVM suite *may* compile on other platforms, but it is not guaranteed to do so. If compilation is successful, the LLVM utilities should be able to assemble, disassemble, analyze, and optimize LLVM bitcode. Code generation should work as well, although the generated native code may not work on your platform.

Software

Compiling LLVM requires that you have several software packages installed. The table below lists those required packages. The Package column is the usual name for the software package that LLVM depends on. The Version column provides “known to work” versions of the package. The Notes column describes how LLVM uses the package and provides other details.

Package	Version	Notes
GNU Make	3.79, 3.79.1	Makefile/build processor
GCC	$\geq 4.8.0$	C/C++ compiler ¹
python	≥ 2.7	Automated test suite ²
zlib	$\geq 1.2.3.4$	Compression library ³

Note:

1. Only the C and C++ languages are needed so there’s no need to build the other languages for LLVM’s purposes. See *below* for specific version info.
 2. Only needed if you want to run the automated test suite in the `llvm/test` directory.
 3. Optional, adds compression / uncompression capabilities to selected LLVM tools.
-

Additionally, your compilation host is expected to have the usual plethora of Unix utilities. Specifically:

- **ar** — archive library builder
- **bzip2** — bzip2 command for distribution generation
- **bunzip2** — bunzip2 command for distribution checking
- **chmod** — change permissions on a file
- **cat** — output concatenation utility

- **cp** — copy files
- **date** — print the current date/time
- **echo** — print to standard output
- **egrep** — extended regular expression search utility
- **find** — find files/dirs in a file system
- **grep** — regular expression search utility
- **gzip** — gzip command for distribution generation
- **gunzip** — gunzip command for distribution checking
- **install** — install directories/files
- **mkdir** — create a directory
- **mv** — move (rename) files
- **ranlib** — symbol table builder for archive libraries
- **rm** — remove (delete) files and directories
- **sed** — stream editor for transforming output
- **sh** — Bourne shell for make build scripts
- **tar** — tape archive for distribution generation
- **test** — test things in file system
- **unzip** — unzip command for distribution checking
- **zip** — zip command for distribution generation

Host C++ Toolchain, both Compiler and Standard Library

LLVM is very demanding of the host C++ compiler, and as such tends to expose bugs in the compiler. We are also planning to follow improvements and developments in the C++ language and library reasonably closely. As such, we require a modern host C++ toolchain, both compiler and standard library, in order to build LLVM.

For the most popular host toolchains we check for specific minimum versions in our build systems:

- Clang 3.1
- GCC 4.8
- Visual Studio 2015 (Update 3)

Anything older than these toolchains *may* work, but will require forcing the build system with a special option and is not really a supported host platform. Also note that older versions of these compilers have often crashed or miscompiled LLVM.

For less widely used host toolchains such as ICC or xIC, be aware that a very recent version may be required to support all of the C++ features used in LLVM.

We track certain versions of software that are *known* to fail when used as part of the host toolchain. These even include linkers at times.

GNU ld 2.16.X. Some 2.16.X versions of the ld linker will produce very long warning messages complaining that some “.gnu.linkonce.t.*” symbol was defined in a discarded section. You can safely ignore these messages as they are erroneous and the linkage is correct. These messages disappear using ld 2.17.

GNU binutils 2.17: Binutils 2.17 contains a [bug](#) which causes huge link times (minutes instead of seconds) when building LLVM. We recommend upgrading to a newer version (2.17.50.0.4 or later).

GNU Binutils 2.19.1 Gold: This version of Gold contained a [bug](#) which causes intermittent failures when building LLVM with position independent code. The symptom is an error about cyclic dependencies. We recommend upgrading to a newer version of Gold.

Getting a Modern Host C++ Toolchain

This section mostly applies to Linux and older BSDs. On Mac OS X, you should have a sufficiently modern Xcode, or you will likely need to upgrade until you do. Windows does not have a “system compiler”, so you must install either Visual Studio 2015 or a recent version of mingw64. FreeBSD 10.0 and newer have a modern Clang as the system compiler.

However, some Linux distributions and some other or older BSDs sometimes have extremely old versions of GCC. These steps attempt to help you upgrade your compiler even on such a system. However, if at all possible, we encourage you to use a recent version of a distribution with a modern system compiler that meets these requirements. Note that it is tempting to install a prior version of Clang and libc++ to be the host compiler, however libc++ was not well tested or set up to build on Linux until relatively recently. As a consequence, this guide suggests just using libstdc++ and a modern GCC as the initial host in a bootstrap, and then using Clang (and potentially libc++).

The first step is to get a recent GCC toolchain installed. The most common distribution on which users have struggled with the version requirements is Ubuntu Precise, 12.04 LTS. For this distribution, one easy option is to install the [toolchain testing PPA](#) and use it to install a modern GCC. There is a really nice discussion of this on the [ask ubuntu stack exchange](#). However, not all users can use PPAs and there are many other distributions, so it may be necessary (or just useful, if you’re here you *are* doing compiler development after all) to build and install GCC from source. It is also quite easy to do these days.

Easy steps for installing GCC 4.8.2:

```
% wget https://ftp.gnu.org/gnu/gcc/gcc-4.8.2/gcc-4.8.2.tar.bz2
% wget https://ftp.gnu.org/gnu/gcc/gcc-4.8.2/gcc-4.8.2.tar.bz2.sig
% wget https://ftp.gnu.org/gnu/gnu-keyring.gpg
% signature_invalid=`gpg --verify --no-default-keyring --keyring ./gnu-keyring.gpg_
↳gcc-4.8.2.tar.bz2.sig`
% if [ $signature_invalid ]; then echo "Invalid signature" ; exit 1 ; fi
% tar -xvjf gcc-4.8.2.tar.bz2
% cd gcc-4.8.2
% ./contrib/download_prerequisites
% cd ..
% mkdir gcc-4.8.2-build
% cd gcc-4.8.2-build
% $PWD/./gcc-4.8.2/configure --prefix=$HOME/toolchains --enable-languages=c,c++
% make -j$(nproc)
% make install
```

For more details, check out the excellent [GCC wiki entry](#), where I got most of this information from.

Once you have a GCC toolchain, configure your build of LLVM to use the new toolchain for your host compiler and C++ standard library. Because the new version of libstdc++ is not on the system library search path, you need to pass extra linker flags so that it can be found at link time (-L) and at runtime (-rpath). If you are using CMake, this invocation should produce working binaries:

```
% mkdir build
% cd build
% CC=$HOME/toolchains/bin/gcc CXX=$HOME/toolchains/bin/g++ \
  cmake .. -DCMAKE_CXX_LINK_FLAGS="-Wl,-rpath,$HOME/toolchains/lib64 -L$HOME/
↳toolchains/lib64"
```

If you fail to set `rpath`, most LLVM binaries will fail on startup with a message from the loader similar to `libstdc++.so.6: version `GLIBCXX_3.4.20' not found`. This means you need to tweak the `-rpath` linker flag.

When you build Clang, you will need to give *it* access to modern C++11 standard library in order to use it as your new host in part of a bootstrap. There are two easy ways to do this, either build (and install) `libc++` along with Clang and then use it with the `-stdlib=libc++` compile and link flag, or install Clang into the same prefix (`$HOME/toolchains` above) as GCC. Clang will look within its own prefix for `libstdc++` and use it if found. You can also add an explicit prefix for Clang to look in for a GCC toolchain with the `--gcc-toolchain=/opt/my/gcc/prefix` flag, passing it to both compile and link commands when using your just-built-Clang to bootstrap.

2.7.4 Getting Started with LLVM

The remainder of this guide is meant to get you up and running with LLVM and to give you some basic information about the LLVM environment.

The later sections of this guide describe the *general layout* of the LLVM source tree, a *simple example* using the LLVM tool chain, and *links* to find more information about LLVM or to get help via e-mail.

Terminology and Notation

Throughout this manual, the following names are used to denote paths specific to the local system and working environment. *These are not environment variables you need to set but just strings used in the rest of this document below*. In any of the examples below, simply replace each of these names with the appropriate pathname on your local system. All these paths are absolute:

`SRC_ROOT`

This is the top level directory of the LLVM source tree.

`OBJ_ROOT`

This is the top level directory of the LLVM object tree (i.e. the tree where object files and compiled programs will be placed. It can be the same as `SRC_ROOT`).

Unpacking the LLVM Archives

If you have the LLVM distribution, you will need to unpack it before you can begin to compile it. LLVM is distributed as a set of two files: the LLVM suite and the LLVM GCC front end compiled for your platform. There is an additional test suite that is optional. Each file is a TAR archive that is compressed with the `gzip` program.

The files are as follows, with `x.y` marking the version number:

`llvm-x.y.tar.gz`

Source release for the LLVM libraries and tools.

`llvm-test-x.y.tar.gz`

Source release for the LLVM test-suite.

Checkout LLVM from Subversion

If you have access to our Subversion repository, you can get a fresh copy of the entire source code. All you need to do is check it out from Subversion as follows:

- `cd where-you-want-llvm-to-live`
- Read-Only: `svn co http://llvm.org/svn/llvm-project/llvm/trunk llvm`
- Read-Write: `svn co https://user@llvm.org/svn/llvm-project/llvm/trunk llvm`

This will create an ‘llvm’ directory in the current directory and fully populate it with the LLVM source code, Makefiles, test directories, and local copies of documentation files.

If you want to get a specific release (as opposed to the most recent revision), you can checkout it from the ‘tags’ directory (instead of ‘trunk’). The following releases are located in the following subdirectories of the ‘tags’ directory:

- Release 3.4: **RELEASE_34/final**
- Release 3.3: **RELEASE_33/final**
- Release 3.2: **RELEASE_32/final**
- Release 3.1: **RELEASE_31/final**
- Release 3.0: **RELEASE_30/final**
- Release 2.9: **RELEASE_29/final**
- Release 2.8: **RELEASE_28**
- Release 2.7: **RELEASE_27**
- Release 2.6: **RELEASE_26**
- Release 2.5: **RELEASE_25**
- Release 2.4: **RELEASE_24**
- Release 2.3: **RELEASE_23**
- Release 2.2: **RELEASE_22**
- Release 2.1: **RELEASE_21**
- Release 2.0: **RELEASE_20**
- Release 1.9: **RELEASE_19**
- Release 1.8: **RELEASE_18**
- Release 1.7: **RELEASE_17**
- Release 1.6: **RELEASE_16**
- Release 1.5: **RELEASE_15**
- Release 1.4: **RELEASE_14**
- Release 1.3: **RELEASE_13**
- Release 1.2: **RELEASE_12**
- Release 1.1: **RELEASE_11**
- Release 1.0: **RELEASE_1**

If you would like to get the LLVM test suite (a separate package as of 1.4), you get it from the Subversion repository:

```
% cd llvm/projects
% svn co http://llvm.org/svn/llvm-project/test-suite/trunk test-suite
```

By placing it in the `llvm/projects`, it will be automatically configured by the LLVM cmake configuration.

Git Mirror

Git mirrors are available for a number of LLVM subprojects. These mirrors sync automatically with each Subversion commit and contain all necessary git-svn marks (so, you can recreate git-svn metadata locally). Note that right now mirrors reflect only `trunk` for each project. You can do the read-only Git clone of LLVM via:

```
% git clone http://llvm.org/git/llvm.git
```

If you want to check out clang too, run:

```
% cd llvm/tools
% git clone http://llvm.org/git/clang.git
```

If you want to check out compiler-rt (required to build the sanitizers), run:

```
% cd llvm/projects
% git clone http://llvm.org/git/compiler-rt.git
```

If you want to check out libomp (required for OpenMP support), run:

```
% cd llvm/projects
% git clone http://llvm.org/git/openmp.git
```

If you want to check out libcxx and libcxxabi (optional), run:

```
% cd llvm/projects
% git clone http://llvm.org/git/libcxx.git
% git clone http://llvm.org/git/libcxxabi.git
```

If you want to check out the Test Suite Source Code (optional), run:

```
% cd llvm/projects
% git clone http://llvm.org/git/test-suite.git
```

Since the upstream repository is in Subversion, you should use `git pull --rebase` instead of `git pull` to avoid generating a non-linear history in your clone. To configure `git pull` to pass `--rebase` by default on the master branch, run the following command:

```
% git config branch.master.rebase true
```

Sending patches with Git

Please read [Developer Policy](#), too.

Assume `master` points the upstream and `mybranch` points your working branch, and `mybranch` is rebased onto `master`. At first you may check sanity of whitespaces:

```
% git diff --check master..mybranch
```

The easiest way to generate a patch is as below:

```
% git diff master..mybranch > /path/to/mybranch.diff
```

It is a little different from svn-generated diff. git-diff-generated diff has prefixes like `a/` and `b/`. Don't worry, most developers might know it could be accepted with `patch -p1 -N`.

But you may generate patchset with `git-format-patch`. It generates by-each-commit patchset. To generate patch files to attach to your article:

```
% git format-patch --no-attach master..mybranch -o /path/to/your/patchset
```

If you would like to send patches directly, you may use `git-send-email` or `git-imap-send`. Here is an example to generate the patchset in Gmail's [Drafts].

```
% git format-patch --attach master..mybranch --stdout | git imap-send
```

Then, your `.git/config` should have [imap] sections.

```
[imap]
  host = imaps://imap.gmail.com
  user = your.gmail.account@gmail.com
  pass = himitsu!
  port = 993
  sslverify = false
; in English
  folder = "[Gmail]/Drafts"
; example for Japanese, "Modified UTF-7" encoded.
  folder = "[Gmail]/&Tgtm+DBN-"
; example for Traditional Chinese
  folder = "[Gmail]/&q016Pw-"
```

For developers to work with git-svn

To set up clone from which you can submit code using `git-svn`, run:

```
% git clone http://llvm.org/git/llvm.git
% cd llvm
% git svn init https://llvm.org/svn/llvm-project/llvm/trunk --username=<username>
% git config svn-remote.svn.fetch :refs/remotes/origin/master
% git svn rebase -l  # -l avoids fetching ahead of the git mirror.

# If you have clang too:
% cd tools
% git clone http://llvm.org/git/clang.git
% cd clang
% git svn init https://llvm.org/svn/llvm-project/cfe/trunk --username=<username>
% git config svn-remote.svn.fetch :refs/remotes/origin/master
% git svn rebase -l
```

Likewise for `compiler-rt`, `libomp` and `test-suite`.

To update this clone without generating `git-svn` tags that conflict with the upstream Git repo, run:

```
% git fetch && (cd tools/clang && git fetch)  # Get matching revisions of both trees.
% git checkout master
% git svn rebase -l
% (cd tools/clang &&
  git checkout master &&
  git svn rebase -l)
```

Likewise for `compiler-rt`, `libomp` and `test-suite`.

This leaves your working directories on their master branches, so you'll need to `checkout` each working branch individually and `rebase` it on top of its parent branch.

For those who wish to be able to update an llvm repo/revert patches easily using git-svn, please look in the directory for the scripts `git-svnup` and `git-svnrevert`.

To perform the aforementioned update steps go into your source directory and just type `git-svnup` or `git svnup` and everything will just work.

If one wishes to revert a commit with git-svn, but do not want the git hash to escape into the commit message, one can use the script `git-svnrevert` or `git svnrevert` which will take in the git hash for the commit you want to revert, look up the appropriate svn revision, and output a message where all references to the git hash have been replaced with the svn revision.

To commit back changes via git-svn, use `git svn dcommit`:

```
% git svn dcommit
```

Note that git-svn will create one SVN commit for each Git commit you have pending, so squash and edit each commit before executing `dcommit` to make sure they all conform to the coding standards and the developers' policy.

On success, `dcommit` will rebase against the HEAD of SVN, so to avoid conflict, please make sure your current branch is up-to-date (via `fetch/rebase`) before proceeding.

The git-svn metadata can get out of sync after you mess around with branches and `dcommit`. When that happens, `git svn dcommit` stops working, complaining about files with uncommitted changes. The fix is to rebuild the metadata:

```
% rm -rf .git/svn
% git svn rebase -l
```

Please, refer to the Git-SVN manual (`man git-svn`) for more information.

For developers to work with a git monorepo

Note: This set-up is using an unofficial mirror hosted on GitHub, use with caution.

To set up a clone of all the llvm projects using a unified repository:

```
% export TOP_LEVEL_DIR=`pwd`
% git clone https://github.com/llvm-project/llvm-project-20170507/ llvm-project
% cd llvm-project
% git config branch.master.rebase true
```

You can configure various build directory from this clone, starting with a build of LLVM alone:

```
% cd $TOP_LEVEL_DIR
% mkdir llvm-build && cd llvm-build
% cmake -GNinja ../llvm-project/llvm
```

Or lldb:

```
% cd $TOP_LEVEL_DIR
% mkdir lldb-build && cd lldb-build
% cmake -GNinja ../llvm-project/llvm -DLLVM_ENABLE_PROJECTS=lldb
```

Or a combination of multiple projects:

```
% cd $TOP_LEVEL_DIR
% mkdir clang-build && cd clang-build
% cmake -GNinja ../llvm-project/llvm -DLLVM_ENABLE_PROJECTS="clang;libcxx;libcxxabi"
```

A helper script is provided in `llvm/utils/git-svn/git-llvm`. After you add it to your path, you can push committed changes upstream with `git llvm push`.

```
% export PATH=$PATH:$TOP_LEVEL_DIR/llvm-project/llvm/utils/git-svn/
% git llvm push
```

While this is using SVN under the hood, it does not require any interaction from you with `git-svn`. After a few minutes, `git pull` should get back the changes as they were committed. Note that a current limitation is that `git` does not directly record file rename, and thus it is propagated to SVN as a combination of delete-add instead of a file rename.

The SVN revision of each monorepo commit can be found in the commit notes. `git` does not fetch notes by default. The following commands will fetch the notes and configure `git` to fetch future notes. Use `git notes show $commit` to look up the SVN revision of a `git` commit. The notes show up `git log`, and searching the log is currently the recommended way to look up the `git` commit for a given SVN revision.

```
% git config --add remote.origin.fetch +refs/notes/commits:refs/notes/commits
% git fetch
```

If you are using *arc* to interact with Phabricator, you need to manually put it at the root of the checkout:

```
% cd $TOP_LEVEL_DIR
% cp llvm/.arcconfig ./
% mkdir -p .git/info/
% echo .arcconfig >> .git/info/exclude
```

Local LLVM Configuration

Once checked out from the Subversion repository, the LLVM suite source code must be configured before being built. This process uses CMake. Unlike the normal `configure` script, CMake generates the build files in whatever format you request as well as various `*.inc` files, and `llvm/include/Config/config.h`.

Variables are passed to `cmake` on the command line using the format `-D<variable name>=<value>`. The following variables are some common options used by people developing LLVM.

Variable	Purpose
CMAKE_C_COMPILER	Specifies which C compiler to use. By default, this will be /usr/bin/cc.
CMAKE_CXX_COMPILER	Specifies which C++ compiler to use. By default, this will be /usr/bin/c++.
CMAKE_BUILD_TYPE	Specifies what type of build you are trying to generate files for. Valid options are Debug, Release, RelWithDebInfo, and MinSizeRel. Default is Debug.
CMAKE_INSTALL_PREFIX	Specifies the PREFIX directory to target when running the install action of the build files.
LLVM_TARGETS_TO_BUILD	Unlimited list controlling which targets will be built and linked into llc. This is equivalent to the --enable-targets option in the configure script. The default list is defined as LLVM_ALL_TARGETS, and can be set to include out-of-tree targets. The default value includes: AArch64, AMDGPU, ARM, BPF, Hexagon, Mips, MSP430, NVPTX, PowerPC, Sparc, SystemZ, X86, XCore.
LLVM_ENABLE_DOXYGEN	Enables documentation from the source code. This is disabled by default because it is slow and generates a lot of output.
LLVM_ENABLE_SPHINX	Enables Sphinx documentation from the source code. This is disabled by default because it is slow and generates a lot of output. Sphinx version 1.5 or later recommended.
LLVM_BUILD_LIBRARY	Specifies the LLVM library to build. This library contains a default set of LLVM components that can be overridden with LLVM_DYLIB_COMPONENTS. The default contains most of LLVM and is defined in tools/llvm-shlib/CMakeLists.txt.
LLVM_OPTIMIZED_DEBUGGER	Enables the use of the LLVM debugger that gets used during the LLVM build. This can dramatically speed up debug builds.

To configure LLVM, follow these steps:

1. Change directory into the object root directory:

```
% cd OBJ_ROOT
```

2. Run the cmake:

```
% cmake -G "Unix Makefiles" -DCMAKE_INSTALL_PREFIX=prefix=/install/path
[other options] SRC_ROOT
```

Compiling the LLVM Suite Source Code

Unlike with autotools, with CMake your build type is defined at configuration. If you want to change your build type, you can re-run cmake with the following invocation:

```
% cmake -G "Unix Makefiles" -DCMAKE_BUILD_TYPE=type SRC_ROOT
```

Between runs, CMake preserves the values set for all options. CMake has the following build types defined:

Debug

These builds are the default. The build system will compile the tools and libraries unoptimized, with debugging information, and asserts enabled.

Release

For these builds, the build system will compile the tools and libraries with optimizations enabled and not generate debug info. CMakes default optimization level is -O3. This can be configured by setting the CMAKE_CXX_FLAGS_RELEASE variable on the CMake command line.

RelWithDebInfo

These builds are useful when debugging. They generate optimized binaries with debug information. CMake's default optimization level is `-O2`. This can be configured by setting the `CMAKE_CXX_FLAGS_RELWITHDEBINFO` variable on the CMake command line.

Once you have LLVM configured, you can build it by entering the `OBJ_ROOT` directory and issuing the following command:

```
% make
```

If the build fails, please [check here](#) to see if you are using a version of GCC that is known not to compile LLVM.

If you have multiple processors in your machine, you may wish to use some of the parallel build options provided by GNU Make. For example, you could use the command:

```
% make -j2
```

There are several special targets which are useful when working with the LLVM source code:

`make clean`

Removes all files generated by the build. This includes object files, generated C/C++ files, libraries, and executables.

`make install`

Installs LLVM header files, libraries, tools, and documentation in a hierarchy under `$PREFIX`, specified with `CMAKE_INSTALL_PREFIX`, which defaults to `/usr/local`.

`make docs-llvm-html`

If configured with `-DLLVM_ENABLE_SPHINX=On`, this will generate a directory at `OBJ_ROOT/docs/html` which contains the HTML formatted documentation.

Cross-Compiling LLVM

It is possible to cross-compile LLVM itself. That is, you can create LLVM executables and libraries to be hosted on a platform different from the platform where they are built (a Canadian Cross build). To generate build files for cross-compiling CMake provides a variable `CMAKE_TOOLCHAIN_FILE` which can define compiler flags and variables used during the CMake test operations.

The result of such a build is executables that are not runnable on the build host but can be executed on the target. As an example the following CMake invocation can generate build files targeting iOS. This will work on Mac OS X with the latest Xcode:

```
% cmake -G "Ninja" -DCMAKE_OSX_ARCHITECTURES="armv7;armv7s;arm64"
-D CMAKE_TOOLCHAIN_FILE=<PATH_TO_LLVM>/cmake/platforms/iOS.cmake
-D CMAKE_BUILD_TYPE=Release -DLLVM_BUILD_RUNTIME=Off -DLLVM_INCLUDE_TESTS=Off
-D LLVM_INCLUDE_EXAMPLES=Off -DLLVM_ENABLE_BACKTRACES=Off [options]
<PATH_TO_LLVM>
```

Note: There are some additional flags that need to be passed when building for iOS due to limitations in the iOS SDK.

Check [How To Cross-Compile Clang/LLVM using Clang/LLVM](#) and Clang docs on how to cross-compile in general for more information about cross-compiling.

The Location of LLVM Object Files

The LLVM build system is capable of sharing a single LLVM source tree among several LLVM builds. Hence, it is possible to build LLVM for several different platforms or configurations using the same source tree.

- Change directory to where the LLVM object files should live:

```
% cd OBJ_ROOT
```

- Run cmake:

```
% cmake -G "Unix Makefiles" SRC_ROOT
```

The LLVM build will create a structure underneath *OBJ_ROOT* that matches the LLVM source tree. At each level where source files are present in the source tree there will be a corresponding CMakeFiles directory in the *OBJ_ROOT*. Underneath that directory there is another directory with a name ending in *.dir* under which you'll find object files for each source.

For example:

```
% cd llvm_build_dir
% find lib/Support/ -name APFloat*
lib/Support/CMakeFiles/LLVMSupport.dir/APFloat.cpp.o
```

Optional Configuration Items

If you're running on a Linux system that supports the `binfmt_misc` module, and you have root access on the system, you can set your system up to execute LLVM bitcode files directly. To do this, use commands like this (the first command may not be required if you are already using the module):

```
% mount -t binfmt_misc none /proc/sys/fs/binfmt_misc
% echo ':llvm:M::BC::/path/to/lli:' > /proc/sys/fs/binfmt_misc/register
% chmod u+x hello.bc (if needed)
% ./hello.bc
```

This allows you to execute LLVM bitcode files directly. On Debian, you can also use this command instead of the 'echo' command above:

```
% sudo update-binfmts --install llvm /path/to/lli --magic 'BC'
```

2.7.5 Directory Layout

One useful source of information about the LLVM source base is the LLVM `doxygen` documentation available at <http://llvm.org/doxygen/>. The following is a brief introduction to code layout:

llvm/examples

Simple examples using the LLVM IR and JIT.

llvm/include

Public header files exported from the LLVM library. The three main subdirectories:

```
llvm/include/llvm
```

All LLVM-specific header files, and subdirectories for different portions of LLVM: Analysis, CodeGen, Target, Transforms, etc...

```
llvm/include/llvm/Support
```

Generic support libraries provided with LLVM but not necessarily specific to LLVM. For example, some C++ STL utilities and a Command Line option processing library store header files here.

`llvm/include/llvm/Config`

Header files configured by the `configure` script. They wrap “standard” UNIX and C header files. Source code can include these header files which automatically take care of the conditional `#includes` that the `configure` script generates.

llvm/lib

Most source files are here. By putting code in libraries, LLVM makes it easy to share code among the *tools*.

`llvm/lib/IR/`

Core LLVM source files that implement core classes like `Instruction` and `BasicBlock`.

`llvm/lib/AsmParser/`

Source code for the LLVM assembly language parser library.

`llvm/lib/Bitcode/`

Code for reading and writing bitcode.

`llvm/lib/Analysis/`

A variety of program analyses, such as Call Graphs, Induction Variables, Natural Loop Identification, etc.

`llvm/lib/Transforms/`

IR-to-IR program transformations, such as Aggressive Dead Code Elimination, Sparse Conditional Constant Propagation, Inlining, Loop Invariant Code Motion, Dead Global Elimination, and many others.

`llvm/lib/Target/`

Files describing target architectures for code generation. For example, `llvm/lib/Target/X86` holds the X86 machine description.

`llvm/lib/CodeGen/`

The major parts of the code generator: Instruction Selector, Instruction Scheduling, and Register Allocation.

`llvm/lib/MC/`

(FIXME: T.B.D.)?

`llvm/lib/ExecutionEngine/`

Libraries for directly executing bitcode at runtime in interpreted and JIT-compiled scenarios.

`llvm/lib/Support/`

Source code that corresponding to the header files in `llvm/include/ADT/` and `llvm/include/Support/`.

llvm/projects

Projects not strictly part of LLVM but shipped with LLVM. This is also the directory for creating your own LLVM-based projects which leverage the LLVM build system.

llvm/test

Feature and regression tests and other sanity checks on LLVM infrastructure. These are intended to run quickly and cover a lot of territory without being exhaustive.

test-suite

A comprehensive correctness, performance, and benchmarking test suite for LLVM. Comes in a separate Subversion module because not every LLVM user is interested in such a comprehensive suite. For details see the [Testing Guide](#) document.

llvm/tools

Executables built out of the libraries above, which form the main part of the user interface. You can always get help for a tool by typing `tool_name -help`. The following is a brief introduction to the most important tools. More detailed information is in the [Command Guide](#).

bugpoint

`bugpoint` is used to debug optimization passes or code generation backends by narrowing down the given test case to the minimum number of passes and/or instructions that still cause a problem, whether it is a crash or miscompilation. See [HowToSubmitABug.html](#) for more information on using `bugpoint`.

llvm-ar

The archiver produces an archive containing the given LLVM bitcode files, optionally with an index for faster lookup.

llvm-as

The assembler transforms the human readable LLVM assembly to LLVM bitcode.

llvm-dis

The disassembler transforms the LLVM bitcode to human readable LLVM assembly.

llvm-link

`llvm-link`, not surprisingly, links multiple LLVM modules into a single program.

lli

`lli` is the LLVM interpreter, which can directly execute LLVM bitcode (although very slowly...). For architectures that support it (currently x86, Sparc, and PowerPC), by default, `lli` will function as a Just-In-Time compiler (if the functionality was compiled in), and will execute the code *much* faster than the interpreter.

llc

`llc` is the LLVM backend compiler, which translates LLVM bitcode to a native code assembly file.

opt

`opt` reads LLVM bitcode, applies a series of LLVM to LLVM transformations (which are specified on the command line), and outputs the resultant bitcode. '`opt -help`' is a good way to get a list of the program transformations available in LLVM.

`opt` can also run a specific analysis on an input LLVM bitcode file and print the results. Primarily useful for debugging analyses, or familiarizing yourself with what an analysis does.

llvm/utils

Utilities for working with LLVM source code; some are part of the build process because they are code generators for parts of the infrastructure.

codegen-diff

codegen-diff finds differences between code that LLC generates and code that LLI generates. This is useful if you are debugging one of them, assuming that the other generates correct output. For the full user manual, run ``perldoc codegen-diff'``.

emacs/

Emacs and XEmacs syntax highlighting for LLVM assembly files and TableGen description files. See the README for information on using them.

getsrsrcs.sh

Finds and outputs all non-generated source files, useful if one wishes to do a lot of development across directories and does not want to find each file. One way to use it is to run, for example: `xemacs `utils/getsources.sh`` from the top of the LLVM source tree.

llvmgrep

Performs an `egrep -H -n` on each source file in LLVM and passes to it a regular expression provided on llvmgrep's command line. This is an efficient way of searching the source base for a particular regular expression.

makellvm

Compiles all files in the current directory, then compiles and links the tool that is the first argument. For example, assuming you are in `llvm/lib/Target/Sparc`, if `makellvm` is in your path, running `makellvm llc` will make a build of the current directory, switch to directory `llvm/tools/llc` and build it, causing a re-linking of LLC.

TableGen/

Contains the tool used to generate register descriptions, instruction set descriptions, and even assemblers from common TableGen description files.

vim/

vim syntax-highlighting for LLVM assembly files and TableGen description files. See the README for how to use them.

2.7.6 An Example Using the LLVM Tool Chain

This section gives an example of using LLVM with the Clang front end.

Example with clang

1. First, create a simple C file, name it 'hello.c':

```
#include <stdio.h>

int main() {
    printf("hello world\n");
    return 0;
}
```

- Next, compile the C file into a native executable:

```
% clang hello.c -o hello
```

Note: Clang works just like GCC by default. The standard `-S` and `-c` arguments work as usual (producing a native `.s` or `.o` file, respectively).

- Next, compile the C file into an LLVM bitcode file:

```
% clang -O3 -emit-llvm hello.c -c -o hello.bc
```

The `-emit-llvm` option can be used with the `-S` or `-c` options to emit an LLVM `.ll` or `.bc` file (respectively) for the code. This allows you to use the [standard LLVM tools](#) on the bitcode file.

- Run the program in both forms. To run the program, use:

```
% ./hello
```

and

```
% lli hello.bc
```

The second examples shows how to invoke the LLVM JIT, *lli*.

- Use the `llvm-dis` utility to take a look at the LLVM assembly code:

```
% llvm-dis < hello.bc | less
```

- Compile the program to native assembly using the LLC code generator:

```
% llc hello.bc -o hello.s
```

- Assemble the native assembly language file into a program:

```
% /opt/SUNWsprow/bin/cc -xarch=v9 hello.s -o hello.native # On Solaris
% gcc hello.s -o hello.native # On others
```

- Execute the native code program:

```
% ./hello.native
```

Note that using clang to compile directly to native code (i.e. when the `-emit-llvm` option is not present) does steps 6/7/8 for you.

2.7.7 Common Problems

If you are having problems building or using LLVM, or if you have any other general questions about LLVM, please consult the [Frequently Asked Questions](#) page.

2.7.8 Links

This document is just an **introduction** on how to use LLVM to do some simple things... there are many more interesting and complicated things that you can do that aren't documented here (but we'll gladly accept a patch if you want to write something up!). For more information about LLVM, check out:

- [LLVM Homepage](#)
- [LLVM Doxygen Tree](#)
- [Starting a Project that Uses LLVM](#)

2.8 Getting Started with the LLVM System using Microsoft Visual Studio

- *Overview*
- *Requirements*
 - *Hardware*
 - *Software*
- *Getting Started*
- *An Example Using the LLVM Tool Chain*
- *Common Problems*
- *Links*

2.8.1 Overview

Welcome to LLVM on Windows! This document only covers LLVM on Windows using Visual Studio, not mingw or cygwin. In order to get started, you first need to know some basic information.

There are many different projects that compose LLVM. The first piece is the LLVM suite. This contains all of the tools, libraries, and header files needed to use LLVM. It contains an assembler, disassembler, bitcode analyzer and bitcode optimizer. It also contains basic regression tests that can be used to test the LLVM tools and the Clang front end.

The second piece is the [Clang](#) front end. This component compiles C, C++, Objective C, and Objective C++ code into LLVM bitcode. Clang typically uses LLVM libraries to optimize the bitcode and emit machine code. LLVM fully supports the COFF object file format, which is compatible with all other existing Windows toolchains.

The last major part of LLVM, the execution Test Suite, does not run on Windows, and this document does not discuss it.

Additional information about the LLVM directory structure and tool chain can be found on the main [Getting Started with the LLVM System](#) page.

2.8.2 Requirements

Before you begin to use the LLVM system, review the requirements given below. This may save you some trouble by knowing ahead of time what hardware and software you will need.

Hardware

Any system that can adequately run Visual Studio 2015 is fine. The LLVM source tree and object files, libraries and executables will consume approximately 3GB.

Software

You will need Visual Studio 2015 or higher, with the latest Update installed.

You will also need the [CMake](#) build system since it generates the project files you will use to build with.

If you would like to run the LLVM tests you will need [Python](#). Version 2.7 and newer are known to work. You will need [GnuWin32](#) tools, too.

Do not install the LLVM directory tree into a path containing spaces (e.g. `C:\Documents and Settings\...`) as the configure step will fail.

2.8.3 Getting Started

Here's the short story for getting up and running quickly with LLVM:

1. Read the documentation.
2. Seriously, read the documentation.
3. Remember that you were warned twice about reading the documentation.
4. Get the Source Code
 - With the distributed files:
 - (a) `cd <where-you-want-llvm-to-live>`
 - (b) `gunzip --stdout llvm-VERSION.tar.gz | tar -xvf -` (*or use WinZip*)
 - (c) `cd llvm`
 - With anonymous Subversion access:
 - (a) `cd <where-you-want-llvm-to-live>`
 - (b) `svn co http://llvm.org/svn/llvm-project/llvm/trunk llvm`
 - (c) `cd llvm`
5. Use [CMake](#) to generate up-to-date project files:
 - Once CMake is installed then the simplest way is to just start the CMake GUI, select the directory where you have LLVM extracted to, and the default options should all be fine. One option you may really want to change, regardless of anything else, might be the `CMAKE_INSTALL_PREFIX` setting to select a directory to `INSTALL` to once compiling is complete, although installation is not mandatory for using LLVM. Another important option is `LLVM_TARGETS_TO_BUILD`, which controls the LLVM target architectures that are included on the build.
 - If CMake complains that it cannot find the compiler, make sure that you have the Visual Studio C++ Tools installed, not just Visual Studio itself (trying to create a C++ project in Visual Studio will generally download the C++ tools if they haven't already been).
 - See the [LLVM CMake guide](#) for detailed information about how to configure the LLVM build.
 - CMake generates project files for all build types. To select a specific build type, use the Configuration manager from the VS IDE or the `/property:Configuration` command line option when using MSBuild.
 - By default, the Visual Studio project files generated by CMake use the 32-bit toolset. If you are developing on a 64-bit version of Windows and want to use the 64-bit toolset, pass the `-T:Host=x64` flag when generating the Visual Studio solution. This requires CMake 3.8.0 or later.
6. Start Visual Studio

- In the directory you created the project files will have an `llvm.sln` file, just double-click on that to open Visual Studio.

7. Build the LLVM Suite:

- The projects may still be built individually, but to build them all do not just select all of them in batch build (as some are meant as configuration projects), but rather select and build just the `ALL_BUILD` project to build everything, or the `INSTALL` project, which first builds the `ALL_BUILD` project, then installs the LLVM headers, libs, and other useful things to the directory set by the `CMAKE_INSTALL_PREFIX` setting when you first configured CMake.
- The Fibonacci project is a sample program that uses the JIT. Modify the project's debugging properties to provide a numeric command line argument or run it from the command line. The program will print the corresponding fibonacci value.

8. Test LLVM in Visual Studio:

- If `%PATH%` does not contain GnuWin32, you may specify `LLVM_LIT_TOOLS_DIR` on CMake for the path to GnuWin32.
- You can run LLVM tests by merely building the project "check". The test results will be shown in the VS output window.

9. Test LLVM on the command line:

- The LLVM tests can be run by changing directory to the `llvm` source directory and running:

```
C:\..\llvm> python ..\build\bin\llvm-lit --param build_config=Win32 --param_
↪ build_mode=Debug --param llvm_site_config=../build/test/lit.site.cfg test
```

This example assumes that Python is in your PATH variable, you have built a Win32 Debug version of `llvm` with a standard out of line build. You should not see any unexpected failures, but will see many unsupported tests and expected failures.

A specific test or test directory can be run with:

```
C:\..\llvm> python ..\build\bin\llvm-lit --param build_config=Win32 --param_
↪ build_mode=Debug --param llvm_site_config=../build/test/lit.site.cfg test/
↪ path/to/test
```

2.8.4 An Example Using the LLVM Tool Chain

1. First, create a simple C file, name it 'hello.c':

```
#include <stdio.h>
int main() {
    printf("hello world\n");
    return 0;
}
```

2. Next, compile the C file into an LLVM bytecode file:

```
C:\..\> clang -c hello.c -emit-llvm -o hello.bc
```

This will create the result file `hello.bc` which is the LLVM bytecode that corresponds the compiled program and the library facilities that it required. You can execute this file directly using `lli` tool, compile it to native assembly with the `llc`, optimize or analyze it further with the `opt` tool, etc.

Alternatively you can directly output an executable with `clang` with:

```
C:\> clang hello.c -o hello.exe
```

The `-o hello.exe` is required because clang currently outputs `a.out` when neither `-o` nor `-c` are given.

3. Run the program using the just-in-time compiler:

```
C:\> lli hello.bc
```

4. Use the `llvm-dis` utility to take a look at the LLVM assembly code:

```
C:\> llvm-dis < hello.bc | more
```

5. Compile the program to object code using the LLC code generator:

```
C:\> llc -filetype=obj hello.bc
```

6. Link to binary using Microsoft link:

```
C:\> link hello.obj -defaultlib:libcmt
```

7. Execute the native code program:

```
C:\> hello.exe
```

2.8.5 Common Problems

If you are having problems building or using LLVM, or if you have any other general questions about LLVM, please consult the [Frequently Asked Questions](#) page.

2.8.6 Links

This document is just an **introduction** to how to use LLVM to do some simple things... there are many more interesting and complicated things that you can do that aren't documented here (but we'll gladly accept a patch if you want to write something up!). For more information about LLVM, check out:

- [LLVM homepage](#)
- [LLVM doxygen tree](#)

2.9 Frequently Asked Questions (FAQ)

- *License*
 - *Does the University of Illinois Open Source License really qualify as an “open source” license?*
 - *Can I modify LLVM source code and redistribute the modified source?*
 - *Can I modify the LLVM source code and redistribute binaries or other tools based on it, without redistributing the source?*
- *Source Code*
 - *In what language is LLVM written?*

- *How portable is the LLVM source code?*
 - *What API do I use to store a value to one of the virtual registers in LLVM IR's SSA representation?*
- *Source Languages*
 - *What source languages are supported?*
 - *I'd like to write a self-hosting LLVM compiler. How should I interface with the LLVM middle-end optimizers and back-end code generators?*
 - *What support is there for a higher level source language constructs for building a compiler?*
 - *I don't understand the `GetElementPtr` instruction. Help!*
- *Using the C and C++ Front Ends*
 - *Can I compile C or C++ code to platform-independent LLVM bytecode?*
- *Questions about code generated by the demo page*
 - *What is this `llvm.global_ctors` and `_GLOBAL__I_a...` stuff that happens when I `#include <iostream>`?*
 - *Where did all of my code go??*
 - *What is this “`undef`” thing that shows up in my code?*
 - *Why does `instcombine` + `simplifycfg` turn a call to a function with a mismatched calling convention into “unreachable”? Why not make the verifier reject it?*

2.9.1 License

Does the University of Illinois Open Source License really qualify as an “open source” license?

Yes, the license is [certified](#) by the Open Source Initiative (OSI).

Can I modify LLVM source code and redistribute the modified source?

Yes. The modified source distribution must retain the copyright notice and follow the three bulleted conditions listed in the [LLVM license](#).

Can I modify the LLVM source code and redistribute binaries or other tools based on it, without redistributing the source?

Yes. This is why we distribute LLVM under a less restrictive license than GPL, as explained in the first question above.

2.9.2 Source Code

In what language is LLVM written?

All of the LLVM tools and libraries are written in C++ with extensive use of the STL.

How portable is the LLVM source code?

The LLVM source code should be portable to most modern Unix-like operating systems. Most of the code is written in standard C++ with operating system services abstracted to a support library. The tools required to build and test LLVM have been ported to a plethora of platforms.

Some porting problems may exist in the following areas:

- The autoconf/makefile build system relies heavily on UNIX shell tools, like the Bourne Shell and sed. Porting to systems without these tools (MacOS 9, Plan 9) will require more effort.

What API do I use to store a value to one of the virtual registers in LLVM IR's SSA representation?

In short: you can't. It's actually kind of a silly question once you grok what's going on. Basically, in code like:

```
%result = add i32 %foo, %bar
```

, `%result` is just a name given to the `Value` of the `add` instruction. In other words, `%result` is the `add` instruction. The “assignment” doesn't explicitly “store” anything to any “virtual register”; the “=” is more like the mathematical sense of equality.

Longer explanation: In order to generate a textual representation of the IR, some kind of name has to be given to each instruction so that other instructions can textually reference it. However, the isomorphic in-memory representation that you manipulate from C++ has no such restriction since instructions can simply keep pointers to any other `Value`'s that they reference. In fact, the names of dummy numbered temporaries like `%1` are not explicitly represented in the in-memory representation at all (see `Value::getName()`).

2.9.3 Source Languages

What source languages are supported?

LLVM currently has full support for C and C++ source languages through [Clang](#). Many other language frontends have been written using LLVM, and an incomplete list is available at [projects with LLVM](#).

I'd like to write a self-hosting LLVM compiler. How should I interface with the LLVM middle-end optimizers and back-end code generators?

Your compiler front-end will communicate with LLVM by creating a module in the LLVM intermediate representation (IR) format. Assuming you want to write your language's compiler in the language itself (rather than C++), there are 3 major ways to tackle generating LLVM IR from a front-end:

1. Call into the LLVM libraries code using your language's FFI (foreign function interface).

- *for*: best tracks changes to the LLVM IR, .ll syntax, and .bc format
- *for*: enables running LLVM optimization passes without a emit/parse overhead
- *for*: adapts well to a JIT context
- *against*: lots of ugly glue code to write

2. Emit LLVM assembly from your compiler's native language.

- *for*: very straightforward to get started
- *against*: the .ll parser is slower than the bitcode reader when interfacing to the middle end
- *against*: it may be harder to track changes to the IR

3. Emit LLVM bitcode from your compiler's native language.

- *for*: can use the more-efficient bitcode reader when interfacing to the middle end
- *against*: you'll have to re-engineer the LLVM IR object model and bitcode writer in your language
- *against*: it may be harder to track changes to the IR

If you go with the first option, the C bindings in `include/llvm-c` should help a lot, since most languages have strong support for interfacing with C. The most common hurdle with calling C from managed code is interfacing with the garbage collector. The C interface was designed to require very little memory management, and so is straightforward in this regard.

What support is there for a higher level source language constructs for building a compiler?

Currently, there isn't much. LLVM supports an intermediate representation which is useful for code representation but will not support the high level (abstract syntax tree) representation needed by most compilers. There are no facilities for lexical nor semantic analysis.

I don't understand the `GetElementPtr` instruction. Help!

See [The Often Misunderstood GEP Instruction](#).

2.9.4 Using the C and C++ Front Ends

Can I compile C or C++ code to platform-independent LLVM bitcode?

No. C and C++ are inherently platform-dependent languages. The most obvious example of this is the preprocessor. A very common way that C code is made portable is by using the preprocessor to include platform-specific code. In practice, information about other platforms is lost after preprocessing, so the result is inherently dependent on the platform that the preprocessing was targeting.

Another example is `sizeof`. It's common for `sizeof(long)` to vary between platforms. In most C front-ends, `sizeof` is expanded to a constant immediately, thus hard-wiring a platform-specific detail.

Also, since many platforms define their ABIs in terms of C, and since LLVM is lower-level than C, front-ends currently must emit platform-specific IR in order to have the result conform to the platform ABI.

2.9.5 Questions about code generated by the demo page

What is this `llvm.global_ctors` and `_GLOBAL__I_a...` stuff that happens when I `#include <iostream>`?

If you `#include` the `<iostream>` header into a C++ translation unit, the file will probably use the `std::cin/std::cout/...` global objects. However, C++ does not guarantee an order of initialization between static objects in different translation units, so if a static ctor/dtor in your `.cpp` file used `std::cout`, for example, the object would not necessarily be automatically initialized before your use.

To make `std::cout` and friends work correctly in these scenarios, the STL that we use declares a static object that gets created in every translation unit that includes `<iostream>`. This object has a static constructor and destructor that initializes and destroys the global `iostream` objects before they could possibly be used in the file. The code that you see in the `.ll` file corresponds to the constructor and destructor registration code.

If you would like to make it easier to *understand* the LLVM code generated by the compiler in the demo page, consider using `printf()` instead of `iostreams` to print values.

Where did all of my code go??

If you are using the LLVM demo page, you may often wonder what happened to all of the code that you typed in. Remember that the demo script is running the code through the LLVM optimizers, so if your code doesn't actually do anything useful, it might all be deleted.

To prevent this, make sure that the code is actually needed. For example, if you are computing some expression, return the value from the function instead of leaving it in a local variable. If you really want to constrain the optimizer, you can read from and assign to `volatile` global variables.

What is this “undef” thing that shows up in my code?

`undef` is the LLVM way of representing a value that is not defined. You can get these if you do not initialize a variable before you use it. For example, the C function:

```
int X() { int i; return i; }
```

Is compiled to “`ret i32 undef`” because “`i`” never has a value specified for it.

Why does `instcombine + simplifycfg` turn a call to a function with a mismatched calling convention into “unreachable”? Why not make the verifier reject it?

This is a common problem run into by authors of front-ends that are using custom calling conventions: you need to make sure to set the right calling convention on both the function and on each call to the function. For example, this code:

```
define fastcc void @foo() {
    ret void
}
define void @bar() {
    call void @foo()
    ret void
}
```

Is optimized to:

```
define fastcc void @foo() {
    ret void
}
define void @bar() {
    unreachable
}
```

... with “`opt -instcombine -simplifycfg`”. This often bites people because “all their code disappears”. Setting the calling convention on the caller and callee is required for indirect calls to work, so people often ask why not make the verifier reject this sort of thing.

The answer is that this code has undefined behavior, but it is not illegal. If we made it illegal, then every transformation that could potentially create this would have to ensure that it doesn't, and there is valid code that can create this sort of construct (in dead code). The sorts of things that can cause this to happen are fairly contrived, but we still need to accept them. Here's an example:

```
define fastcc void @foo() {
    ret void
}
define internal void @bar(void() * %FP, i1 %cond) {
```

```

    br i1 %cond, label %T, label %F
T:
    call void %FP()
    ret void
F:
    call fastcc void %FP()
    ret void
}
define void @test() {
    %X = or i1 false, false
    call void @bar(void()) * @foo, i1 %X
    ret void
}

```

In this example, “test” always passes @foo/false into bar, which ensures that it is dynamically called with the right calling conv (thus, the code is perfectly well defined). If you run this through the inliner, you get this (the explicit “or” is there so that the inliner doesn’t dead code eliminate a bunch of stuff):

```

define fastcc void @foo() {
    ret void
}
define void @test() {
    %X = or i1 false, false
    br i1 %X, label %T.i, label %F.i
T.i:
    call void @foo()
    br label %bar.exit
F.i:
    call fastcc void @foo()
    br label %bar.exit
bar.exit:
    ret void
}

```

Here you can see that the inlining pass made an undefined call to @foo with the wrong calling convention. We really don’t want to make the inliner have to know about this sort of thing, so it needs to be valid code. In this case, dead code elimination can trivially remove the undefined code. However, if %X was an input argument to @test, the inliner would produce this:

```

define fastcc void @foo() {
    ret void
}

define void @test(i1 %X) {
    br i1 %X, label %T.i, label %F.i
T.i:
    call void @foo()
    br label %bar.exit
F.i:
    call fastcc void @foo()
    br label %bar.exit
bar.exit:
    ret void
}

```

The interesting thing about this is that %X *must* be false for the code to be well-defined, but no amount of dead code elimination will be able to delete the broken call as unreachable. However, since `instcombine/simplifycfg`

turns the undefined call into unreachable, we end up with a branch on a condition that goes to unreachable: a branch to unreachable can never happen, so “`-inline -instcombine -simplifycfg`” is able to produce:

```
define fastcc void @foo() {
    ret void
}
define void @test(i1 %X) {
F.i:
    call fastcc void @foo()
    ret void
}
```

2.10 The LLVM Lexicon

Note: This document is a work in progress!

2.10.1 Definitions

A

ADCE Aggressive Dead Code Elimination

AST Abstract Syntax Tree.

Due to Clang’s influence (mostly the fact that parsing and semantic analysis are so intertwined for C and especially C++), the typical working definition of AST in the LLVM community is roughly “the compiler’s first complete symbolic (as opposed to textual) representation of an input program”. As such, an “AST” might be a more general graph instead of a “tree” (consider the symbolic representation for the type of a typical “linked list node”). This working definition is closer to what some authors call an “annotated abstract syntax tree”.

Consult your favorite compiler book or search engine for more details.

B

BB Vectorization Basic-Block Vectorization

BDCE Bit-tracking dead code elimination. Some bit-wise instructions (shifts, ands, ors, etc.) “kill” some of their input bits – that is, they make it such that those bits can be either zero or one without affecting control or data flow of a program. The BDCE pass removes instructions that only compute these dead bits.

BURS Bottom Up Rewriting System — A method of instruction selection for code generation. An example is the [BURG](#) tool.

C

CFI Call Frame Information. Used in DWARF debug info and in C++ unwind info to show how the function prolog lays out the stack frame.

CIE Common Information Entry. A kind of CFI used to reduce the size of FDEs. The compiler creates a CIE which contains the information common across all the FDEs. Each FDE then points to its CIE.

CSE Common Subexpression Elimination. An optimization that removes common subexpression computation. For example $(a+b) * (a+b)$ has two subexpressions that are the same: $(a+b)$. This optimization would perform the addition only once and then perform the multiply (but only if it's computationally correct/safe).

D

DAG Directed Acyclic Graph

Derived Pointer A pointer to the interior of an object, such that a garbage collector is unable to use the pointer for reachability analysis. While a derived pointer is live, the corresponding object pointer must be kept in a root, otherwise the collector might free the referenced object. With copying collectors, derived pointers pose an additional hazard that they may be invalidated at any *safe point*. This term is used in opposition to *object pointer*.

DSA Data Structure Analysis

DSE Dead Store Elimination

F

FCA First Class Aggregate

FDE Frame Description Entry. A kind of CFI used to describe the stack frame of one function.

G

GC Garbage Collection. The practice of using reachability analysis instead of explicit memory management to reclaim unused memory.

GVN Global Value Numbering. GVN is a pass that partitions values computed by a function into congruence classes. Values ending up in the same congruence class are guaranteed to be the same for every execution of the program. In that respect, congruency is a compile-time approximation of equivalence of values at runtime.

H

Heap In garbage collection, the region of memory which is managed using reachability analysis.

I

IPA Inter-Procedural Analysis. Refers to any variety of code analysis that occurs between procedures, functions or compilation units (modules).

IPO Inter-Procedural Optimization. Refers to any variety of code optimization that occurs between procedures, functions or compilation units (modules).

ISel Instruction Selection

L

LCSSA Loop-Closed Static Single Assignment Form

LGTM “Looks Good To Me”. In a review thread, this indicates that the reviewer thinks that the patch is okay to commit.

LICM Loop Invariant Code Motion

LSDA Language Specific Data Area. C++ “zero cost” unwinding is built on top a generic unwinding mechanism. As the unwinder walks each frame, it calls a “personality” function to do language specific analysis. Each function’s FDE points to an optional LSDA which is passed to the personality function. For C++, the LSDA contain info about the type and location of catch statements in that function.

Load-VN Load Value Numbering

LTO Link-Time Optimization

M

MC Machine Code

N

NFC “No functional change”. Used in a commit message to indicate that a patch is a pure refactoring/cleanup. Usually used in the first line, so it is visible without opening the actual commit email.

O

Object Pointer A pointer to an object such that the garbage collector is able to trace references contained within the object. This term is used in opposition to *derived pointer*.

P

PR Problem report. A bug filed on the [LLVM Bug Tracking System](#).

PRE Partial Redundancy Elimination

R

RAUW

Replace All Uses With. The functions `User::replaceUsesOfWith()`, `Value::replaceAllUsesWith()`, and `Constant::replaceUsesOfWithOnConstant()` implement the replacement of one Value with another by iterating over its def/use chain and fixing up all of the pointers to point to the new value. See also [def/use chains](#).

Reassociation Rearranging associative expressions to promote better redundancy elimination and other optimization. For example, changing $(A+B-A)$ into $(B+A-A)$, permitting it to be optimized into $(B+0)$ then (B) .

Root In garbage collection, a pointer variable lying outside of the *heap* from which the collector begins its reachability analysis. In the context of code generation, “root” almost always refers to a “stack root” — a local or temporary variable within an executing function.

RPO Reverse postorder

S

Safe Point In garbage collection, it is necessary to identify *stack roots* so that reachability analysis may proceed. It may be infeasible to provide this information for every instruction, so instead the information may be calculated only at designated safe points. With a copying collector, *derived pointers* must not be retained across safe points and *object pointers* must be reloaded from stack roots.

SDISel Selection DAG Instruction Selection.

SCC Strongly Connected Component

SCCP Sparse Conditional Constant Propagation

SLP Superword-Level Parallelism, same as *Basic-Block Vectorization*.

Splat Splat refers to a vector of identical scalar elements.

The term is based on the PowerPC AltiVec instructions that provided this functionality in hardware. For example, “vsplth” and the corresponding software intrinsic “vec_splat()”. Examples of other hardware names for this action include “duplicate” (ARM) and “broadcast” (x86).

SRoA Scalar Replacement of Aggregates

SSA Static Single Assignment

Stack Map In garbage collection, metadata emitted by the code generator which identifies *roots* within the stack frame of an executing function.

T

TBAA Type-Based Alias Analysis

2.11 How To Add Your Build Configuration To LLVM Buildbot Infrastructure

2.11.1 Introduction

This document contains information about adding a build configuration and builds slave to private slave builder to LLVM Buildbot Infrastructure.

2.11.2 Buildmasters

There are two buildmasters running.

- The main buildmaster at <http://lab.llvm.org:8011>. All builders attached to this machine will notify commit authors every time they break the build.
- The staging buildbot at <http://lab.llvm.org:8014>. All builders attached to this machine will be completely silent by default when the build is broken. Builders for experimental backends should generally be attached to this buildmaster.

2.11.3 Steps To Add Builder To LLVM Buildbot

Volunteers can provide their build machines to work as build slaves to public LLVM Buildbot.

Here are the steps you can follow to do so:

1. Check the existing build configurations to make sure the one you are interested in is not covered yet or gets built on your computer much faster than on the existing one. We prefer faster builds so developers will get feedback sooner after changes get committed.
2. The computer you will be registering with the LLVM buildbot infrastructure should have all dependencies installed and you can actually build your configuration successfully. Please check what degree of parallelism (-j param) would give the fastest build. You can build multiple configurations on one computer.
3. Install builds slave (currently we are using buildbot version 0.8.5). Depending on the platform, builds slave could be available to download and install with your package manager, or you can download it directly from <http://trac.buildbot.net> and install it manually.
4. Create a designated user account, your builds slave will be running under, and set appropriate permissions.
5. Choose the builds slave root directory (all builds will be placed under it), builds slave access name and password the build master will be using to authenticate your builds slave.
6. Create a builds slave in context of that builds slave account. Point it to the **lab.llvm.org** port **9990** (see [Buildbot documentation](#), [Creating a slave](#) for more details) by running the following command:

```
$ builds slave create-slave <builds slave-root-directory> \
    lab.llvm.org:9990 \
    <builds slave-access-name> <builds slave-access-password>
```

To point a slave to silent master please use lab.llvm.org:9994 instead of lab.llvm.org:9990.

7. Fill the builds slave description and admin name/e-mail. Here is an example of the builds slave description:

```
Windows 7 x64
Core i7 (2.66GHz), 16GB of RAM

g++.exe (TDM-1 mingw32) 4.4.0
GNU Binutils 2.19.1
cmake version 2.8.4
Microsoft(R) 32-bit C/C++ Optimizing Compiler Version 16.00.40219.01 for 80x86
```

8. Make sure you can actually start the builds slave successfully. Then set up your builds slave to start automatically at the start up time. See the buildbot documentation for help. You may want to restart your computer to see if it works.
9. Send a patch which adds your build slave and your builder to zorg.
 - slaves are added to buildbot/osuosl/master/config/slaves.py
 - builders are added to buildbot/osuosl/master/config/builders.py

Please make sure your builder name and its builddir are unique through the file.

It is possible to whitelist email addresses to unconditionally receive notifications on build failure; for this you'll need to add an InformativeMailNotifier to buildbot/osuosl/master/config/status.py. This is particularly useful for the staging buildmaster which is silent otherwise.

10. Send the builds slave access name and the access password directly to [Galina Kistanova](#), and wait till she will let you know that your changes are applied and buildmaster is reconfigured.

11. Check the status of your builds slave on the [Waterfall Display](http://lab.llvm.org:8011/buildslaves/<your-buildslave-name>) to make sure it is connected, and <http://lab.llvm.org:8011/buildslaves/<your-buildslave-name>> to see if administrator contact and slave information are correct.
12. Wait for the first build to succeed and enjoy.

2.12 yaml2obj

yaml2obj takes a YAML description of an object file and converts it to a binary file.

```
$ yaml2obj input-file
```

Outputs the binary to stdout.

2.12.1 COFF Syntax

Here's a sample COFF file.

```
header:
  Machine: IMAGE_FILE_MACHINE_I386 # (0x14C)

sections:
  - Name: .text
    Characteristics: [ IMAGE_SCN_CNT_CODE
                      , IMAGE_SCN_ALIGN_16BYTES
                      , IMAGE_SCN_MEM_EXECUTE
                      , IMAGE_SCN_MEM_READ
                      ] # 0x60500020
    SectionData:
      ↪ "\x83\xEC\x0C\xC7\x44\x24\x08\x00\x00\x00\x00\xC7\x04\x24\x00\x00\x00\x00\xE8\x00\x00\x00\x00\xE8\x00\x00\x00\x00"
      ↪ # |....D$......$......D$....../

symbols:
  - Name: .text
    Value: 0
    SectionNumber: 1
    SimpleType: IMAGE_SYM_TYPE_NULL # (0)
    ComplexType: IMAGE_SYM_DTYPE_NULL # (0)
    StorageClass: IMAGE_SYM_CLASS_STATIC # (3)
    NumberOfAuxSymbols: 1
    AuxiliaryData:
      ↪ "\x24\x00\x00\x00\x03\x00\x00\x00\x00\x00\x00\x00\x01\x00\x00\x00\x00\x00" # |$.
      ↪ ...../

  - Name: _main
    Value: 0
    SectionNumber: 1
    SimpleType: IMAGE_SYM_TYPE_NULL # (0)
    ComplexType: IMAGE_SYM_DTYPE_NULL # (0)
    StorageClass: IMAGE_SYM_CLASS_EXTERNAL # (2)
```

Here's a simplified [Kwalify](#) schema with an extension to allow alternate types.

```
type: map
mapping:
```

```

header:
  type: map
  mapping:
    Machine: [ {type: str, enum:
                  [ IMAGE_FILE_MACHINE_UNKNOWN
                    , IMAGE_FILE_MACHINE_AM33
                    , IMAGE_FILE_MACHINE_AMD64
                    , IMAGE_FILE_MACHINE_ARM
                    , IMAGE_FILE_MACHINE_ARMNT
                    , IMAGE_FILE_MACHINE_ARM64
                    , IMAGE_FILE_MACHINE_EBC
                    , IMAGE_FILE_MACHINE_I386
                    , IMAGE_FILE_MACHINE_IA64
                    , IMAGE_FILE_MACHINE_M32R
                    , IMAGE_FILE_MACHINE_MIPS16
                    , IMAGE_FILE_MACHINE_MIPSFPU
                    , IMAGE_FILE_MACHINE_MIPSFPU16
                    , IMAGE_FILE_MACHINE_POWERPC
                    , IMAGE_FILE_MACHINE_POWERPCFP
                    , IMAGE_FILE_MACHINE_R4000
                    , IMAGE_FILE_MACHINE_SH3
                    , IMAGE_FILE_MACHINE_SH3DSP
                    , IMAGE_FILE_MACHINE_SH4
                    , IMAGE_FILE_MACHINE_SH5
                    , IMAGE_FILE_MACHINE_THUMB
                    , IMAGE_FILE_MACHINE_WCEMIPSV2
                  ]}
              , {type: int}
            ]
    Characteristics:
      - type: seq
        sequence:
          - type: str
            enum: [ IMAGE_FILE_RELOCS_STRIPPED
                  , IMAGE_FILE_EXECUTABLE_IMAGE
                  , IMAGE_FILE_LINE_NUMS_STRIPPED
                  , IMAGE_FILE_LOCAL_SYMS_STRIPPED
                  , IMAGE_FILE_AGGRESSIVE_WS_TRIM
                  , IMAGE_FILE_LARGE_ADDRESS_AWARE
                  , IMAGE_FILE_BYTES_REVERSED_LO
                  , IMAGE_FILE_32BIT_MACHINE
                  , IMAGE_FILE_DEBUG_STRIPPED
                  , IMAGE_FILE_REMOVABLE_RUN_FROM_SWAP
                  , IMAGE_FILE_NET_RUN_FROM_SWAP
                  , IMAGE_FILE_SYSTEM
                  , IMAGE_FILE_DLL
                  , IMAGE_FILE_UP_SYSTEM_ONLY
                  , IMAGE_FILE_BYTES_REVERSED_HI
                ]
          - type: int
    sections:
      type: seq
      sequence:
        - type: map
          mapping:
            Name: {type: str}
            Characteristics:
              - type: seq

```

```
sequence:
  - type: str
    enum: [ IMAGE_SCN_TYPE_NO_PAD
            , IMAGE_SCN_CNT_CODE
            , IMAGE_SCN_CNT_INITIALIZED_DATA
            , IMAGE_SCN_CNT_UNINITIALIZED_DATA
            , IMAGE_SCN_LNK_OTHER
            , IMAGE_SCN_LNK_INFO
            , IMAGE_SCN_LNK_REMOVE
            , IMAGE_SCN_LNK_COMDAT
            , IMAGE_SCN_GPREL
            , IMAGE_SCN_MEM_PURGEABLE
            , IMAGE_SCN_MEM_16BIT
            , IMAGE_SCN_MEM_LOCKED
            , IMAGE_SCN_MEM_PRELOAD
            , IMAGE_SCN_ALIGN_1BYTES
            , IMAGE_SCN_ALIGN_2BYTES
            , IMAGE_SCN_ALIGN_4BYTES
            , IMAGE_SCN_ALIGN_8BYTES
            , IMAGE_SCN_ALIGN_16BYTES
            , IMAGE_SCN_ALIGN_32BYTES
            , IMAGE_SCN_ALIGN_64BYTES
            , IMAGE_SCN_ALIGN_128BYTES
            , IMAGE_SCN_ALIGN_256BYTES
            , IMAGE_SCN_ALIGN_512BYTES
            , IMAGE_SCN_ALIGN_1024BYTES
            , IMAGE_SCN_ALIGN_2048BYTES
            , IMAGE_SCN_ALIGN_4096BYTES
            , IMAGE_SCN_ALIGN_8192BYTES
            , IMAGE_SCN_LNK_NRELOC_OVFL
            , IMAGE_SCN_MEM_DISCARDABLE
            , IMAGE_SCN_MEM_NOT_CACHED
            , IMAGE_SCN_MEM_NOT_PAGED
            , IMAGE_SCN_MEM_SHARED
            , IMAGE_SCN_MEM_EXECUTE
            , IMAGE_SCN_MEM_READ
            , IMAGE_SCN_MEM_WRITE
          ]
    - type: int
      SectionData: {type: str}
symbols:
  type: seq
  sequence:
    - type: map
      mapping:
        Name: {type: str}
        Value: {type: int}
        SectionNumber: {type: int}
        SimpleType: [ {type: str, enum: [ IMAGE_SYM_TYPE_NULL
                                          , IMAGE_SYM_TYPE_VOID
                                          , IMAGE_SYM_TYPE_CHAR
                                          , IMAGE_SYM_TYPE_SHORT
                                          , IMAGE_SYM_TYPE_INT
                                          , IMAGE_SYM_TYPE_LONG
                                          , IMAGE_SYM_TYPE_FLOAT
                                          , IMAGE_SYM_TYPE_DOUBLE
                                          , IMAGE_SYM_TYPE_STRUCT
                                          , IMAGE_SYM_TYPE_UNION
```



```

        , IMAGE_SYM_TYPE_ENUM
        , IMAGE_SYM_TYPE_MOE
        , IMAGE_SYM_TYPE_BYTE
        , IMAGE_SYM_TYPE_WORD
        , IMAGE_SYM_TYPE_UINT
        , IMAGE_SYM_TYPE_DWORD
    ]}

    , {type: int}
]
ComplexType: [ {type: str, enum: [ IMAGE_SYM_DTYPE_NULL
                                , IMAGE_SYM_DTYPE_POINTER
                                , IMAGE_SYM_DTYPE_FUNCTION
                                , IMAGE_SYM_DTYPE_ARRAY
                                ]}

    , {type: int}
]
StorageClass: [ {type: str, enum:
                [ IMAGE_SYM_CLASS_END_OF_FUNCTION
                , IMAGE_SYM_CLASS_NULL
                , IMAGE_SYM_CLASS_AUTOMATIC
                , IMAGE_SYM_CLASS_EXTERNAL
                , IMAGE_SYM_CLASS_STATIC
                , IMAGE_SYM_CLASS_REGISTER
                , IMAGE_SYM_CLASS_EXTERNAL_DEF
                , IMAGE_SYM_CLASS_LABEL
                , IMAGE_SYM_CLASS_UNDEFINED_LABEL
                , IMAGE_SYM_CLASS_MEMBER_OF_STRUCT
                , IMAGE_SYM_CLASS_ARGUMENT
                , IMAGE_SYM_CLASS_STRUCT_TAG
                , IMAGE_SYM_CLASS_MEMBER_OF_UNION
                , IMAGE_SYM_CLASS_UNION_TAG
                , IMAGE_SYM_CLASS_TYPE_DEFINITION
                , IMAGE_SYM_CLASS_UNDEFINED_STATIC
                , IMAGE_SYM_CLASS_ENUM_TAG
                , IMAGE_SYM_CLASS_MEMBER_OF_ENUM
                , IMAGE_SYM_CLASS_REGISTER_PARAM
                , IMAGE_SYM_CLASS_BIT_FIELD
                , IMAGE_SYM_CLASS_BLOCK
                , IMAGE_SYM_CLASS_FUNCTION
                , IMAGE_SYM_CLASS_END_OF_STRUCT
                , IMAGE_SYM_CLASS_FILE
                , IMAGE_SYM_CLASS_SECTION
                , IMAGE_SYM_CLASS_WEAK_EXTERNAL
                , IMAGE_SYM_CLASS_CLR_TOKEN
                ]}

    , {type: int}
]

```

2.13 How to submit an LLVM bug report

2.13.1 Introduction - Got bugs?

If you're working with LLVM and run into a bug, we definitely want to know about it. This document describes what you can do to increase the odds of getting it fixed quickly.

Basically you have to do two things at a minimum. First, decide whether the bug *crashes the compiler* (or an LLVM pass), or if the compiler is *miscompiling* the program (i.e., the compiler successfully produces an executable, but it doesn't run right). Based on what type of bug it is, follow the instructions in the linked section to narrow down the bug so that the person who fixes it will be able to find the problem more easily.

Once you have a reduced test-case, go to [the LLVM Bug Tracking System](#) and fill out the form with the necessary details (note that you don't need to pick a category, just use the “new-bugs” category if you're not sure). The bug description should contain the following information:

- All information necessary to reproduce the problem.
- The reduced test-case that triggers the bug.
- The location where you obtained LLVM (if not from our Subversion repository).

Thanks for helping us make LLVM better!

2.13.2 Crashing Bugs

More often than not, bugs in the compiler cause it to crash—often due to an assertion failure of some sort. The most important piece of the puzzle is to figure out if it is crashing in the GCC front-end or if it is one of the LLVM libraries (e.g. the optimizer or code generator) that has problems.

To figure out which component is crashing (the front-end, optimizer or code generator), run the `clang` command line as you were when the crash occurred, but with the following extra command line options:

- `-O0 -emit-llvm`: If `clang` still crashes when passed these options (which disable the optimizer and code generator), then the crash is in the front-end. Jump ahead to the section on *front-end bugs*.
- `-emit-llvm`: If `clang` crashes with this option (which disables the code generator), you found an optimizer bug. Jump ahead to *compile-time optimization bugs*.
- Otherwise, you have a code generator crash. Jump ahead to *code generator bugs*.

Front-end bugs

If the problem is in the front-end, you should re-run the same `clang` command that resulted in the crash, but add the `-save-temps` option. The compiler will crash again, but it will leave behind a `foo.i` file (containing preprocessed C source code) and possibly `foo.s` for each compiled `foo.c` file. Send us the `foo.i` file, along with the options you passed to `clang`, and a brief description of the error it caused.

The `delta` tool helps to reduce the preprocessed file down to the smallest amount of code that still replicates the problem. You're encouraged to use `delta` to reduce the code to make the developers' lives easier. [This website](#) has instructions on the best way to use `delta`.

Compile-time optimization bugs

If you find that a bug crashes in the optimizer, compile your test-case to a `.bc` file by passing “`-emit-llvm -O0 -c -o foo.bc`”. Then run:

```
opt -O3 -debug-pass=Arguments foo.bc -disable-output
```

This command should do two things: it should print out a list of passes, and then it should crash in the same way as `clang`. If it doesn't crash, please follow the instructions for a *front-end bug*.

If this does crash, then you should be able to debug this with the following `bugpoint` command:

```
bugpoint foo.bc <list of passes printed by opt>
```

Please run this, then file a bug with the instructions and reduced .bc files that bugpoint emits. If something goes wrong with bugpoint, please submit the “foo.bc” file and the list of passes printed by opt.

Code generator bugs

If you find a bug that crashes clang in the code generator, compile your source file to a .bc file by passing “-emit-llvm -c -o foo.bc” to clang (in addition to the options you already pass). Once you have foo.bc, one of the following commands should fail:

1. `llc foo.bc`
2. `llc foo.bc -relocation-model=pic`
3. `llc foo.bc -relocation-model=static`

If none of these crash, please follow the instructions for a *front-end bug*. If one of these do crash, you should be able to reduce this with one of the following bugpoint command lines (use the one corresponding to the command above that failed):

1. `bugpoint -run-llc foo.bc`
2. `bugpoint -run-llc foo.bc --tool-args -relocation-model=pic`
3. `bugpoint -run-llc foo.bc --tool-args -relocation-model=static`

Please run this, then file a bug with the instructions and reduced .bc file that bugpoint emits. If something goes wrong with bugpoint, please submit the “foo.bc” file and the option that llc crashes with.

2.13.3 Miscompilations

If clang successfully produces an executable, but that executable doesn’t run right, this is either a bug in the code or a bug in the compiler. The first thing to check is to make sure it is not using undefined behavior (e.g. reading a variable before it is defined). In particular, check to see if the program `valgrind`’s clean, passes `purify`, or some other memory checker tool. Many of the “LLVM bugs” that we have chased down ended up being bugs in the program being compiled, not LLVM.

Once you determine that the program itself is not buggy, you should choose which code generator you wish to compile the program with (e.g. LLC or the JIT) and optionally a series of LLVM passes to run. For example:

```
bugpoint -run-llc [... optzn passes ...] file-to-test.bc --args -- [program arguments]
```

bugpoint will try to narrow down your list of passes to the one pass that causes an error, and simplify the bitcode file as much as it can to assist you. It will print a message letting you know how to reproduce the resulting error.

2.13.4 Incorrect code generation

Similarly to debugging incorrect compilation by mis-behaving passes, you can debug incorrect code generation by either LLC or the JIT, using bugpoint. The process bugpoint follows in this case is to try to narrow the code down to a function that is miscompiled by one or the other method, but since for correctness, the entire program must be run, bugpoint will compile the code it deems to not be affected with the C Backend, and then link in the shared object it generates.

To debug the JIT:

```
bugpoint -run-jit -output=[correct output file] [bitcode file] \
--tool-args -- [arguments to pass to lli] \
--args -- [program arguments]
```

Similarly, to debug the LLC, one would run:

```
bugpoint -run-llc -output=[correct output file] [bitcode file] \
--tool-args -- [arguments to pass to llc] \
--args -- [program arguments]
```

Special note: if you are debugging MultiSource or SPEC tests that already exist in the `llvm/test` hierarchy, there is an easier way to debug the JIT, LLC, and CBE, using the pre-written Makefile targets, which will pass the program options specified in the Makefiles:

```
cd llvm/test/../../program
make bugpoint-jit
```

At the end of a successful `bugpoint` run, you will be presented with two bitcode files: a *safe* file which can be compiled with the C backend and the *test* file which either LLC or the JIT mis-codegenerates, and thus causes the error.

To reproduce the error that `bugpoint` found, it is sufficient to do the following:

1. Regenerate the shared object from the safe bitcode file:

```
llc -march=c safe.bc -o safe.c
gcc -shared safe.c -o safe.so
```

2. If debugging LLC, compile test bitcode native and link with the shared object:

```
llc test.bc -o test.s
gcc test.s safe.so -o test.llc
./test.llc [program options]
```

3. If debugging the JIT, load the shared object and supply the test bitcode:

```
lli -load=safe.so test.bc [program options]
```

2.14 Sphinx Quickstart Template

2.14.1 Introduction and Quickstart

This document is meant to get you writing documentation as fast as possible even if you have no previous experience with Sphinx. The goal is to take someone in the state of “I want to write documentation and get it added to LLVM’s docs” and turn that into useful documentation mailed to `llvm-commits` with as little nonsense as possible.

You can find this document in `docs/SphinxQuickstartTemplate.rst`. You should copy it, open the new file in your text editor, write your docs, and then send the new document to `llvm-commits` for review.

Focus on *content*. It is easy to fix the Sphinx (`reStructuredText`) syntax later if necessary, although `reStructuredText` tries to imitate common plain-text conventions so it should be quite natural. A basic knowledge of `reStructuredText` syntax is useful when writing the document, so the last ~half of this document (starting with [Example Section](#)) gives examples which should cover 99% of use cases.

Let me say that again: focus on *content*. But if you really need to verify Sphinx’s output, see `docs/README.txt` for information.

Once you have finished with the content, please send the `.rst` file to `llvm-commits` for review.

2.14.2 Guidelines

Try to answer the following questions in your first section:

1. Why would I want to read this document?
2. What should I know to be able to follow along with this document?
3. What will I have learned by the end of this document?

Common names for the first section are `Introduction`, `Overview`, or `Background`.

If possible, make your document a “how to”. Give it a name `HowTo*.rst` like the other “how to” documents. This format is usually the easiest for another person to understand and also the most useful.

You generally should not be writing documentation other than a “how to” unless there is already a “how to” about your topic. The reason for this is that without a “how to” document to read first, it is difficult for a person to understand a more advanced document.

Focus on content (yes, I had to say it again).

The rest of this document shows example `reStructuredText` markup constructs that are meant to be read by you in your text editor after you have copied this file into a new file for the documentation you are about to write.

2.14.3 Example Section

Your text can be *emphasized*, **bold**, or `monospace`.

Use blank lines to separate paragraphs.

Headings (like `Example Section` just above) give your document its structure. Use the same kind of adornments (e.g. `=====` vs. `-----`) as are used in this document. The adornment must be the same length as the text above it. For Vim users, variations of `yyPvR=` might be handy.

Example Subsection

Make a link [like this](#). There is also a more sophisticated syntax which [can be more readable](#) for longer links since it disrupts the flow less. You can put the `.. _`link text`: <URL>` block pretty much anywhere later in the document.

Lists can be made like this:

1. A list starting with `# .` will be automatically numbered.
2. This is a second list element.
 - (a) Use indentation to create nested lists.

You can also use unordered lists.

- Stuff.
 - Deeper stuff.
- More stuff.

Example Subsubsection

You can make blocks of code like this:

```
int main() {  
    return 0;  
}
```

For a shell session, use a `console` code block (some existing docs use `bash`):

```
$ echo "Goodbye cruel world!"  
$ rm -rf /
```

If you need to show LLVM IR use the `llvm` code block.

```
define i32 @test1() {  
entry:  
    ret i32 0  
}
```

Some other common code blocks you might need are `c`, `objc`, `make`, and `cmake`. If you need something beyond that, you can look at the [full list](#) of supported code blocks.

However, don't waste time fiddling with syntax highlighting when you could be adding meaningful content. When in doubt, show preformatted text without any syntax highlighting like this:

```
      .  
      +:.  
    .::: ::  
  .+++++: ::+:..  
    .:+      :  
  ::::::::::: .+.  
  ..:+      ::      :  
.....+:+      ..      :  
  :++      ..      :  
  .+++++:      :  
  ..      .+      ::  
      +.:      .::+.  
      ...+      .:  
      .++:..  
      ...
```

Hopefully you won't need to be this deep

If you need to do fancier things than what has been shown in this document, you can mail the list or check Sphinx's [reStructuredText Primer](#).

2.15 Code Reviews with Phabricator

- *Sign up*
- *Requesting a review via the command line*

- *Requesting a review via the web interface*
- *Reviewing code with Phabricator*
- *Committing a change*
 - *Subversion and Arcanist*
 - *git-svn and Arcanist*
- *Abandoning a change*
- *Status*

If you prefer to use a web user interface for code reviews, you can now submit your patches for Clang and LLVM at [LLVM's Phabricator](#) instance.

While Phabricator is a useful tool for some, the relevant -commits mailing list is the system of record for all LLVM code review. The mailing list should be added as a subscriber on all reviews, and Phabricator users should be prepared to respond to free-form comments in mail sent to the commits list.

2.15.1 Sign up

To get started with Phabricator, navigate to <http://reviews.llvm.org> and click the power icon in the top right. You can register with a GitHub account, a Google account, or you can create your own profile.

Make *sure* that the email address registered with Phabricator is subscribed to the relevant -commits mailing list. If you are not subscribed to the commit list, all mail sent by Phabricator on your behalf will be held for moderation.

Note that if you use your Subversion user name as Phabricator user name, Phabricator will automatically connect your submits to your Phabricator user in the [Code Repository Browser](#).

2.15.2 Requesting a review via the command line

Phabricator has a tool called *Arcanist* to upload patches from the command line. To get you set up, follow the [Arcanist Quick Start](#) instructions.

You can learn more about how to use arc to interact with Phabricator in the [Arcanist User Guide](#).

2.15.3 Requesting a review via the web interface

The tool to create and review patches in Phabricator is called *Differential*.

Note that you can upload patches created through various diff tools, including git and svn. To make reviews easier, please always include **as much context as possible** with your diff! Don't worry, Phabricator will automatically send a diff with a smaller context in the review email, but having the full file in the web interface will help the reviewer understand your code.

To get a full diff, use one of the following commands (or just use Arcanist to upload your patch):

- `git show HEAD -U999999 > mypatch.patch`
- `git format-patch -U999999 @{u}`
- `svn diff --diff-cmd=diff -x -U999999`

To upload a new patch:

- Click *Differential*.

- Click + *Create Diff*.
- Paste the text diff or browse to the patch file. Click *Create Diff*.
- Leave the Repository field blank.
- Leave the drop down on *Create a new Revision...* and click *Continue*.
- Enter a descriptive title and summary. The title and summary are usually in the form of a *commit message*.
- Add reviewers (see below for advice) and subscribe mailing lists that you want to be included in the review. If your patch is for LLVM, add llvm-commits as a Subscriber; if your patch is for Clang, add cfe-commits.
- Leave the Repository and Project fields blank.
- Click *Save*.

To submit an updated patch:

- Click *Differential*.
- Click + *Create Diff*.
- Paste the updated diff or browse to the updated patch file. Click *Create Diff*.
- Select the review you want to from the *Attach To* dropdown and click *Continue*.
- Leave the Repository and Project fields blank.
- Add comments about the changes in the new diff. Click *Save*.

Choosing reviewers: You typically pick one or two people as initial reviewers. This choice is not crucial, because you are merely suggesting and not requiring them to participate. Many people will see the email notification on cfe-commits or llvm-commits, and if the subject line suggests the patch is something they should look at, they will.

Here are a couple of ways to pick the initial reviewer(s):

- Use `svn blame` and the commit log to find names of people who have recently modified the same area of code that you are modifying.
- Look in `CODE_OWNERS.TXT` to see who might be responsible for that area.
- If you've discussed the change on a dev list, the people who participated might be appropriate reviewers.

Even if you think the code owner is the busiest person in the world, it's still okay to put them as a reviewer. Being the code owner means they have accepted responsibility for making sure the review happens.

2.15.4 Reviewing code with Phabricator

Phabricator allows you to add inline comments as well as overall comments to a revision. To add an inline comment, select the lines of code you want to comment on by clicking and dragging the line numbers in the diff pane. When you have added all your comments, scroll to the bottom of the page and click the Submit button.

You can add overall comments in the text box at the bottom of the page. When you're done, click the Submit button.

Phabricator has many useful features, for example allowing you to select diffs between different versions of the patch as it was reviewed in the *Revision Update History*. Most features are self descriptive - explore, and if you have a question, drop by on #llvm in IRC to get help.

Note that as e-mail is the system of reference for code reviews, and some people prefer it over a web interface, we do not generate automated mail when a review changes state, for example by clicking "Accept Revision" in the web interface. Thus, please type LGTM into the comment box to accept a change from Phabricator.

2.15.5 Committing a change

Once a patch has been reviewed and approved on Phabricator it can then be committed to trunk. If you do not have commit access, someone has to commit the change for you (with attribution). It is sufficient to add a comment to the approved review indicating you cannot commit the patch yourself. If you have commit access, there are multiple workflows to commit the change. Whichever method you follow it is recommended that your commit message ends with the line:

```
Differential Revision: <URL>
```

where <URL> is the URL for the code review, starting with `http://reviews.llvm.org/`.

This allows people reading the version history to see the review for context. This also allows Phabricator to detect the commit, close the review, and add a link from the review to the commit.

Note that if you use the Arcanist tool the `Differential Revision` line will be added automatically. If you don't want to use Arcanist, you can add the `Differential Revision` line (as the last line) to the commit message yourself.

Using the Arcanist tool can simplify the process of committing reviewed code as it will retrieve reviewers, the `Differential Revision`, etc from the review and place it in the commit message. Several methods of using Arcanist to commit code are given below. If you do not wish to use Arcanist then simply commit the reviewed patch as you would normally.

Note that if you commit the change without using Arcanist and forget to add the `Differential Revision` line to your commit message then it is recommended that you close the review manually. In the web UI, under "Leap Into Action" put the SVN revision number in the Comment, set the Action to "Close Revision" and click Submit. Note the review must have been Accepted first.

Subversion and Arcanist

On a clean Subversion working copy run the following (where <Revision> is the Phabricator review number):

```
arc patch D<Revision>
arc commit --revision D<Revision>
```

The first command will take the latest version of the reviewed patch and apply it to the working copy. The second command will commit this revision to trunk.

git-svn and Arcanist

This presumes that the git repository has been configured as described in *For developers to work with git-svn*.

On a clean Git repository on an up to date `master` branch run the following (where <Revision> is the Phabricator review number):

```
arc patch D<Revision>
```

This will create a new branch called `arcpatch-D<Revision>` based on the current `master` and will create a commit corresponding to `D<Revision>` with a commit message derived from information in the Phabricator review.

Check you are happy with the commit message and amend it if necessary. Now switch to the `master` branch and add the new commit to it and commit it to trunk. This can be done by running the following:

```
git checkout master
git merge --ff-only arcpatch-D<Revision>
git svn dcommit
```

2.15.6 Abandoning a change

If you decide you should not commit the patch, you should explicitly abandon the review so that reviewers don't think it is still open. In the web UI, scroll to the bottom of the page where normally you would enter an overall comment. In the drop-down Action list, which defaults to "Comment," you should select "Abandon Revision" and then enter a comment explaining why. Click the Submit button to finish closing the review.

2.15.7 Status

Please let us know whether you like it and what could be improved! We're still working on setting up a bug tracker, but you can email klimek-at-google-dot-com and chandlerc-at-gmail-dot-com and CC the [llvm-dev](#) mailing list with questions until then. We also could use help implementing improvements. This sadly is really painful and hard because the Phabricator codebase is in PHP and not as testable as you might like. However, we've put exactly what we're deploying up on an [llvm-reviews GitHub project](#) where folks can hack on it and post pull requests. We're looking into what the right long-term hosting for this is, but note that it is a derivative of an existing open source project, and so not trivially a good fit for an official LLVM project.

2.16 LLVM Testing Infrastructure Guide

- *Overview*
- *Requirements*
- *LLVM testing infrastructure organization*
 - *Regression tests*
 - *test-suite*
 - *Debugging Information tests*
- *Quick start*
 - *Regression tests*
 - *Debugging Information tests*
- *Regression test structure*
 - *Writing new regression tests*
 - *Extra files*
 - *Fragile tests*
 - *Platform-Specific Tests*
 - *Constraining test execution*
 - *Substitutions*
 - *Options*
 - *Other Features*
- *test-suite Overview*
 - *test-suite Quickstart*

- *test-suite Makefiles*

2.16.1 LLVM test-suite Guide

- *Overview*
- *Test suite Structure*
- *Running the test suite via CMake*
- *Running the test suite via Makefiles (deprecated)*
 - *Configuring External Tests*
 - *Running different tests*
 - *Generating test output*
 - *Writing custom tests for the test suite*

Overview

This document describes the features of the Makefile-based LLVM test-suite as well as the cmake based replacement. This way of interacting with the test-suite is deprecated in favor of running the test-suite using LNT, but may continue to prove useful for some users. See the Testing Guide's *test-suite Quickstart* section for more information.

Test suite Structure

The `test-suite` module contains a number of programs that can be compiled with LLVM and executed. These programs are compiled using the native compiler and various LLVM backends. The output from the program compiled with the native compiler is assumed correct; the results from the other programs are compared to the native program output and pass if they match.

When executing tests, it is usually a good idea to start out with a subset of the available tests or programs. This makes test run times smaller at first and later on this is useful to investigate individual test failures. To run some test only on a subset of programs, simply change directory to the programs you want tested and run `gmake` there. Alternatively, you can run a different test using the `TEST` variable to change what tests or run on the selected programs (see below for more info).

In addition for testing correctness, the `test-suite` directory also performs timing tests of various LLVM optimizations. It also records compilation times for the compilers and the JIT. This information can be used to compare the effectiveness of LLVM's optimizations and code generation.

`test-suite` tests are divided into three types of tests: MultiSource, SingleSource, and External.

- `test-suite/SingleSource`

The SingleSource directory contains test programs that are only a single source file in size. These are usually small benchmark programs or small programs that calculate a particular value. Several such programs are grouped together in each directory.

- `test-suite/MultiSource`

The MultiSource directory contains subdirectories which contain entire programs with multiple source files. Large benchmarks and whole applications go here.

- test-suite/External

The External directory contains Makefiles for building code that is external to (i.e., not distributed with) LLVM. The most prominent members of this directory are the SPEC 95 and SPEC 2000 benchmark suites. The External directory does not contain these actual tests, but only the Makefiles that know how to properly compile these programs from somewhere else. The presence and location of these external programs is configured by the test-suite configure script.

Each tree is then subdivided into several categories, including applications, benchmarks, regression tests, code that is strange grammatically, etc. These organizations should be relatively self explanatory.

Some tests are known to fail. Some are bugs that we have not fixed yet; others are features that we haven't added yet (or may never add). In the regression tests, the result for such tests will be XFAIL (eXpected FAILure). In this way, you can tell the difference between an expected and unexpected failure.

The tests in the test suite have no such feature at this time. If the test passes, only warnings and other miscellaneous output will be generated. If a test fails, a large <program> FAILED message will be displayed. This will help you separate benign warnings from actual test failures.

Running the test suite via CMake

To run the test suite, you need to use the following steps:

1. The test suite uses the lit test runner to run the test-suite, you need to have lit installed first. Check out LLVM and install lit:

```
% svn co http://llvm.org/svn/llvm-project/llvm/trunk llvm
% cd llvm/utils/lit
% sudo python setup.py install # Or without sudo, install in virtual-env.
running install
running bdist_egg
running egg_info
writing lit.egg-info/PKG-INFO
...
% lit --version
lit 0.5.0dev
```

2. Check out the test-suite module with:

```
% svn co http://llvm.org/svn/llvm-project/test-suite/trunk test-suite
```

3. Use CMake to configure the test suite in a new directory. You cannot build the test suite in the source tree.

```
% mkdir test-suite-build
% cd test-suite-build
% cmake ../test-suite
```

4. Build the benchmarks, using the makefiles CMake generated.

```
% make
Scanning dependencies of target timeit-target
[ 0%] Building C object tools/CMakeFiles/timeit-target.dir/timeit.c.o
[ 0%] Linking C executable timeit-target
[ 0%] Built target timeit-target
Scanning dependencies of target fpcmp-host
[ 0%] [TEST_SUITE_HOST_CC] Building host executable fpcmp
[ 0%] Built target fpcmp-host
Scanning dependencies of target timeit-host
```

```
[ 0%] [TEST_SUITE_HOST_CC] Building host executable timeit
[ 0%] Built target timeit-host
```

1. Run the tests with lit:

```
% lit -v -j 1 . -o results.json
-- Testing: 474 tests, 1 threads --
PASS: test-suite :: MultiSource/Applications/ALAC/decode/alaconvert-decode.test (1_
↳ of 474)
***** TEST 'test-suite :: MultiSource/Applications/ALAC/decode/alaconvert-
↳ decode.test' RESULTS *****
compile_time: 0.2192
exec_time: 0.0462
hash: "59620e187c6ac38b36382685ccd2b63b"
size: 83348
*****
PASS: test-suite :: MultiSource/Applications/ALAC/encode/alaconvert-encode.test (2_
↳ of 474)
```

Running the test suite via Makefiles (deprecated)

First, all tests are executed within the LLVM object directory tree. They *are not* executed inside of the LLVM source tree. This is because the test suite creates temporary files during execution.

To run the test suite, you need to use the following steps:

1. cd into the `llvm/projects` directory in your source tree.
2. Check out the `test-suite` module with:

```
% svn co http://llvm.org/svn/llvm-project/test-suite/trunk test-suite
```

This will get the test suite into `llvm/projects/test-suite`.

3. Configure and build `llvm`.
4. Configure and build `llvm-gcc`.
5. Install `llvm-gcc` somewhere.
6. *Re-configure* `llvm` from the top level of each build tree (LLVM object directory tree) in which you want to run the test suite, just as you do before building LLVM.

During the *re-configuration*, you must either: (1) have `llvm-gcc` you just built in your path, or (2) specify the directory where your just-built `llvm-gcc` is installed using `--with-llvmgccdir=$LLVM_GCC_DIR`.

You must also tell the configure machinery that the test suite is available so it can be configured for your build tree:

```
% cd $LLVM_OBJ_ROOT ; $LLVM_SRC_ROOT/configure [--with-llvmgccdir=$LLVM_GCC_DIR]
```

[Remember that `$LLVM_GCC_DIR` is the directory where you *installed* `llvm-gcc`, not its `src` or `obj` directory.]

7. You can now run the test suite from your build tree as follows:

```
% cd $LLVM_OBJ_ROOT/projects/test-suite
% make
```

Note that the second and third steps only need to be done once. After you have the suite checked out and configured, you don't need to do it again (unless the test code or configure script changes).

Configuring External Tests

In order to run the External tests in the `test-suite` module, you must specify `-with-externals`. This must be done during the *re-configuration* step (see above), and the `llvm` re-configuration must recognize the previously-built `llvm-gcc`. If any of these is missing or neglected, the External tests won't work.

- `-with-externals`
- `-with-externals=<directory>`

This tells LLVM where to find any external tests. They are expected to be in specifically named subdirectories of `<directory>`. If `directory` is left unspecified, `configure` uses the default value `/home/vadve/shared/benchmarks/speccpu2000/benchspec`. Subdirectory names known to LLVM include:

- `spec95`
- `speccpu2000`
- `speccpu2006`
- `povray31`

Others are added from time to time, and can be determined from `configure`.

Running different tests

In addition to the regular “whole program” tests, the `test-suite` module also provides a mechanism for compiling the programs in different ways. If the variable `TEST` is defined on the `gmake` command line, the test system will include a Makefile named `TEST.<value of TEST variable>.Makefile`. This Makefile can modify build rules to yield different results.

For example, the LLVM nightly tester uses `TEST.nightly.Makefile` to create the nightly test reports. To run the nightly tests, run `gmake TEST=nightly`.

There are several `TEST` Makefiles available in the tree. Some of them are designed for internal LLVM research and will not work outside of the LLVM research group. They may still be valuable, however, as a guide to writing your own `TEST` Makefile for any optimization or analysis passes that you develop with LLVM.

Generating test output

There are a number of ways to run the tests and generate output. The most simple one is simply running `gmake` with no arguments. This will compile and run all programs in the tree using a number of different methods and compare results. Any failures are reported in the output, but are likely drowned in the other output. Passes are not reported explicitly.

Somewhat better is running `gmake TEST=sometest test`, which runs the specified test and usually adds per-program summaries to the output (depending on which `sometest` you use). For example, the `nightly` test explicitly outputs `TEST-PASS` or `TEST-FAIL` for every test after each program. Though these lines are still drowned in the output, it's easy to grep the output logs in the Output directories.

Even better are the `report` and `report.format` targets (where `format` is one of `html`, `csv`, `text` or `graphs`). The exact contents of the report are dependent on which `TEST` you are running, but the text results are always shown at the end of the run and the results are always stored in the `report.<type>.format` file (when running with `TEST=<type>`). The `report` also generate a file called `report.<type>.raw.out` containing the output of the entire test run.

Writing custom tests for the test suite

Assuming you can run the test suite, (e.g. “`gmake TEST=nightly report`” should work), it is really easy to run optimizations or code generator components against every program in the tree, collecting statistics or running custom checks for correctness. At base, this is how the nightly tester works, it’s just one example of a general framework.

Lets say that you have an LLVM optimization pass, and you want to see how many times it triggers. First thing you should do is add an LLVM [statistic](#) to your pass, which will tally counts of things you care about.

Following this, you can set up a test and a report that collects these and formats them for easy viewing. This consists of two files, a “`test-suite/TEST.XXX.Makefile`” fragment (where XXX is the name of your test) and a “`test-suite/TEST.XXX.report`” file that indicates how to format the output into a table. There are many example reports of various levels of sophistication included with the test suite, and the framework is very general.

If you are interested in testing an optimization pass, check out the “libcalls” test as an example. It can be run like this:

```
% cd llvm/projects/test-suite/MultiSource/Benchmarks # or some other level
% make TEST=libcalls report
```

This will do a bunch of stuff, then eventually print a table like this:

Name	total	#exit
...		
FreeBench/analyzer/analyzer	51	6
FreeBench/fourinarow/fourinarow	1	1
FreeBench/neural/neural	19	9
FreeBench/pifft/pifft	5	3
MallocBench/cfrac/cfrac	1	*
MallocBench/espresso/espresso	52	12
MallocBench/gs/gs	4	*
Prolangs-C/TimberWolfMC/timberwolfmc	302	*
Prolangs-C/agrep/agrep	33	12
Prolangs-C/allroots/allroots	*	*
Prolangs-C/assembler/assembler	47	*
Prolangs-C/bison/mybison	74	*
...		

This basically is grepping the `-stats` output and displaying it in a table. You can also use the “`TEST=libcalls report.html`” target to get the table in HTML form, similarly for `report.csv` and `report.tex`.

The source for this is in `test-suite/TEST.libcalls.*`. The format is pretty simple: the Makefile indicates how to run the test (in this case, “`opt -simplify-libcalls -stats`”), and the report contains one line for each column of the output. The first value is the header for the column and the second is the regex to grep the output of the command for. There are lots of example reports that can do fancy stuff.

2.16.2 Overview

This document is the reference manual for the LLVM testing infrastructure. It documents the structure of the LLVM testing infrastructure, the tools needed to use it, and how to add and run tests.

2.16.3 Requirements

In order to use the LLVM testing infrastructure, you will need all of the software required to build LLVM, as well as [Python 2.7](#) or later.

If you intend to run the *test-suite*, you will also need a development version of zlib (zlib1g-dev is known to work on several Linux distributions).

2.16.4 LLVM testing infrastructure organization

The LLVM testing infrastructure contains two major categories of tests: regression tests and whole programs. The regression tests are contained inside the LLVM repository itself under `llvm/test` and are expected to always pass – they should be run before every commit.

The whole programs tests are referred to as the “LLVM test suite” (or “test-suite”) and are in the `test-suite` module in subversion. For historical reasons, these tests are also referred to as the “nightly tests” in places, which is less ambiguous than “test-suite” and remains in use although we run them much more often than nightly.

Regression tests

The regression tests are small pieces of code that test a specific feature of LLVM or trigger a specific bug in LLVM. The language they are written in depends on the part of LLVM being tested. These tests are driven by the *Lit* testing tool (which is part of LLVM), and are located in the `llvm/test` directory.

Typically when a bug is found in LLVM, a regression test containing just enough code to reproduce the problem should be written and placed somewhere underneath this directory. For example, it can be a small piece of LLVM IR distilled from an actual application or benchmark.

`test-suite`

The test suite contains whole programs, which are pieces of code which can be compiled and linked into a stand-alone program that can be executed. These programs are generally written in high level languages such as C or C++.

These programs are compiled using a user specified compiler and set of flags, and then executed to capture the program output and timing information. The output of these programs is compared to a reference output to ensure that the program is being compiled correctly.

In addition to compiling and executing programs, whole program tests serve as a way of benchmarking LLVM performance, both in terms of the efficiency of the programs generated as well as the speed with which LLVM compiles, optimizes, and generates code.

The test-suite is located in the `test-suite` Subversion module.

Debugging Information tests

The test suite contains tests to check quality of debugging information. The test are written in C based languages or in LLVM assembly language.

These tests are compiled and run under a debugger. The debugger output is checked to validate of debugging information. See README.txt in the test suite for more information . This test suite is located in the `debuginfo-tests` Subversion module.

2.16.5 Quick start

The tests are located in two separate Subversion modules. The regressions tests are in the main “llvm” module under the directory `llvm/test` (so you get these tests for free with the main LLVM tree). Use `make check-all` to run the regression tests after building LLVM.

The more comprehensive test suite that includes whole programs in C and C++ is in the `test-suite` module. See [test-suite Quickstart](#) for more information on running these tests.

Regression tests

To run all of the LLVM regression tests use the `check-llvm` target:

```
% make check-llvm
```

If you have [Clang](#) checked out and built, you can run the LLVM and Clang tests simultaneously using:

```
% make check-all
```

To run the tests with Valgrind (Memcheck by default), use the `LIT_ARGS` make variable to pass the required options to `lit`. For example, you can use:

```
% make check LIT_ARGS="-v --vg --vg-leak"
```

to enable testing with valgrind and with leak checking enabled.

To run individual tests or subsets of tests, you can use the `llvm-lit` script which is built as part of LLVM. For example, to run the `Integer/BitPacked.ll` test by itself you can run:

```
% llvm-lit ~/llvm/test/Integer/BitPacked.ll
```

or to run all of the ARM CodeGen tests:

```
% llvm-lit ~/llvm/test/CodeGen/ARM
```

For more information on using the `lit` tool, see `llvm-lit --help` or the [lit man page](#).

Debugging Information tests

To run debugging information tests simply checkout the tests inside `clang/test` directory.

```
% cd clang/test
% svn co http://llvm.org/svn/llvm-project/debuginfo-tests/trunk debuginfo-tests
```

These tests are already set up to run as part of clang regression tests.

2.16.6 Regression test structure

The LLVM regression tests are driven by `lit` and are located in the `llvm/test` directory.

This directory contains a large array of small tests that exercise various features of LLVM and to ensure that regressions do not occur. The directory is broken into several sub-directories, each focused on a particular area of LLVM.

Writing new regression tests

The regression test structure is very simple, but does require some information to be set. This information is gathered via `configure` and is written to a file, `test/lit.site.cfg` in the build directory. The `llvm/test` Makefile does this work for you.

In order for the regression tests to work, each directory of tests must have a `lit.local.cfg` file. `lit` looks for this file to determine how to run the tests. This file is just Python code and thus is very flexible, but we’ve standardized it for the LLVM regression tests. If you’re adding a directory of tests, just copy `lit.local.cfg` from another directory to get running. The standard `lit.local.cfg` simply specifies which files to look in for tests. Any directory that contains only directories does not need the `lit.local.cfg` file. Read the [Lit documentation](#) for more information.

Each test file must contain lines starting with “RUN:” that tell `lit` how to run it. If there are no RUN lines, `lit` will issue an error while running a test.

RUN lines are specified in the comments of the test program using the keyword `RUN` followed by a colon, and lastly the command (pipeline) to execute. Together, these lines form the “script” that `lit` executes to run the test case. The syntax of the RUN lines is similar to a shell’s syntax for pipelines including I/O redirection and variable substitution. However, even though these lines may *look* like a shell script, they are not. RUN lines are interpreted by `lit`. Consequently, the syntax differs from shell in a few ways. You can specify as many RUN lines as needed.

`lit` performs substitution on each RUN line to replace LLVM tool names with the full paths to the executable built for each tool (in `$(LLVM_OBJ_ROOT)/$(BuildMode)/bin`). This ensures that `lit` does not invoke any stray LLVM tools in the user’s path during testing.

Each RUN line is executed on its own, distinct from other lines unless its last character is `\`. This continuation character causes the RUN line to be concatenated with the next one. In this way you can build up long pipelines of commands without making huge line lengths. The lines ending in `\` are concatenated until a RUN line that doesn’t end in `\` is found. This concatenated set of RUN lines then constitutes one execution. `lit` will substitute variables and arrange for the pipeline to be executed. If any process in the pipeline fails, the entire line (and test case) fails too.

Below is an example of legal RUN lines in a `.ll` file:

```
; RUN: llvm-as < %s | llvm-dis > %t1
; RUN: llvm-dis < %s.bc-13 > %t2
; RUN: diff %t1 %t2
```

As with a Unix shell, the RUN lines permit pipelines and I/O redirection to be used.

There are some quoting rules that you must pay attention to when writing your RUN lines. In general nothing needs to be quoted. `lit` won’t strip off any quote characters so they will get passed to the invoked program. To avoid this use curly braces to tell `lit` that it should treat everything enclosed as one value.

In general, you should strive to keep your RUN lines as simple as possible, using them only to run tools that generate textual output you can then examine. The recommended way to examine output to figure out if the test passes is using the [FileCheck tool](#). *[The usage of `grep` in RUN lines is deprecated - please do not send or commit patches that use it.]*

Put related tests into a single file rather than having a separate file per test. Check if there are files already covering your feature and consider adding your code there instead of creating a new file.

Extra files

If your test requires extra files besides the file containing the `RUN:` lines, the idiomatic place to put them is in a subdirectory `Inputs`. You can then refer to the extra files as `%S/Inputs/foo.bar`.

For example, consider `test/Linker/ident.ll`. The directory structure is as follows:

```
test/
  Linker/
    ident.ll
    Inputs/
      ident.a.ll
      ident.b.ll
```

For convenience, these are the contents:

```

;;;; ident.ll:

; RUN: llvm-link %S/Inputs/ident.a.ll %S/Inputs/ident.b.ll -S | FileCheck %s

; Verify that multiple input llvm.ident metadata are linked together.

; CHECK-DAG: !llvm.ident = !{!0, !1, !2}
; CHECK-DAG: "Compiler V1"
; CHECK-DAG: "Compiler V2"
; CHECK-DAG: "Compiler V3"

;;;; Inputs/ident.a.ll:

!llvm.ident = !{!0, !1}
!0 = metadata !{metadata !"Compiler V1"}
!1 = metadata !{metadata !"Compiler V2"}

;;;; Inputs/ident.b.ll:

!llvm.ident = !{!0}
!0 = metadata !{metadata !"Compiler V3"}

```

For symmetry reasons, `ident.ll` is just a dummy file that doesn't actually participate in the test besides holding the `RUN:` lines.

Note: Some existing tests use `RUN: true` in extra files instead of just putting the extra files in an `Inputs/` directory. This pattern is deprecated.

Fragile tests

It is easy to write a fragile test that would fail spuriously if the tool being tested outputs a full path to the input file. For example, `opt` by default outputs a `ModuleID`:

```

$ cat example.ll
define i32 @main() nounwind {
    ret i32 0
}

$ opt -S /path/to/example.ll
; ModuleID = '/path/to/example.ll'

define i32 @main() nounwind {
    ret i32 0
}

```

`ModuleID` can unexpectedly match against `CHECK` lines. For example:

```

; RUN: opt -S %s | FileCheck

define i32 @main() nounwind {
    ; CHECK-NOT: load
    ret i32 0
}

```

This test will fail if placed into a `download` directory.

To make your tests robust, always use `opt ... < %s` in the RUN line. `opt` does not output a `ModuleID` when input comes from stdin.

Platform-Specific Tests

Whenever adding tests that require the knowledge of a specific platform, either related to code generated, specific output or back-end features, you must make sure to isolate the features, so that buildbots that run on different architectures (and don't even compile all back-ends), don't fail.

The first problem is to check for target-specific output, for example sizes of structures, paths and architecture names, for example:

- Tests containing Windows paths will fail on Linux and vice-versa.
- Tests that check for `x86_64` somewhere in the text will fail anywhere else.
- Tests where the debug information calculates the size of types and structures.

Also, if the test rely on any behaviour that is coded in any back-end, it must go in its own directory. So, for instance, code generator tests for ARM go into `test/CodeGen/ARM` and so on. Those directories contain a special `lit` configuration file that ensure all tests in that directory will only run if a specific back-end is compiled and available.

For instance, on `test/CodeGen/ARM`, the `lit.local.cfg` is:

```
config.suffixes = ['.ll', '.c', '.cpp', '.test']
if not 'ARM' in config.root.targets:
    config.unsupported = True
```

Other platform-specific tests are those that depend on a specific feature of a specific sub-architecture, for example only to Intel chips that support AVX2.

For instance, `test/CodeGen/X86/psubus.ll` tests three sub-architecture variants:

```
; RUN: llc -mcpu=core2 < %s | FileCheck %s -check-prefix=SSE2
; RUN: llc -mcpu=corei7-avx < %s | FileCheck %s -check-prefix=AVX1
; RUN: llc -mcpu=core-avx2 < %s | FileCheck %s -check-prefix=AVX2
```

And the checks are different:

```
; SSE2: @test1
; SSE2: psubusw LCPI0_0(%rip), %xmm0
; AVX1: @test1
; AVX1: vpsubusw LCPI0_0(%rip), %xmm0, %xmm0
; AVX2: @test1
; AVX2: vpsubusw LCPI0_0(%rip), %xmm0, %xmm0
```

So, if you're testing for a behaviour that you know is platform-specific or depends on special features of sub-architectures, you must add the specific triple, test with the specific `FileCheck` and put it into the specific directory that will filter out all other architectures.

Constraining test execution

Some tests can be run only in specific configurations, such as with debug builds or on particular platforms. Use `REQUIRES` and `UNSUPPORTED` to control when the test is enabled.

Some tests are expected to fail. For example, there may be a known bug that the test detect. Use `XFAIL` to mark a test as an expected failure. An `XFAIL` test will be successful if its execution fails, and will be a failure if its execution succeeds.

```
; This test will be only enabled in the build with asserts.
; REQUIRES: asserts
; This test is disabled on Linux.
; UNSUPPORTED: -linux-
; This test is expected to fail on PowerPC.
; XFAIL: powerpc
```

REQUIRES and UNSUPPORTED and XFAIL all accept a comma-separated list of boolean expressions. The values in each expression may be:

- Features added to `config.available_features` by configuration files such as `lit.cfg`.
- Substrings of the target triple (UNSUPPORTED and XFAIL only).

REQUIRES enables the test if all expressions are true.

UNSUPPORTED disables the test if any expression is true.

XFAIL expects the test to fail if any expression is true.

As a special case, XFAIL: * is expected to fail everywhere.

```
; This test is disabled on Windows,
; and is disabled on Linux, except for Android Linux.
; UNSUPPORTED: windows, linux && !android
; This test is expected to fail on both PowerPC and ARM.
; XFAIL: powerpc || arm
```

Substitutions

Besides replacing LLVM tool names the following substitutions are performed in RUN lines:

%% Replaced by a single `%`. This allows escaping other substitutions.

%s File path to the test case's source. This is suitable for passing on the command line as the input to an LLVM tool.

Example: `/home/user/llvm/test/MC/ELF/foo_test.s`

%S Directory path to the test case's source.

Example: `/home/user/llvm/test/MC/ELF`

%t File path to a temporary file name that could be used for this test case. The file name won't conflict with other test cases. You can append to it if you need multiple temporaries. This is useful as the destination of some redirected output.

Example: `/home/user/llvm.build/test/MC/ELF/Output/foo_test.s.tmp`

%T Directory of `%t`.

Example: `/home/user/llvm.build/test/MC/ELF/Output`

%{pathsep}

Expands to the path separator, i.e. `:` (or `;` on Windows).

%/s, %/S, %/t, %/T:

Act like the corresponding substitution above but replace any `\` character with a `/`. This is useful to normalize path separators.

Example: %s: C:\Desktop Files/foo_test.s.tmp

Example: %/s: C:/Desktop Files/foo_test.s.tmp

%:s, %:S, %:t, %:T:

Act like the corresponding substitution above but remove colons at the beginning of Windows paths. This is useful to allow concatenation of absolute paths on Windows to produce a legal path.

Example: %s: C:\Desktop Files\foo_test.s.tmp

Example: %:s: C\Desktop Files\foo_test.s.tmp

LLVM-specific substitutions:

%shlibext The suffix for the host platforms shared library files. This includes the period as the first character.

Example: .so (Linux), .dylib (OS X), .dll (Windows)

%exeext The suffix for the host platforms executable files. This includes the period as the first character.

Example: .exe (Windows), empty on Linux.

%(line),%(line+<number>),%(line-<number>) The number of the line where this substitution is used, with an optional integer offset. This can be used in tests with multiple RUN lines, which reference test file's line numbers.

Clang-specific substitutions:

%clang Invokes the Clang driver.

%clang_cpp Invokes the Clang driver for C++.

%clang_cl Invokes the CL-compatible Clang driver.

%clangxx Invokes the G++-compatible Clang driver.

%clang_cc1 Invokes the Clang frontend.

%itanium_abi_triple,%ms_abi_triple These substitutions can be used to get the current target triple adjusted to the desired ABI. For example, if the test suite is running with the i686-pc-win32 target, %itanium_abi_triple will expand to i686-pc-mingw32. This allows a test to run with a specific ABI without constraining it to a specific triple.

To add more substitutions, look at test/lit.cfg or lit.local.cfg.

Options

The llvm lit configuration allows to customize some things with user options:

llc,opt,... Substitute the respective llvm tool name with a custom command line. This allows to specify custom paths and default arguments for these tools. Example:

% llvm-lit "-Dlhc=llc -verify-machineinstrs"

run_long_tests Enable the execution of long running tests.

llvm_site_config Load the specified lit configuration instead of the default one.

Other Features

To make RUN line writing easier, there are several helper programs. These helpers are in the PATH when running tests, so you can just call them using their name. For example:

not This program runs its arguments and then inverts the result code from it. Zero result codes become 1. Non-zero result codes become 0.

To make the output more useful, **lit** will scan the lines of the test case for ones that contain a pattern that matches `PR[0-9]+`. This is the syntax for specifying a PR (Problem Report) number that is related to the test case. The number after “PR” specifies the LLVM bugzilla number. When a PR number is specified, it will be used in the pass/fail reporting. This is useful to quickly get some context when a test fails.

Finally, any line that contains “END.” will cause the special interpretation of lines to terminate. This is generally done right after the last RUN: line. This has two side effects:

1. it prevents special interpretation of lines that are part of the test program, not the instructions to the test case, and
2. it speeds things up for really big test cases by avoiding interpretation of the remainder of the file.

2.16.7 test-suite Overview

The `test-suite` module contains a number of programs that can be compiled and executed. The `test-suite` includes reference outputs for all of the programs, so that the output of the executed program can be checked for correctness.

`test-suite` tests are divided into three types of tests: `MultiSource`, `SingleSource`, and `External`.

- `test-suite/SingleSource`

The `SingleSource` directory contains test programs that are only a single source file in size. These are usually small benchmark programs or small programs that calculate a particular value. Several such programs are grouped together in each directory.

- `test-suite/MultiSource`

The `MultiSource` directory contains subdirectories which contain entire programs with multiple source files. Large benchmarks and whole applications go here.

- `test-suite/External`

The `External` directory contains Makefiles for building code that is external to (i.e., not distributed with) LLVM. The most prominent members of this directory are the SPEC 95 and SPEC 2000 benchmark suites. The `External` directory does not contain these actual tests, but only the Makefiles that know how to properly compile these programs from somewhere else. When using LNT, use the `--test-externals` option to include these tests in the results.

test-suite Quickstart

The modern way of running the `test-suite` is focused on testing and benchmarking complete compilers using the [LNT](#) testing infrastructure.

For more information on using LNT to execute the `test-suite`, please see the [LNT Quickstart](#) documentation.

test-suite Makefiles

Historically, the `test-suite` was executed using a complicated setup of Makefiles. The LNT based approach above is recommended for most users, but there are some testing scenarios which are not supported by the LNT approach. In addition, LNT currently uses the Makefile setup under the covers and so developers who are interested in how LNT works under the hood may want to understand the Makefile based setup.

For more information on the `test-suite` Makefile setup, please see the [Test Suite Makefile Guide](#).

2.17 LLVM Tutorial: Table of Contents

2.17.1 Kaleidoscope: Implementing a Language with LLVM

Kaleidoscope: Tutorial Introduction and the Lexer

- *Tutorial Introduction*
- *The Basic Language*
- *The Lexer*

Tutorial Introduction

Welcome to the “Implementing a language with LLVM” tutorial. This tutorial runs through the implementation of a simple language, showing how fun and easy it can be. This tutorial will get you up and started as well as help to build a framework you can extend to other languages. The code in this tutorial can also be used as a playground to hack on other LLVM specific things.

The goal of this tutorial is to progressively unveil our language, describing how it is built up over time. This will let us cover a fairly broad range of language design and LLVM-specific usage issues, showing and explaining the code for it all along the way, without overwhelming you with tons of details up front.

It is useful to point out ahead of time that this tutorial is really about teaching compiler techniques and LLVM specifically, *not* about teaching modern and sane software engineering principles. In practice, this means that we’ll take a number of shortcuts to simplify the exposition. For example, the code uses global variables all over the place, doesn’t use nice design patterns like [visitors](#), etc... but it is very simple. If you dig in and use the code as a basis for future projects, fixing these deficiencies shouldn’t be hard.

I’ve tried to put this tutorial together in a way that makes chapters easy to skip over if you are already familiar with or are uninterested in the various pieces. The structure of the tutorial is:

- **Chapter #1:** Introduction to the Kaleidoscope language, and the definition of its Lexer - This shows where we are going and the basic functionality that we want it to do. In order to make this tutorial maximally understandable and hackable, we choose to implement everything in C++ instead of using lexer and parser generators. LLVM obviously works just fine with such tools, feel free to use one if you prefer.
- **Chapter #2:** Implementing a Parser and AST - With the lexer in place, we can talk about parsing techniques and basic AST construction. This tutorial describes recursive descent parsing and operator precedence parsing. Nothing in Chapters 1 or 2 is LLVM-specific, the code doesn’t even link in LLVM at this point. :)
- **Chapter #3:** Code generation to LLVM IR - With the AST ready, we can show off how easy generation of LLVM IR really is.
- **Chapter #4:** Adding JIT and Optimizer Support - Because a lot of people are interested in using LLVM as a JIT, we’ll dive right into it and show you the 3 lines it takes to add JIT support. LLVM is also useful in many other ways, but this is one simple and “sexy” way to show off its power. :)
- **Chapter #5:** Extending the Language: Control Flow - With the language up and running, we show how to extend it with control flow operations (if/then/else and a ‘for’ loop). This gives us a chance to talk about simple SSA construction and control flow.
- **Chapter #6:** Extending the Language: User-defined Operators - This is a silly but fun chapter that talks about extending the language to let the user program define their own arbitrary unary and binary operators (with assignable precedence!). This lets us build a significant piece of the “language” as library routines.

- [Chapter #7](#): Extending the Language: Mutable Variables - This chapter talks about adding user-defined local variables along with an assignment operator. The interesting part about this is how easy and trivial it is to construct SSA form in LLVM: no, LLVM does *not* require your front-end to construct SSA form!
- [Chapter #8](#): Compiling to Object Files - This chapter explains how to take LLVM IR and compile it down to object files.
- [Chapter #9](#): Extending the Language: Debug Information - Having built a decent little programming language with control flow, functions and mutable variables, we consider what it takes to add debug information to standalone executables. This debug information will allow you to set breakpoints in Kaleidoscope functions, print out argument variables, and call functions - all from within the debugger!
- [Chapter #10](#): Conclusion and other useful LLVM tidbits - This chapter wraps up the series by talking about potential ways to extend the language, but also includes a bunch of pointers to info about “special topics” like adding garbage collection support, exceptions, debugging, support for “spaghetti stacks”, and a bunch of other tips and tricks.

By the end of the tutorial, we’ll have written a bit less than 1000 lines of non-comment, non-blank, lines of code. With this small amount of code, we’ll have built up a very reasonable compiler for a non-trivial language including a hand-written lexer, parser, AST, as well as code generation support with a JIT compiler. While other systems may have interesting “hello world” tutorials, I think the breadth of this tutorial is a great testament to the strengths of LLVM and why you should consider it if you’re interested in language or compiler design.

A note about this tutorial: we expect you to extend the language and play with it on your own. Take the code and go crazy hacking away at it, compilers don’t need to be scary creatures - it can be a lot of fun to play with languages!

The Basic Language

This tutorial will be illustrated with a toy language that we’ll call “[Kaleidoscope](#)” (derived from “meaning beautiful, form, and view”). Kaleidoscope is a procedural language that allows you to define functions, use conditionals, math, etc. Over the course of the tutorial, we’ll extend Kaleidoscope to support the if/then/else construct, a for loop, user defined operators, JIT compilation with a simple command line interface, etc.

Because we want to keep things simple, the only datatype in Kaleidoscope is a 64-bit floating point type (aka ‘double’ in C parlance). As such, all values are implicitly double precision and the language doesn’t require type declarations. This gives the language a very nice and simple syntax. For example, the following simple example computes [Fibonacci numbers](#):

```
# Compute the x'th fibonacci number.
def fib(x)
  if x < 3 then
    1
  else
    fib(x-1)+fib(x-2)

# This expression will compute the 40th number.
fib(40)
```

We also allow Kaleidoscope to call into standard library functions (the LLVM JIT makes this completely trivial). This means that you can use the ‘extern’ keyword to define a function before you use it (this is also useful for mutually recursive functions). For example:

```
extern sin(arg);
extern cos(arg);
extern atan2(arg1 arg2);

atan2(sin(.4), cos(42))
```

A more interesting example is included in Chapter 6 where we write a little Kaleidoscope application that displays a [Mandelbrot Set](#) at various levels of magnification.

Lets dive into the implementation of this language!

The Lexer

When it comes to implementing a language, the first thing needed is the ability to process a text file and recognize what it says. The traditional way to do this is to use a “lexer” (aka ‘scanner’) to break the input up into “tokens”. Each token returned by the lexer includes a token code and potentially some metadata (e.g. the numeric value of a number). First, we define the possibilities:

```
// The lexer returns tokens [0-255] if it is an unknown character, otherwise one
// of these for known things.
enum Token {
    tok_eof = -1,

    // commands
    tok_def = -2,
    tok_extern = -3,

    // primary
    tok_identifier = -4,
    tok_number = -5,
};

static std::string IdentifierStr; // Filled in if tok_identifier
static double NumVal;           // Filled in if tok_number
```

Each token returned by our lexer will either be one of the Token enum values or it will be an ‘unknown’ character like ‘+’, which is returned as its ASCII value. If the current token is an identifier, the IdentifierStr global variable holds the name of the identifier. If the current token is a numeric literal (like 1.0), NumVal holds its value. Note that we use global variables for simplicity, this is not the best choice for a real language implementation :).

The actual implementation of the lexer is a single function named `gettok`. The `gettok` function is called to return the next token from standard input. Its definition starts as:

```
/// gettok - Return the next token from standard input.
static int gettok() {
    static int LastChar = ' ';

    // Skip any whitespace.
    while (isspace(LastChar))
        LastChar = getchar();
```

`gettok` works by calling the C `getchar()` function to read characters one at a time from standard input. It eats them as it recognizes them and stores the last character read, but not processed, in `LastChar`. The first thing that it has to do is ignore whitespace between tokens. This is accomplished with the loop above.

The next thing `gettok` needs to do is recognize identifiers and specific keywords like “def”. Kaleidoscope does this with this simple loop:

```
if (isalpha(LastChar)) { // identifier: [a-zA-Z][a-zA-Z0-9]*
    IdentifierStr = LastChar;
    while (isalnum((LastChar = getchar())))
        IdentifierStr += LastChar;
```

```

if (IdentifierStr == "def")
    return tok_def;
if (IdentifierStr == "extern")
    return tok_extern;
return tok_identifier;
}

```

Note that this code sets the ‘IdentifierStr’ global whenever it lexes an identifier. Also, since language keywords are matched by the same loop, we handle them here inline. Numeric values are similar:

```

if (isdigit>LastChar) || LastChar == '.') {    // Number: [0-9.]+
    std::string NumStr;
    do {
        NumStr += LastChar;
        LastChar = getchar();
    } while (isdigit>LastChar) || LastChar == '.');

    NumVal = strtod(NumStr.c_str(), 0);
    return tok_number;
}

```

This is all pretty straight-forward code for processing input. When reading a numeric value from input, we use the C `strtod` function to convert it to a numeric value that we store in `NumVal`. Note that this isn’t doing sufficient error checking: it will incorrectly read “1.23.45.67” and handle it as if you typed in “1.23”. Feel free to extend it :). Next we handle comments:

```

if (LastChar == '#') {
    // Comment until end of line.
    do
        LastChar = getchar();
    while (LastChar != EOF && LastChar != '\n' && LastChar != '\r');

    if (LastChar != EOF)
        return gettok();
}

```

We handle comments by skipping to the end of the line and then return the next token. Finally, if the input doesn’t match one of the above cases, it is either an operator character like ‘+’ or the end of the file. These are handled with this code:

```

// Check for end of file.  Don't eat the EOF.
if (LastChar == EOF)
    return tok_eof;

// Otherwise, just return the character as its ascii value.
int ThisChar = LastChar;
LastChar = getchar();
return ThisChar;
}

```

With this, we have the complete lexer for the basic Kaleidoscope language (the [full code listing](#) for the Lexer is available in the [next chapter](#) of the tutorial). Next we’ll [build a simple parser that uses this](#) to build an Abstract Syntax Tree. When we have that, we’ll include a driver so that you can use the lexer and parser together.

Next: [Implementing a Parser and AST](#)

Kaleidoscope: Implementing a Parser and AST

- *Chapter 2 Introduction*
- *The Abstract Syntax Tree (AST)*
- *Parser Basics*
- *Basic Expression Parsing*
- *Binary Expression Parsing*
- *Parsing the Rest*
- *The Driver*
- *Conclusions*
- *Full Code Listing*

Chapter 2 Introduction

Welcome to Chapter 2 of the “[Implementing a language with LLVM](#)” tutorial. This chapter shows you how to use the lexer, built in [Chapter 1](#), to build a full [parser](#) for our Kaleidoscope language. Once we have a parser, we’ll define and build an [Abstract Syntax Tree](#) (AST).

The parser we will build uses a combination of [Recursive Descent Parsing](#) and [Operator-Precedence Parsing](#) to parse the Kaleidoscope language (the latter for binary expressions and the former for everything else). Before we get to parsing though, let’s talk about the output of the parser: the Abstract Syntax Tree.

The Abstract Syntax Tree (AST)

The AST for a program captures its behavior in such a way that it is easy for later stages of the compiler (e.g. code generation) to interpret. We basically want one object for each construct in the language, and the AST should closely model the language. In Kaleidoscope, we have expressions, a prototype, and a function object. We’ll start with expressions first:

```
/// ExprAST - Base class for all expression nodes.
class ExprAST {
public:
    virtual ~ExprAST() {}
};

/// NumberExprAST - Expression class for numeric literals like "1.0".
class NumberExprAST : public ExprAST {
    double Val;

public:
    NumberExprAST(double Val) : Val(Val) {}
};
```

The code above shows the definition of the base ExprAST class and one subclass which we use for numeric literals. The important thing to note about this code is that the NumberExprAST class captures the numeric value of the literal as an instance variable. This allows later phases of the compiler to know what the stored numeric value is.

Right now we only create the AST, so there are no useful accessor methods on them. It would be very easy to add a virtual method to pretty print the code, for example. Here are the other expression AST node definitions that we'll use in the basic form of the Kaleidoscope language:

```
/// VariableExprAST - Expression class for referencing a variable, like "a".
class VariableExprAST : public ExprAST {
    std::string Name;

public:
    VariableExprAST(const std::string &Name) : Name(Name) {}
};

/// BinaryExprAST - Expression class for a binary operator.
class BinaryExprAST : public ExprAST {
    char Op;
    std::unique_ptr<ExprAST> LHS, RHS;

public:
    BinaryExprAST(char op, std::unique_ptr<ExprAST> LHS,
                  std::unique_ptr<ExprAST> RHS)
        : Op(op), LHS(std::move(LHS)), RHS(std::move(RHS)) {}
};

/// CallExprAST - Expression class for function calls.
class CallExprAST : public ExprAST {
    std::string Callee;
    std::vector<std::unique_ptr<ExprAST>> Args;

public:
    CallExprAST(const std::string &Callee,
                std::vector<std::unique_ptr<ExprAST>> Args)
        : Callee(Callee), Args(std::move(Args)) {}
};
```

This is all (intentionally) rather straight-forward: variables capture the variable name, binary operators capture their opcode (e.g. '+'), and calls capture a function name as well as a list of any argument expressions. One thing that is nice about our AST is that it captures the language features without talking about the syntax of the language. Note that there is no discussion about precedence of binary operators, lexical structure, etc.

For our basic language, these are all of the expression nodes we'll define. Because it doesn't have conditional control flow, it isn't Turing-complete; we'll fix that in a later installment. The two things we need next are a way to talk about the interface to a function, and a way to talk about functions themselves:

```
/// PrototypeAST - This class represents the "prototype" for a function,
/// which captures its name, and its argument names (thus implicitly the number
/// of arguments the function takes).
class PrototypeAST {
    std::string Name;
    std::vector<std::string> Args;

public:
    PrototypeAST(const std::string &name, std::vector<std::string> Args)
        : Name(name), Args(std::move(Args)) {}

    const std::string &getName() const { return Name; }
};

/// FunctionAST - This class represents a function definition itself.
```

```
class FunctionAST {
    std::unique_ptr<PrototypeAST> Proto;
    std::unique_ptr<ExprAST> Body;

public:
    FunctionAST(std::unique_ptr<PrototypeAST> Proto,
                std::unique_ptr<ExprAST> Body)
        : Proto(std::move(Proto)), Body(std::move(Body)) {}
};
```

In Kaleidoscope, functions are typed with just a count of their arguments. Since all values are double precision floating point, the type of each argument doesn't need to be stored anywhere. In a more aggressive and realistic language, the "ExprAST" class would probably have a type field.

With this scaffolding, we can now talk about parsing expressions and function bodies in Kaleidoscope.

Parser Basics

Now that we have an AST to build, we need to define the parser code to build it. The idea here is that we want to parse something like "x+y" (which is returned as three tokens by the lexer) into an AST that could be generated with calls like this:

```
auto LHS = llvm::make_unique<VariableExprAST>("x");
auto RHS = llvm::make_unique<VariableExprAST>("y");
auto Result = std::make_unique<BinaryExprAST>('+', std::move(LHS),
                                                std::move(RHS));
```

In order to do this, we'll start by defining some basic helper routines:

```
/// CurTok/getNextToken - Provide a simple token buffer. CurTok is the current
/// token the parser is looking at. getNextToken reads another token from the
/// lexer and updates CurTok with its results.
static int CurTok;
static int getNextToken() {
    return CurTok = gettok();
}
```

This implements a simple token buffer around the lexer. This allows us to look one token ahead at what the lexer is returning. Every function in our parser will assume that CurTok is the current token that needs to be parsed.

```
/// LogError* - These are little helper functions for error handling.
std::unique_ptr<ExprAST> LogError(const char *Str) {
    fprintf(stderr, "LogError: %s\n", Str);
    return nullptr;
}
std::unique_ptr<PrototypeAST> LogErrorP(const char *Str) {
    LogError(Str);
    return nullptr;
}
```

The LogError routines are simple helper routines that our parser will use to handle errors. The error recovery in our parser will not be the best and is not particular user-friendly, but it will be enough for our tutorial. These routines make it easier to handle errors in routines that have various return types: they always return null.

With these basic helper functions, we can implement the first piece of our grammar: numeric literals.

Basic Expression Parsing

We start with numeric literals, because they are the simplest to process. For each production in our grammar, we'll define a function which parses that production. For numeric literals, we have:

```
/// numberexpr ::= number
static std::unique_ptr<ExprAST> ParseNumberExpr() {
    auto Result = llvm::make_unique<NumberExprAST>(NumVal);
    getNextToken(); // consume the number
    return std::move(Result);
}
```

This routine is very simple: it expects to be called when the current token is a `tok_number` token. It takes the current number value, creates a `NumberExprAST` node, advances the lexer to the next token, and finally returns.

There are some interesting aspects to this. The most important one is that this routine eats all of the tokens that correspond to the production and returns the lexer buffer with the next token (which is not part of the grammar production) ready to go. This is a fairly standard way to go for recursive descent parsers. For a better example, the parenthesis operator is defined like this:

```
/// parenexpr ::= '(' expression ')'
static std::unique_ptr<ExprAST> ParseParenExpr() {
    getNextToken(); // eat (.
    auto V = ParseExpression();
    if (!V)
        return nullptr;

    if (CurTok != ')')
        return LogError("expected ')'");
    getNextToken(); // eat ).
    return V;
}
```

This function illustrates a number of interesting things about the parser:

- 1) It shows how we use the `LogError` routines. When called, this function expects that the current token is a '(' token, but after parsing the subexpression, it is possible that there is no ')' waiting. For example, if the user types in "(4 x" instead of "(4)", the parser should emit an error. Because errors can occur, the parser needs a way to indicate that they happened: in our parser, we return null on an error.
- 2) Another interesting aspect of this function is that it uses recursion by calling `ParseExpression` (we will soon see that `ParseExpression` can call `ParseParenExpr`). This is powerful because it allows us to handle recursive grammars, and keeps each production very simple. Note that parentheses do not cause construction of AST nodes themselves. While we could do it this way, the most important role of parentheses are to guide the parser and provide grouping. Once the parser constructs the AST, parentheses are not needed.

The next simple production is for handling variable references and function calls:

```
/// identifierexpr
/// ::= identifier
/// ::= identifier '(' expression* ')'
static std::unique_ptr<ExprAST> ParseIdentifierExpr() {
    std::string IdName = IdentifierStr;

    getNextToken(); // eat identifier.

    if (CurTok != '(') // Simple variable ref.
        return llvm::make_unique<VariableExprAST>(IdName);
```

```
// Call.
getNextToken(); // eat (
std::vector<std::unique_ptr<ExprAST>> Args;
if (CurTok != ')') {
    while (1) {
        if (auto Arg = ParseExpression())
            Args.push_back(std::move(Arg));
        else
            return nullptr;

        if (CurTok == ')')
            break;

        if (CurTok != ',')
            return LogError("Expected ')' or ',' in argument list");
        getNextToken();
    }
}

// Eat the ')'.
getNextToken();

return llvm::make_unique<CallExprAST>(IdName, std::move(Args));
}
```

This routine follows the same style as the other routines. (It expects to be called if the current token is a `tok_identifier` token). It also has recursion and error handling. One interesting aspect of this is that it uses *look-ahead* to determine if the current identifier is a stand alone variable reference or if it is a function call expression. It handles this by checking to see if the token after the identifier is a `'('` token, constructing either a `VariableExprAST` or `CallExprAST` node as appropriate.

Now that we have all of our simple expression-parsing logic in place, we can define a helper function to wrap it together into one entry point. We call this class of expressions “primary” expressions, for reasons that will become more clear [later in the tutorial](#). In order to parse an arbitrary primary expression, we need to determine what sort of expression it is:

```
/// primary
/// ::= identifierexpr
/// ::= numberexpr
/// ::= parenexpr
static std::unique_ptr<ExprAST> ParsePrimary() {
    switch (CurTok) {
    default:
        return LogError("unknown token when expecting an expression");
    case tok_identifier:
        return ParseIdentifierExpr();
    case tok_number:
        return ParseNumberExpr();
    case '(':
        return ParseParenExpr();
    }
}
```

Now that you see the definition of this function, it is more obvious why we can assume the state of `CurTok` in the various functions. This uses look-ahead to determine which sort of expression is being inspected, and then parses it with a function call.

Now that basic expressions are handled, we need to handle binary expressions. They are a bit more complex.

Binary Expression Parsing

Binary expressions are significantly harder to parse because they are often ambiguous. For example, when given the string “ $x+y*z$ ”, the parser can choose to parse it as either “ $(x+y)*z$ ” or “ $x+(y*z)$ ”. With common definitions from mathematics, we expect the later parse, because “ $*$ ” (multiplication) has higher *precedence* than “ $+$ ” (addition).

There are many ways to handle this, but an elegant and efficient way is to use [Operator-Precedence Parsing](#). This parsing technique uses the precedence of binary operators to guide recursion. To start with, we need a table of precedences:

```

/// BinopPrecedence - This holds the precedence for each binary operator that is
/// defined.
static std::map<char, int> BinopPrecedence;

/// GetTokPrecedence - Get the precedence of the pending binary operator token.
static int GetTokPrecedence() {
    if (!isascii(CurTok))
        return -1;

    // Make sure it's a declared binop.
    int TokPrec = BinopPrecedence[CurTok];
    if (TokPrec <= 0) return -1;
    return TokPrec;
}

int main() {
    // Install standard binary operators.
    // 1 is lowest precedence.
    BinopPrecedence['<'] = 10;
    BinopPrecedence['+'] = 20;
    BinopPrecedence['-'] = 20;
    BinopPrecedence['*'] = 40; // highest.
    ...
}

```

For the basic form of Kaleidoscope, we will only support 4 binary operators (this can obviously be extended by you, our brave and intrepid reader). The `GetTokPrecedence` function returns the precedence for the current token, or -1 if the token is not a binary operator. Having a map makes it easy to add new operators and makes it clear that the algorithm doesn’t depend on the specific operators involved, but it would be easy enough to eliminate the map and do the comparisons in the `GetTokPrecedence` function. (Or just use a fixed-size array).

With the helper above defined, we can now start parsing binary expressions. The basic idea of operator precedence parsing is to break down an expression with potentially ambiguous binary operators into pieces. Consider, for example, the expression “ $a+b+(c+d)*e*f+g$ ”. Operator precedence parsing considers this as a stream of primary expressions separated by binary operators. As such, it will first parse the leading primary expression “ a ”, then it will see the pairs $[+, b]$, $[+, (c+d)]$, $[*, e]$, $[*, f]$ and $[+, g]$. Note that because parentheses are primary expressions, the binary expression parser doesn’t need to worry about nested subexpressions like $(c+d)$ at all.

To start, an expression is a primary expression potentially followed by a sequence of `[binop,primaryexpr]` pairs:

```

/// expression
/// ::= primary binoprhs
///
static std::unique_ptr<ExprAST> ParseExpression() {
    auto LHS = ParsePrimary();

```

```

if (!LHS)
    return nullptr;

return ParseBinOpRHS(0, std::move(LHS));
}

```

ParseBinOpRHS is the function that parses the sequence of pairs for us. It takes a precedence and a pointer to an expression for the part that has been parsed so far. Note that “x” is a perfectly valid expression: As such, “binoprhs” is allowed to be empty, in which case it returns the expression that is passed into it. In our example above, the code passes the expression for “a” into ParseBinOpRHS and the current token is “+”.

The precedence value passed into ParseBinOpRHS indicates the *minimal operator precedence* that the function is allowed to eat. For example, if the current pair stream is [+ , x] and ParseBinOpRHS is passed in a precedence of 40, it will not consume any tokens (because the precedence of ‘+’ is only 20). With this in mind, ParseBinOpRHS starts with:

```

/// binoprhs
/// ::= ('+' primary)*
static std::unique_ptr<ExprAST> ParseBinOpRHS(int ExprPrec,
                                              std::unique_ptr<ExprAST> LHS) {
    // If this is a binop, find its precedence.
    while (1) {
        int TokPrec = GetTokPrecedence();

        // If this is a binop that binds at least as tightly as the current binop,
        // consume it, otherwise we are done.
        if (TokPrec < ExprPrec)
            return LHS;
    }
}

```

This code gets the precedence of the current token and checks to see if it is too low. Because we defined invalid tokens to have a precedence of -1, this check implicitly knows that the pair-stream ends when the token stream runs out of binary operators. If this check succeeds, we know that the token is a binary operator and that it will be included in this expression:

```

// Okay, we know this is a binop.
int BinOp = CurTok;
getNextToken(); // eat binop

// Parse the primary expression after the binary operator.
auto RHS = ParsePrimary();
if (!RHS)
    return nullptr;

```

As such, this code eats (and remembers) the binary operator and then parses the primary expression that follows. This builds up the whole pair, the first of which is [+ , b] for the running example.

Now that we parsed the left-hand side of an expression and one pair of the RHS sequence, we have to decide which way the expression associates. In particular, we could have “(a+b) binop unparsed” or “a + (b binop unparsed)”. To determine this, we look ahead at “binop” to determine its precedence and compare it to BinOp’s precedence (which is ‘+’ in this case):

```

// If BinOp binds less tightly with RHS than the operator after RHS, let
// the pending operator take RHS as its LHS.
int NextPrec = GetTokPrecedence();
if (TokPrec < NextPrec) {

```

If the precedence of the binop to the right of “RHS” is lower or equal to the precedence of our current operator, then we know that the parentheses associate as “(a+b) binop ...”. In our example, the current operator is “+” and the next

operator is “+”, we know that they have the same precedence. In this case we’ll create the AST node for “a+b”, and then continue parsing:

```

    ... if body omitted ...
}

// Merge LHS/RHS.
LHS = llvm::make_unique<BinaryExprAST>(BinOp, std::move(LHS),
                                       std::move(RHS));
} // loop around to the top of the while loop.
}

```

In our example above, this will turn “a+b+” into “(a+b)” and execute the next iteration of the loop, with “+” as the current token. The code above will eat, remember, and parse “(c+d)” as the primary expression, which makes the current pair equal to [+ , (c+d)]. It will then evaluate the ‘if’ conditional above with “*” as the binop to the right of the primary. In this case, the precedence of “*” is higher than the precedence of “+” so the if condition will be entered.

The critical question left here is “how can the if condition parse the right hand side in full”? In particular, to build the AST correctly for our example, it needs to get all of “(c+d)*e*f” as the RHS expression variable. The code to do this is surprisingly simple (code from the above two blocks duplicated for context):

```

// If BinOp binds less tightly with RHS than the operator after RHS, let
// the pending operator take RHS as its LHS.
int NextPrec = GetTokPrecedence();
if (TokPrec < NextPrec) {
    RHS = ParseBinOpRHS(TokPrec+1, std::move(RHS));
    if (!RHS)
        return nullptr;
}
// Merge LHS/RHS.
LHS = llvm::make_unique<BinaryExprAST>(BinOp, std::move(LHS),
                                       std::move(RHS));
} // loop around to the top of the while loop.
}

```

At this point, we know that the binary operator to the RHS of our primary has higher precedence than the binop we are currently parsing. As such, we know that any sequence of pairs whose operators are all higher precedence than “+” should be parsed together and returned as “RHS”. To do this, we recursively invoke the `ParseBinOpRHS` function specifying “TokPrec+1” as the minimum precedence required for it to continue. In our example above, this will cause it to return the AST node for “(c+d)*e*f” as RHS, which is then set as the RHS of the ‘+’ expression.

Finally, on the next iteration of the while loop, the “+g” piece is parsed and added to the AST. With this little bit of code (14 non-trivial lines), we correctly handle fully general binary expression parsing in a very elegant way. This was a whirlwind tour of this code, and it is somewhat subtle. I recommend running through it with a few tough examples to see how it works.

This wraps up handling of expressions. At this point, we can point the parser at an arbitrary token stream and build an expression from it, stopping at the first token that is not part of the expression. Next up we need to handle function definitions, etc.

Parsing the Rest

The next thing missing is handling of function prototypes. In Kaleidoscope, these are used both for ‘extern’ function declarations as well as function body definitions. The code to do this is straight-forward and not very interesting (once you’ve survived expressions):

```
/// prototype
/// ::= id '(' id* ')'
static std::unique_ptr<PrototypeAST> ParsePrototype() {
    if (CurTok != tok_identifier)
        return LogErrorP("Expected function name in prototype");

    std::string FnName = IdentifierStr;
    getNextToken();

    if (CurTok != '(')
        return LogErrorP("Expected '(' in prototype");

    // Read the list of argument names.
    std::vector<std::string> ArgNames;
    while (getNextToken() == tok_identifier)
        ArgNames.push_back(IdentifierStr);
    if (CurTok != ')')
        return LogErrorP("Expected ')' in prototype");

    // success.
    getNextToken(); // eat ')'.

    return llvm::make_unique<PrototypeAST>(FnName, std::move(ArgNames));
}
```

Given this, a function definition is very simple, just a prototype plus an expression to implement the body:

```
/// definition ::= 'def' prototype expression
static std::unique_ptr<FunctionAST> ParseDefinition() {
    getNextToken(); // eat def.
    auto Proto = ParsePrototype();
    if (!Proto) return nullptr;

    if (auto E = ParseExpression())
        return llvm::make_unique<FunctionAST>(std::move(Proto), std::move(E));
    return nullptr;
}
```

In addition, we support ‘extern’ to declare functions like ‘sin’ and ‘cos’ as well as to support forward declaration of user functions. These ‘extern’s are just prototypes with no body:

```
/// external ::= 'extern' prototype
static std::unique_ptr<PrototypeAST> ParseExtern() {
    getNextToken(); // eat extern.
    return ParsePrototype();
}
```

Finally, we’ll also let the user type in arbitrary top-level expressions and evaluate them on the fly. We will handle this by defining anonymous nullary (zero argument) functions for them:

```
/// toplevelexpr ::= expression
static std::unique_ptr<FunctionAST> ParseTopLevelExpr() {
    if (auto E = ParseExpression()) {
        // Make an anonymous proto.
        auto Proto = llvm::make_unique<PrototypeAST>("", std::vector<std::string>());
        return llvm::make_unique<FunctionAST>(std::move(Proto), std::move(E));
    }
    return nullptr;
}
```

```
}

```

Now that we have all the pieces, let’s build a little driver that will let us actually *execute* this code we’ve built!

The Driver

The driver for this simply invokes all of the parsing pieces with a top-level dispatch loop. There isn’t much interesting here, so I’ll just include the top-level loop. See [below](#) for full code in the “Top-Level Parsing” section.

```
/// top ::= definition | external | expression | ';'
static void MainLoop() {
    while (1) {
        fprintf(stderr, "ready> ");
        switch (CurTok) {
            case tok_eof:
                return;
            case ';': // ignore top-level semicolons.
                getNextToken();
                break;
            case tok_def:
                HandleDefinition();
                break;
            case tok_extern:
                HandleExtern();
                break;
            default:
                HandleTopLevelExpression();
                break;
        }
    }
}
```

The most interesting part of this is that we ignore top-level semicolons. Why is this, you ask? The basic reason is that if you type “4 + 5” at the command line, the parser doesn’t know whether that is the end of what you will type or not. For example, on the next line you could type “def foo...” in which case 4+5 is the end of a top-level expression. Alternatively you could type “* 6”, which would continue the expression. Having top-level semicolons allows you to type “4+5;”, and the parser will know you are done.

Conclusions

With just under 400 lines of commented code (240 lines of non-comment, non-blank code), we fully defined our minimal language, including a lexer, parser, and AST builder. With this done, the executable will validate Kaleidoscope code and tell us if it is grammatically invalid. For example, here is a sample interaction:

```
$ ./a.out
ready> def foo(x y) x+foo(y, 4.0);
Parsed a function definition.
ready> def foo(x y) x+y y;
Parsed a function definition.
Parsed a top-level expr
ready> def foo(x y) x+y );
Parsed a function definition.
Error: unknown token when expecting an expression
ready> extern sin(a);
```

```
ready> Parsed an extern
ready> ^D
$
```

There is a lot of room for extension here. You can define new AST nodes, extend the language in many ways, etc. In the [next installment](#), we will describe how to generate LLVM Intermediate Representation (IR) from the AST.

Full Code Listing

Here is the complete code listing for this and the previous chapter. Note that it is fully self-contained: you don't need LLVM or any external libraries at all for this. (Besides the C and C++ standard libraries, of course.) To build this, just compile with:

```
# Compile
clang++ -g -O3 toy.cpp
# Run
./a.out
```

Here is the code:

```
#include "llvm/ADT/STLExtras.h"
#include <algorithm>
#include <cctype>
#include <cstdio>
#include <cstdlib>
#include <map>
#include <memory>
#include <string>
#include <vector>

//===-----//
// Lexer
//===-----//

// The lexer returns tokens [0-255] if it is an unknown character, otherwise one
// of these for known things.
enum Token {
    tok_eof = -1,

    // commands
    tok_def = -2,
    tok_extern = -3,

    // primary
    tok_identifier = -4,
    tok_number = -5
};

static std::string IdentifierStr; // Filled in if tok_identifier
static double NumVal;           // Filled in if tok_number

/// gettok - Return the next token from standard input.
static int gettok() {
    static int LastChar = ' ';

    // Skip any whitespace.
```

```

while (isspace>LastChar))
    LastChar = getchar();

if (isalpha>LastChar)) { // identifier: [a-zA-Z][a-zA-Z0-9]*
    IdentifierStr = LastChar;
    while (isalnum((LastChar = getchar()))))
        IdentifierStr += LastChar;

    if (IdentifierStr == "def")
        return tok_def;
    if (IdentifierStr == "extern")
        return tok_extern;
    return tok_identifier;
}

if (isdigit>LastChar) || LastChar == '.') { // Number: [0-9.]+
    std::string NumStr;
    do {
        NumStr += LastChar;
        LastChar = getchar();
    } while (isdigit>LastChar) || LastChar == '.');

    NumVal = strtod(NumStr.c_str(), nullptr);
    return tok_number;
}

if (LastChar == '#') {
    // Comment until end of line.
    do
        LastChar = getchar();
    while (LastChar != EOF && LastChar != '\n' && LastChar != '\r');

    if (LastChar != EOF)
        return gettok();
}

// Check for end of file. Don't eat the EOF.
if (LastChar == EOF)
    return tok_eof;

// Otherwise, just return the character as its ascii value.
int ThisChar = LastChar;
LastChar = getchar();
return ThisChar;
}

//===-----
// Abstract Syntax Tree (aka Parse Tree)
//===-----

namespace {

/// ExprAST - Base class for all expression nodes.
class ExprAST {
public:
    virtual ~ExprAST() = default;
};

```

```
/// NumberExprAST - Expression class for numeric literals like "1.0".
class NumberExprAST : public ExprAST {
    double Val;

public:
    NumberExprAST(double Val) : Val(Val) {}
};

/// VariableExprAST - Expression class for referencing a variable, like "a".
class VariableExprAST : public ExprAST {
    std::string Name;

public:
    VariableExprAST(const std::string &Name) : Name(Name) {}
};

/// BinaryExprAST - Expression class for a binary operator.
class BinaryExprAST : public ExprAST {
    char Op;
    std::unique_ptr<ExprAST> LHS, RHS;

public:
    BinaryExprAST(char Op, std::unique_ptr<ExprAST> LHS,
                  std::unique_ptr<ExprAST> RHS)
        : Op(Op), LHS(std::move(LHS)), RHS(std::move(RHS)) {}
};

/// CallExprAST - Expression class for function calls.
class CallExprAST : public ExprAST {
    std::string Callee;
    std::vector<std::unique_ptr<ExprAST>> Args;

public:
    CallExprAST(const std::string &Callee,
                std::vector<std::unique_ptr<ExprAST>> Args)
        : Callee(Callee), Args(std::move(Args)) {}
};

/// PrototypeAST - This class represents the "prototype" for a function,
/// which captures its name, and its argument names (thus implicitly the number
/// of arguments the function takes).
class PrototypeAST {
    std::string Name;
    std::vector<std::string> Args;

public:
    PrototypeAST(const std::string &Name, std::vector<std::string> Args)
        : Name(Name), Args(std::move(Args)) {}

    const std::string &getName() const { return Name; }
};

/// FunctionAST - This class represents a function definition itself.
class FunctionAST {
    std::unique_ptr<PrototypeAST> Proto;
    std::unique_ptr<ExprAST> Body;

public:
```



```

FunctionAST(std::unique_ptr<PrototypeAST> Proto,
            std::unique_ptr<ExprAST> Body)
    : Proto(std::move(Proto)), Body(std::move(Body)) {}
};

} // end anonymous namespace

//===----- Parser
// Parser
//===-----

/// CurTok/getNextToken - Provide a simple token buffer. CurTok is the current
/// token the parser is looking at. getNextToken reads another token from the
/// lexer and updates CurTok with its results.
static int CurTok;
static int getNextToken() { return CurTok = gettok(); }

/// BinopPrecedence - This holds the precedence for each binary operator that is
/// defined.
static std::map<char, int> BinopPrecedence;

/// GetTokPrecedence - Get the precedence of the pending binary operator token.
static int GetTokPrecedence() {
    if (!isascii(CurTok))
        return -1;

    // Make sure it's a declared binop.
    int TokPrec = BinopPrecedence[CurTok];
    if (TokPrec <= 0)
        return -1;
    return TokPrec;
}

/// LogError* - These are little helper functions for error handling.
std::unique_ptr<ExprAST> LogError(const char *Str) {
    fprintf(stderr, "Error: %s\n", Str);
    return nullptr;
}

std::unique_ptr<PrototypeAST> LogErrorP(const char *Str) {
    LogError(Str);
    return nullptr;
}

static std::unique_ptr<ExprAST> ParseExpression();

/// numberexpr ::= number
static std::unique_ptr<ExprAST> ParseNumberExpr() {
    auto Result = llvm::make_unique<NumberExprAST>(NumVal);
    getNextToken(); // consume the number
    return std::move(Result);
}

/// parenexpr ::= '(' expression ')'
static std::unique_ptr<ExprAST> ParseParenExpr() {
    getNextToken(); // eat (.
    auto V = ParseExpression();
    if (!V)
        return nullptr;
    return nullptr;
}

```

```

    if (CurTok != ')')
        return LogError("expected ')'");
    getNextToken(); // eat ).
    return V;
}

/// identifierexpr
/// ::= identifier
/// ::= identifier '(' expression* ')'
static std::unique_ptr<ExprAST> ParseIdentifierExpr() {
    std::string IdName = IdentifierStr;

    getNextToken(); // eat identifier.

    if (CurTok != '(') // Simple variable ref.
        return llvm::make_unique<VariableExprAST>(IdName);

    // Call.
    getNextToken(); // eat (
    std::vector<std::unique_ptr<ExprAST>> Args;
    if (CurTok != ')') {
        while (true) {
            if (auto Arg = ParseExpression())
                Args.push_back(std::move(Arg));
            else
                return nullptr;

            if (CurTok == ')')
                break;

            if (CurTok != ',')
                return LogError("Expected ')' or ',' in argument list");
            getNextToken();
        }
    }

    // Eat the ')'.
    getNextToken();

    return llvm::make_unique<CallExprAST>(IdName, std::move(Args));
}

/// primary
/// ::= identifierexpr
/// ::= numberexpr
/// ::= parenexpr
static std::unique_ptr<ExprAST> ParsePrimary() {
    switch (CurTok) {
    default:
        return LogError("unknown token when expecting an expression");
    case tok_identifier:
        return ParseIdentifierExpr();
    case tok_number:
        return ParseNumberExpr();
    case '(':
        return ParseParenExpr();
    }
}

```

```

}

/// binoprhs
/// ::= ('+' primary)*
static std::unique_ptr<ExprAST> ParseBinOpRHS(int ExprPrec,
                                             std::unique_ptr<ExprAST> LHS) {
    // If this is a binop, find its precedence.
    while (true) {
        int TokPrec = GetTokPrecedence();

        // If this is a binop that binds at least as tightly as the current binop,
        // consume it, otherwise we are done.
        if (TokPrec < ExprPrec)
            return LHS;

        // Okay, we know this is a binop.
        int BinOp = CurTok;
        getNextToken(); // eat binop

        // Parse the primary expression after the binary operator.
        auto RHS = ParsePrimary();
        if (!RHS)
            return nullptr;

        // If BinOp binds less tightly with RHS than the operator after RHS, let
        // the pending operator take RHS as its LHS.
        int NextPrec = GetTokPrecedence();
        if (TokPrec < NextPrec) {
            RHS = ParseBinOpRHS(TokPrec + 1, std::move(RHS));
            if (!RHS)
                return nullptr;
        }

        // Merge LHS/RHS.
        LHS = llvm::make_unique<BinaryExprAST>(BinOp, std::move(LHS),
                                              std::move(RHS));
    }
}

/// expression
/// ::= primary binoprhs
///
static std::unique_ptr<ExprAST> ParseExpression() {
    auto LHS = ParsePrimary();
    if (!LHS)
        return nullptr;

    return ParseBinOpRHS(0, std::move(LHS));
}

/// prototype
/// ::= id '(' id* ')'
static std::unique_ptr<PrototypeAST> ParsePrototype() {
    if (CurTok != tok_identifier)
        return LogErrorP("Expected function name in prototype");

    std::string FnName = IdentifierStr;
    getNextToken();

```

```

    if (CurTok != '(')
        return LogErrorP("Expected '(' in prototype");

    std::vector<std::string> ArgNames;
    while (getNextToken() == tok_identifier)
        ArgNames.push_back(IdentifierStr);
    if (CurTok != ')')
        return LogErrorP("Expected ')' in prototype");

    // success.
    getNextToken(); // eat ')'.

    return llvm::make_unique<PrototypeAST>(FnName, std::move(ArgNames));
}

// definition ::= 'def' prototype expression
static std::unique_ptr<FunctionAST> ParseDefinition() {
    getNextToken(); // eat def.
    auto Proto = ParsePrototype();
    if (!Proto)
        return nullptr;

    if (auto E = ParseExpression())
        return llvm::make_unique<FunctionAST>(std::move(Proto), std::move(E));
    return nullptr;
}

// toplevelexpr ::= expression
static std::unique_ptr<FunctionAST> ParseTopLevelExpr() {
    if (auto E = ParseExpression()) {
        // Make an anonymous proto.
        auto Proto = llvm::make_unique<PrototypeAST>("__anon_expr",
                                                    std::vector<std::string>());
        return llvm::make_unique<FunctionAST>(std::move(Proto), std::move(E));
    }
    return nullptr;
}

// external ::= 'extern' prototype
static std::unique_ptr<PrototypeAST> ParseExtern() {
    getNextToken(); // eat extern.
    return ParsePrototype();
}

//=====//
// Top-Level parsing
//=====//

static void HandleDefinition() {
    if (ParseDefinition()) {
        fprintf(stderr, "Parsed a function definition.\n");
    } else {
        // Skip token for error recovery.
        getNextToken();
    }
}

```

```

static void HandleExtern() {
    if (ParseExtern()) {
        fprintf(stderr, "Parsed an extern\n");
    } else {
        // Skip token for error recovery.
        getNextToken();
    }
}

static void HandleTopLevelExpression() {
    // Evaluate a top-level expression into an anonymous function.
    if (ParseTopLevelExpr()) {
        fprintf(stderr, "Parsed a top-level expr\n");
    } else {
        // Skip token for error recovery.
        getNextToken();
    }
}

// top ::= definition | external | expression | ';'
static void MainLoop() {
    while (true) {
        fprintf(stderr, "ready> ");
        switch (CurTok) {
            case tok_eof:
                return;
            case ';': // ignore top-level semicolons.
                getNextToken();
                break;
            case tok_def:
                HandleDefinition();
                break;
            case tok_extern:
                HandleExtern();
                break;
            default:
                HandleTopLevelExpression();
                break;
        }
    }
}

//=====//
// Main driver code.
//=====//

int main() {
    // Install standard binary operators.
    // 1 is lowest precedence.
    BinopPrecedence['<'] = 10;
    BinopPrecedence['+'] = 20;
    BinopPrecedence['-'] = 20;
    BinopPrecedence['*'] = 40; // highest.

    // Prime the first token.
    fprintf(stderr, "ready> ");
    getNextToken();
}

```

```
// Run the main "interpreter loop" now.
MainLoop();

return 0;
}
```

Next: Implementing Code Generation to LLVM IR

Kaleidoscope: Code generation to LLVM IR

- *Chapter 3 Introduction*
- *Code Generation Setup*
- *Expression Code Generation*
- *Function Code Generation*
- *Driver Changes and Closing Thoughts*
- *Full Code Listing*

Chapter 3 Introduction

Welcome to Chapter 3 of the “Implementing a language with LLVM” tutorial. This chapter shows you how to transform the [Abstract Syntax Tree](#), built in Chapter 2, into LLVM IR. This will teach you a little bit about how LLVM does things, as well as demonstrate how easy it is to use. It’s much more work to build a lexer and parser than it is to generate LLVM IR code. :)

Please note: the code in this chapter and later require LLVM 3.7 or later. LLVM 3.6 and before will not work with it. Also note that you need to use a version of this tutorial that matches your LLVM release: If you are using an official LLVM release, use the version of the documentation included with your release or on the [llvm.org releases](http://llvm.org/releases) page.

Code Generation Setup

In order to generate LLVM IR, we want some simple setup to get started. First we define virtual code generation (codegen) methods in each AST class:

```
/// ExprAST - Base class for all expression nodes.
class ExprAST {
public:
    virtual ~ExprAST() {}
    virtual Value *codegen() = 0;
};

/// NumberExprAST - Expression class for numeric literals like "1.0".
class NumberExprAST : public ExprAST {
    double Val;

public:
    NumberExprAST(double Val) : Val(Val) {}
    virtual Value *codegen();
};
```

```
};
...
```

The `codegen()` method says to emit IR for that AST node along with all the things it depends on, and they all return an LLVM Value object. “Value” is the class used to represent a “[Static Single Assignment \(SSA\)](#) register” or “SSA value” in LLVM. The most distinct aspect of SSA values is that their value is computed as the related instruction executes, and it does not get a new value until (and if) the instruction re-executes. In other words, there is no way to “change” an SSA value. For more information, please read up on [Static Single Assignment](#) - the concepts are really quite natural once you grok them.

Note that instead of adding virtual methods to the `ExprAST` class hierarchy, it could also make sense to use a [visitor pattern](#) or some other way to model this. Again, this tutorial won’t dwell on good software engineering practices: for our purposes, adding a virtual method is simplest.

The second thing we want is an “LogError” method like we used for the parser, which will be used to report errors found during code generation (for example, use of an undeclared parameter):

```
static LLVMContext TheContext;
static IRBuilder<> Builder(TheContext);
static std::unique_ptr<Module> TheModule;
static std::map<std::string, Value*> NamedValues;

Value *LogErrorV(const char *Str) {
    LogError(Str);
    return nullptr;
}
```

The static variables will be used during code generation. `TheContext` is an opaque object that owns a lot of core LLVM data structures, such as the type and constant value tables. We don’t need to understand it in detail, we just need a single instance to pass into APIs that require it.

The `Builder` object is a helper object that makes it easy to generate LLVM instructions. Instances of the `IRBuilder` class template keep track of the current place to insert instructions and has methods to create new instructions.

`TheModule` is an LLVM construct that contains functions and global variables. In many ways, it is the top-level structure that the LLVM IR uses to contain code. It will own the memory for all of the IR that we generate, which is why the `codegen()` method returns a raw `Value*`, rather than a `unique_ptr<Value>`.

The `NamedValues` map keeps track of which values are defined in the current scope and what their LLVM representation is. (In other words, it is a symbol table for the code). In this form of Kaleidoscope, the only things that can be referenced are function parameters. As such, function parameters will be in this map when generating code for their function body.

With these basics in place, we can start talking about how to generate code for each expression. Note that this assumes that the `Builder` has been set up to generate code *into* something. For now, we’ll assume that this has already been done, and we’ll just use it to emit code.

Expression Code Generation

Generating LLVM code for expression nodes is very straightforward: less than 45 lines of commented code for all four of our expression nodes. First we’ll do numeric literals:

```
Value *NumberExprAST::codegen() {
    return ConstantFP::get(TheContext, APFloat(Val));
}
```

In the LLVM IR, numeric constants are represented with the `ConstantFP` class, which holds the numeric value in an `APFloat` internally (`APFloat` has the capability of holding floating point constants of Arbitrary Precision). This code basically just creates and returns a `ConstantFP`. Note that in the LLVM IR that constants are all uniqued together and shared. For this reason, the API uses the “`foo::get(...)`” idiom instead of “`new foo(..)`” or “`foo::Create(..)`”.

```
Value *VariableExprAST::codegen() {  
    // Look this variable up in the function.  
    Value *V = NamedValues[Name];  
    if (!V)  
        LogErrorV("Unknown variable name");  
    return V;  
}
```

References to variables are also quite simple using LLVM. In the simple version of Kaleidoscope, we assume that the variable has already been emitted somewhere and its value is available. In practice, the only values that can be in the `NamedValues` map are function arguments. This code simply checks to see that the specified name is in the map (if not, an unknown variable is being referenced) and returns the value for it. In future chapters, we’ll add support for [loop induction variables](#) in the symbol table, and for [local variables](#).

```
Value *BinaryExprAST::codegen() {  
    Value *L = LHS->codegen();  
    Value *R = RHS->codegen();  
    if (!L || !R)  
        return nullptr;  
  
    switch (Op) {  
    case '+':  
        return Builder.CreateFAdd(L, R, "addtmp");  
    case '-':  
        return Builder.CreateFSub(L, R, "subtmp");  
    case '*':  
        return Builder.CreateFMul(L, R, "multmp");  
    case '<':  
        L = Builder.CreateFCmpULT(L, R, "cmptmp");  
        // Convert bool 0/1 to double 0.0 or 1.0  
        return Builder.CreateUIToFP(L, Type::getDoubleTy(TheContext),  
                                    "booltmp");  
    default:  
        return LogErrorV("invalid binary operator");  
    }  
}
```

Binary operators start to get more interesting. The basic idea here is that we recursively emit code for the left-hand side of the expression, then the right-hand side, then we compute the result of the binary expression. In this code, we do a simple switch on the opcode to create the right LLVM instruction.

In the example above, the LLVM builder class is starting to show its value. `IRBuilder` knows where to insert the newly created instruction, all you have to do is specify what instruction to create (e.g. with `CreateFAdd`), which operands to use (`L` and `R` here) and optionally provide a name for the generated instruction.

One nice thing about LLVM is that the name is just a hint. For instance, if the code above emits multiple “`addtmp`” variables, LLVM will automatically provide each one with an increasing, unique numeric suffix. Local value names for instructions are purely optional, but it makes it much easier to read the IR dumps.

[LLVM instructions](#) are constrained by strict rules: for example, the Left and Right operators of an [add instruction](#) must have the same type, and the result type of the add must match the operand types. Because all values in Kaleidoscope are doubles, this makes for very simple code for add, sub and mul.

On the other hand, LLVM specifies that the [fcmp instruction](#) always returns an ‘i1’ value (a one bit integer). The

problem with this is that Kaleidoscope wants the value to be a 0.0 or 1.0 value. In order to get these semantics, we combine the `fcmp` instruction with a `uitofp` instruction. This instruction converts its input integer into a floating point value by treating the input as an unsigned value. In contrast, if we used the `sitofp` instruction, the Kaleidoscope ‘<’ operator would return 0.0 and -1.0, depending on the input value.

```
Value *CallExprAST::codegen() {
    // Look up the name in the global module table.
    Function *CalleeF = TheModule->getFunction(Callee);
    if (!CalleeF)
        return LogErrorV("Unknown function referenced");

    // If argument mismatch error.
    if (CalleeF->arg_size() != Args.size())
        return LogErrorV("Incorrect # arguments passed");

    std::vector<Value *> ArgsV;
    for (unsigned i = 0, e = Args.size(); i != e; ++i) {
        ArgsV.push_back(Args[i]->codegen());
        if (!ArgsV.back())
            return nullptr;
    }

    return Builder.CreateCall(CalleeF, ArgsV, "calltmp");
}
```

Code generation for function calls is quite straightforward with LLVM. The code above initially does a function name lookup in the LLVM Module’s symbol table. Recall that the LLVM Module is the container that holds the functions we are JIT’ing. By giving each function the same name as what the user specifies, we can use the LLVM symbol table to resolve function names for us.

Once we have the function to call, we recursively codegen each argument that is to be passed in, and create an LLVM `call` instruction. Note that LLVM uses the native C calling conventions by default, allowing these calls to also call into standard library functions like “sin” and “cos”, with no additional effort.

This wraps up our handling of the four basic expressions that we have so far in Kaleidoscope. Feel free to go in and add some more. For example, by browsing the [LLVM language reference](#) you’ll find several other interesting instructions that are really easy to plug into our basic framework.

Function Code Generation

Code generation for prototypes and functions must handle a number of details, which make their code less beautiful than expression code generation, but allows us to illustrate some important points. First, let’s talk about code generation for prototypes: they are used both for function bodies and external function declarations. The code starts with:

```
Function *PrototypeAST::codegen() {
    // Make the function type: double(double,double) etc.
    std::vector<Type*> Doubles(Args.size(),
                               Type::getDoubleTy(TheContext));

    FunctionType *FT =
        FunctionType::get(Type::getDoubleTy(TheContext), Doubles, false);

    Function *F =
        Function::Create(FT, Function::ExternalLinkage, Name, TheModule);
}
```

This code packs a lot of power into a few lines. Note first that this function returns a “Function*” instead of a “Value*”. Because a “prototype” really talks about the external interface for a function (not the value computed by an expression), it makes sense for it to return the LLVM Function it corresponds to when codegen’d.

The call to `FunctionType::get` creates the `FunctionType` that should be used for a given `Prototype`. Since all function arguments in Kaleidoscope are of type double, the first line creates a vector of “N” LLVM double types. It then uses the `FunctionType::get` method to create a function type that takes “N” doubles as arguments, returns one double as a result, and that is not vararg (the false parameter indicates this). Note that Types in LLVM are unique just like Constants are, so you don’t “new” a type, you “get” it.

The final line above actually creates the IR Function corresponding to the `Prototype`. This indicates the type, linkage and name to use, as well as which module to insert into. “external linkage” means that the function may be defined outside the current module and/or that it is callable by functions outside the module. The Name passed in is the name the user specified: since “TheModule” is specified, this name is registered in “TheModule”’s symbol table.

```
// Set names for all arguments.
unsigned Idx = 0;
for (auto &Arg : F->args())
    Arg.setName(Args[Idx++]);

return F;
```

Finally, we set the name of each of the function’s arguments according to the names given in the `Prototype`. This step isn’t strictly necessary, but keeping the names consistent makes the IR more readable, and allows subsequent code to refer directly to the arguments for their names, rather than having to look up them up in the `Prototype` AST.

At this point we have a function prototype with no body. This is how LLVM IR represents function declarations. For extern statements in Kaleidoscope, this is as far as we need to go. For function definitions however, we need to codegen and attach a function body.

```
Function *FunctionAST::codegen() {
    // First, check for an existing function from a previous 'extern' declaration.
    Function *TheFunction = TheModule->getFunction(Proto->getName());

    if (!TheFunction)
        TheFunction = Proto->codegen();

    if (!TheFunction)
        return nullptr;

    if (!TheFunction->empty())
        return (Function*)LogErrorV("Function cannot be redefined.");
```

For function definitions, we start by searching `TheModule`’s symbol table for an existing version of this function, in case one has already been created using an ‘extern’ statement. If `Module::getFunction` returns null then no previous version exists, so we’ll codegen one from the `Prototype`. In either case, we want to assert that the function is empty (i.e. has no body yet) before we start.

```
// Create a new basic block to start insertion into.
BasicBlock *BB = BasicBlock::Create(TheContext, "entry", TheFunction);
Builder.SetInsertPoint(BB);

// Record the function arguments in the NamedValues map.
NamedValues.clear();
for (auto &Arg : TheFunction->args())
    NamedValues[Arg.getName()] = &Arg;
```

Now we get to the point where the `Builder` is set up. The first line creates a new **basic block** (named “entry”), which is inserted into `TheFunction`. The second line then tells the builder that new instructions should be inserted into the end of the new basic block. Basic blocks in LLVM are an important part of functions that define the **Control Flow Graph**. Since we don’t have any control flow, our functions will only contain one block at this point. We’ll fix this in Chapter 5 :).

Next we add the function arguments to the `NamedValues` map (after first clearing it out) so that they’re accessible to `VariableExprAST` nodes.

```
if (Value *RetVal = Body->codegen()) {
    // Finish off the function.
    Builder.CreateRet(RetVal);

    // Validate the generated code, checking for consistency.
    verifyFunction(*TheFunction);

    return TheFunction;
}
```

Once the insertion point has been set up and the `NamedValues` map populated, we call the `codegen()` method for the root expression of the function. If no error happens, this emits code to compute the expression into the entry block and returns the value that was computed. Assuming no error, we then create an LLVM `ret instruction`, which completes the function. Once the function is built, we call `verifyFunction`, which is provided by LLVM. This function does a variety of consistency checks on the generated code, to determine if our compiler is doing everything right. Using this is important: it can catch a lot of bugs. Once the function is finished and validated, we return it.

```
// Error reading body, remove function.
TheFunction->eraseFromParent();
return nullptr;
}
```

The only piece left here is handling of the error case. For simplicity, we handle this by merely deleting the function we produced with the `eraseFromParent` method. This allows the user to redefine a function that they incorrectly typed in before: if we didn’t delete it, it would live in the symbol table, with a body, preventing future redefinition.

This code does have a bug, though: If the `FunctionAST::codegen()` method finds an existing IR Function, it does not validate its signature against the definition’s own prototype. This means that an earlier ‘extern’ declaration will take precedence over the function definition’s signature, which can cause `codegen` to fail, for instance if the function arguments are named differently. There are a number of ways to fix this bug, see what you can come up with! Here is a testcase:

```
extern foo(a);           # ok, defines foo.
def foo(b) b;           # Error: Unknown variable name. (decl using 'a' takes precedence).
```

Driver Changes and Closing Thoughts

For now, code generation to LLVM doesn’t really get us much, except that we can look at the pretty IR calls. The sample code inserts calls to `codegen` into the “`HandleDefinition`”, “`HandleExtern`” etc functions, and then dumps out the LLVM IR. This gives a nice way to look at the LLVM IR for simple functions. For example:

```
ready> 4+5;
Read top-level expression:
define double @0() {
entry:
    ret double 9.000000e+00
}
```

Note how the parser turns the top-level expression into anonymous functions for us. This will be handy when we add `JIT support` in the next chapter. Also note that the code is very literally transcribed, no optimizations are being performed except simple constant folding done by `IRBuilder`. We will add `optimizations` explicitly in the next chapter.

```
ready> def foo(a b) a*a + 2*a*b + b*b;
Read function definition:
define double @foo(double %a, double %b) {
entry:
    %multmp = fmul double %a, %a
    %multmp1 = fmul double 2.000000e+00, %a
    %multmp2 = fmul double %multmp1, %b
    %addtmp = fadd double %multmp, %multmp2
    %multmp3 = fmul double %b, %b
    %addtmp4 = fadd double %addtmp, %multmp3
    ret double %addtmp4
}
```

This shows some simple arithmetic. Notice the striking similarity to the LLVM builder calls that we use to create the instructions.

```
ready> def bar(a) foo(a, 4.0) + bar(31337);
Read function definition:
define double @bar(double %a) {
entry:
    %calltmp = call double @foo(double %a, double 4.000000e+00)
    %calltmp1 = call double @bar(double 3.133700e+04)
    %addtmp = fadd double %calltmp, %calltmp1
    ret double %addtmp
}
```

This shows some function calls. Note that this function will take a long time to execute if you call it. In the future we'll add conditional control flow to actually make recursion useful :).

```
ready> extern cos(x);
Read extern:
declare double @cos(double)

ready> cos(1.234);
Read top-level expression:
define double @1() {
entry:
    %calltmp = call double @cos(double 1.234000e+00)
    ret double %calltmp
}
```

This shows an extern for the libm “cos” function, and a call to it.

```
ready> ^D
; ModuleID = 'my cool jit'

define double @0() {
entry:
    %addtmp = fadd double 4.000000e+00, 5.000000e+00
    ret double %addtmp
}

define double @foo(double %a, double %b) {
entry:
    %multmp = fmul double %a, %a
    %multmp1 = fmul double 2.000000e+00, %a
    %multmp2 = fmul double %multmp1, %b
    %addtmp = fadd double %multmp, %multmp2
}
```

```

    %multmp3 = fmul double %b, %b
    %addtmp4 = fadd double %addtmp, %multmp3
    ret double %addtmp4
}

define double @bar(double %a) {
entry:
    %calltmp = call double @foo(double %a, double 4.000000e+00)
    %calltmp1 = call double @bar(double 3.133700e+04)
    %addtmp = fadd double %calltmp, %calltmp1
    ret double %addtmp
}

declare double @cos(double)

define double @1() {
entry:
    %calltmp = call double @cos(double 1.234000e+00)
    ret double %calltmp
}

```

When you quit the current demo (by sending an EOF via CTRL+D on Linux or CTRL+Z and ENTER on Windows), it dumps out the IR for the entire module generated. Here you can see the big picture with all the functions referencing each other.

This wraps up the third chapter of the Kaleidoscope tutorial. Up next, we'll describe how to [add JIT codegen and optimizer support](#) to this so we can actually start running code!

Full Code Listing

Here is the complete code listing for our running example, enhanced with the LLVM code generator. Because this uses the LLVM libraries, we need to link them in. To do this, we use the `llvm-config` tool to inform our makefile/command line about which options to use:

```

# Compile
clang++ -g -O3 toy.cpp `llvm-config --cxxflags --ldflags --system-libs --libs core` -
-o toy
# Run
./toy

```

Here is the code:

```

#include "llvm/ADT/APFloat.h"
#include "llvm/ADT/STLExtras.h"
#include "llvm/IR/BasicBlock.h"
#include "llvm/IR/Constants.h"
#include "llvm/IR/DerivedTypes.h"
#include "llvm/IR/Function.h"
#include "llvm/IR/IRBuilder.h"
#include "llvm/IR/LLVMSContext.h"
#include "llvm/IR/Module.h"
#include "llvm/IR/Type.h"
#include "llvm/IR/Verifier.h"
#include <algorithm>
#include <cctype>
#include <cstdio>

```

```

#include <cstdlib>
#include <map>
#include <memory>
#include <string>
#include <vector>

using namespace llvm;

//===-----//
// Lexer
//===-----//

// The lexer returns tokens [0-255] if it is an unknown character, otherwise one
// of these for known things.
enum Token {
    tok_eof = -1,

    // commands
    tok_def = -2,
    tok_extern = -3,

    // primary
    tok_identifier = -4,
    tok_number = -5
};

static std::string IdentifierStr; // Filled in if tok_identifier
static double NumVal;           // Filled in if tok_number

/// gettok - Return the next token from standard input.
static int gettok() {
    static int LastChar = ' ';

    // Skip any whitespace.
    while (isspace(LastChar))
        LastChar = getchar();

    if (isalpha(LastChar)) { // identifier: [a-zA-Z][a-zA-Z0-9]*
        IdentifierStr = LastChar;
        while (isalnum((LastChar = getchar())))
            IdentifierStr += LastChar;

        if (IdentifierStr == "def")
            return tok_def;
        if (IdentifierStr == "extern")
            return tok_extern;
        return tok_identifier;
    }

    if (isdigit(LastChar) || LastChar == '.') { // Number: [0-9.]+
        std::string NumStr;
        do {
            NumStr += LastChar;
            LastChar = getchar();
        } while (isdigit(LastChar) || LastChar == '.');

        NumVal = strtod(NumStr.c_str(), nullptr);
        return tok_number;
    }

```

```

}

if (LastChar == '#') {
    // Comment until end of line.
    do
        LastChar = getchar();
    while (LastChar != EOF && LastChar != '\n' && LastChar != '\r');

    if (LastChar != EOF)
        return gettok();
}

// Check for end of file. Don't eat the EOF.
if (LastChar == EOF)
    return tok_eof;

// Otherwise, just return the character as its ascii value.
int ThisChar = LastChar;
LastChar = getchar();
return ThisChar;
}

//===-----
// Abstract Syntax Tree (aka Parse Tree)
//===-----

namespace {

/// ExprAST - Base class for all expression nodes.
class ExprAST {
public:
    virtual ~ExprAST() = default;

    virtual Value *codegen() = 0;
};

/// NumberExprAST - Expression class for numeric literals like "1.0".
class NumberExprAST : public ExprAST {
    double Val;

public:
    NumberExprAST(double Val) : Val(Val) {}

    Value *codegen() override;
};

/// VariableExprAST - Expression class for referencing a variable, like "a".
class VariableExprAST : public ExprAST {
    std::string Name;

public:
    VariableExprAST(const std::string &Name) : Name(Name) {}

    Value *codegen() override;
};

/// BinaryExprAST - Expression class for a binary operator.
class BinaryExprAST : public ExprAST {

```

```

    char Op;
    std::unique_ptr<ExprAST> LHS, RHS;

public:
    BinaryExprAST(char Op, std::unique_ptr<ExprAST> LHS,
                  std::unique_ptr<ExprAST> RHS)
        : Op(Op), LHS(std::move(LHS)), RHS(std::move(RHS)) {}

    Value *codegen() override;
};

/// CallExprAST - Expression class for function calls.
class CallExprAST : public ExprAST {
    std::string Callee;
    std::vector<std::unique_ptr<ExprAST>> Args;

public:
    CallExprAST(const std::string &Callee,
                std::vector<std::unique_ptr<ExprAST>> Args)
        : Callee(Callee), Args(std::move(Args)) {}

    Value *codegen() override;
};

/// PrototypeAST - This class represents the "prototype" for a function,
/// which captures its name, and its argument names (thus implicitly the number
/// of arguments the function takes).
class PrototypeAST {
    std::string Name;
    std::vector<std::string> Args;

public:
    PrototypeAST(const std::string &Name, std::vector<std::string> Args)
        : Name(Name), Args(std::move(Args)) {}

    Function *codegen();
    const std::string &getName() const { return Name; }
};

/// FunctionAST - This class represents a function definition itself.
class FunctionAST {
    std::unique_ptr<PrototypeAST> Proto;
    std::unique_ptr<ExprAST> Body;

public:
    FunctionAST(std::unique_ptr<PrototypeAST> Proto,
                std::unique_ptr<ExprAST> Body)
        : Proto(std::move(Proto)), Body(std::move(Body)) {}

    Function *codegen();
};

} // end anonymous namespace

//=====
// Parser
//=====

```



```

/// CurTok/getNextToken - Provide a simple token buffer. CurTok is the current
/// token the parser is looking at. getNextToken reads another token from the
/// lexer and updates CurTok with its results.
static int CurTok;
static int getNextToken() { return CurTok = gettok(); }

/// BinopPrecedence - This holds the precedence for each binary operator that is
/// defined.
static std::map<char, int> BinopPrecedence;

/// GetTokPrecedence - Get the precedence of the pending binary operator token.
static int GetTokPrecedence() {
    if (!isascii(CurTok))
        return -1;

    // Make sure it's a declared binop.
    int TokPrec = BinopPrecedence[CurTok];
    if (TokPrec <= 0)
        return -1;
    return TokPrec;
}

/// LogError* - These are little helper functions for error handling.
std::unique_ptr<ExprAST> LogError(const char *Str) {
    fprintf(stderr, "Error: %s\n", Str);
    return nullptr;
}

std::unique_ptr<PrototypeAST> LogErrorP(const char *Str) {
    LogError(Str);
    return nullptr;
}

static std::unique_ptr<ExprAST> ParseExpression();

/// numberexpr ::= number
static std::unique_ptr<ExprAST> ParseNumberExpr() {
    auto Result = llvm::make_unique<NumberExprAST>(NumVal);
    getNextToken(); // consume the number
    return std::move(Result);
}

/// parenexpr ::= '(' expression ')'
static std::unique_ptr<ExprAST> ParseParenExpr() {
    getNextToken(); // eat '('
    auto V = ParseExpression();
    if (!V)
        return nullptr;

    if (CurTok != ')')
        return LogError("expected ')'");
    getNextToken(); // eat ')'
    return V;
}

/// identifierexpr
/// ::= identifier
/// ::= identifier '(' expression* ')'

```

```

static std::unique_ptr<ExprAST> ParseIdentifierExpr() {
    std::string IdName = IdentifierStr;

    getNextToken(); // eat identifier.

    if (CurTok != '(') // Simple variable ref.
        return llvm::make_unique<VariableExprAST>(IdName);

    // Call.
    getNextToken(); // eat (
    std::vector<std::unique_ptr<ExprAST>> Args;
    if (CurTok != ')') {
        while (true) {
            if (auto Arg = ParseExpression())
                Args.push_back(std::move(Arg));
            else
                return nullptr;

            if (CurTok == ')')
                break;

            if (CurTok != ',')
                return LogError("Expected ')' or ',' in argument list");
            getNextToken();
        }
    }

    // Eat the ')'.
    getNextToken();

    return llvm::make_unique<CallExprAST>(IdName, std::move(Args));
}

/// primary
/// ::= identifierexpr
/// ::= numberexpr
/// ::= parenexpr
static std::unique_ptr<ExprAST> ParsePrimary() {
    switch (CurTok) {
    default:
        return LogError("unknown token when expecting an expression");
    case tok_identifier:
        return ParseIdentifierExpr();
    case tok_number:
        return ParseNumberExpr();
    case '(':
        return ParseParenExpr();
    }
}

/// binoprhs
/// ::= ('+' primary)*
static std::unique_ptr<ExprAST> ParseBinOpRHS(int ExprPrec,
                                              std::unique_ptr<ExprAST> LHS) {
    // If this is a binop, find its precedence.
    while (true) {
        int TokPrec = GetTokPrecedence();

```

```

// If this is a binop that binds at least as tightly as the current binop,
// consume it, otherwise we are done.
if (TokPrec < ExprPrec)
    return LHS;

// Okay, we know this is a binop.
int BinOp = CurTok;
getNextToken(); // eat binop

// Parse the primary expression after the binary operator.
auto RHS = ParsePrimary();
if (!RHS)
    return nullptr;

// If BinOp binds less tightly with RHS than the operator after RHS, let
// the pending operator take RHS as its LHS.
int NextPrec = GetTokPrecedence();
if (TokPrec < NextPrec) {
    RHS = ParseBinOpRHS(TokPrec + 1, std::move(RHS));
    if (!RHS)
        return nullptr;
}

// Merge LHS/RHS.
LHS =
    llvm::make_unique<BinaryExprAST>(BinOp, std::move(LHS), std::move(RHS));
}

/// expression
/// ::= primary binoprhs
///
static std::unique_ptr<ExprAST> ParseExpression() {
    auto LHS = ParsePrimary();
    if (!LHS)
        return nullptr;

    return ParseBinOpRHS(0, std::move(LHS));
}

/// prototype
/// ::= id '(' id* ')'
static std::unique_ptr<PrototypeAST> ParsePrototype() {
    if (CurTok != tok_identifier)
        return LogErrorP("Expected function name in prototype");

    std::string FnName = IdentifierStr;
    getNextToken();

    if (CurTok != '(')
        return LogErrorP("Expected '(' in prototype");

    std::vector<std::string> ArgNames;
    while (getNextToken() == tok_identifier)
        ArgNames.push_back(IdentifierStr);
    if (CurTok != ')')
        return LogErrorP("Expected ')' in prototype");

```

```

// success.
getNextToken(); // eat ')'.

return llvm::make_unique<PrototypeAST>(FnName, std::move(ArgNames));
}

/// definition ::= 'def' prototype expression
static std::unique_ptr<FunctionAST> ParseDefinition() {
    getNextToken(); // eat def.
    auto Proto = ParsePrototype();
    if (!Proto)
        return nullptr;

    if (auto E = ParseExpression())
        return llvm::make_unique<FunctionAST>(std::move(Proto), std::move(E));
    return nullptr;
}

/// toplevelexpr ::= expression
static std::unique_ptr<FunctionAST> ParseTopLevelExpr() {
    if (auto E = ParseExpression()) {
        // Make an anonymous proto.
        auto Proto = llvm::make_unique<PrototypeAST>("__anon_expr",
                                                    std::vector<std::string>());
        return llvm::make_unique<FunctionAST>(std::move(Proto), std::move(E));
    }
    return nullptr;
}

/// external ::= 'extern' prototype
static std::unique_ptr<PrototypeAST> ParseExtern() {
    getNextToken(); // eat extern.
    return ParsePrototype();
}

//===-----
// Code Generation
//===-----

static LLVMContext TheContext;
static IRBuilder<> Builder(TheContext);
static std::unique_ptr<Module> TheModule;
static std::map<std::string, Value*> NamedValues;

Value *LogErrorV(const char *Str) {
    LogError(Str);
    return nullptr;
}

Value *NumberExprAST::codegen() {
    return ConstantFP::get(TheContext, APFloat(Val));
}

Value *VariableExprAST::codegen() {
    // Look this variable up in the function.
    Value *V = NamedValues[Name];
    if (!V)
        return LogErrorV("Unknown variable name");
}

```

```

    return V;
}

Value *BinaryExprAST::codegen() {
    Value *L = LHS->codegen();
    Value *R = RHS->codegen();
    if (!L || !R)
        return nullptr;

    switch (Op) {
    case '+':
        return Builder.CreateFAdd(L, R, "addtmp");
    case '-':
        return Builder.CreateFSub(L, R, "subtmp");
    case '*':
        return Builder.CreateFMul(L, R, "multmp");
    case '<':
        L = Builder.CreateFCmpULT(L, R, "cmptmp");
        // Convert bool 0/1 to double 0.0 or 1.0
        return Builder.CreateUIToFP(L, Type::getDoubleTy(TheContext), "booltmp");
    default:
        return LogErrorV("invalid binary operator");
    }
}

Value *CallExprAST::codegen() {
    // Look up the name in the global module table.
    Function *CalleeF = TheModule->getFunction(Callee);
    if (!CalleeF)
        return LogErrorV("Unknown function referenced");

    // If argument mismatch error.
    if (CalleeF->arg_size() != Args.size())
        return LogErrorV("Incorrect # arguments passed");

    std::vector<Value *> ArgsV;
    for (unsigned i = 0, e = Args.size(); i != e; ++i) {
        ArgsV.push_back(Args[i]->codegen());
        if (!ArgsV.back())
            return nullptr;
    }

    return Builder.CreateCall(CalleeF, ArgsV, "calltmp");
}

Function *PrototypeAST::codegen() {
    // Make the function type: double(double,double) etc.
    std::vector<Type *> Doubles(Args.size(), Type::getDoubleTy(TheContext));
    FunctionType *FT =
        FunctionType::get(Type::getDoubleTy(TheContext), Doubles, false);

    Function *F =
        Function::Create(FT, Function::ExternalLinkage, Name, TheModule.get());

    // Set names for all arguments.
    unsigned Idx = 0;
    for (auto &Arg : F->args())
        Arg.setName(Args[Idx++]);
}

```

```

    return F;
}

Function *FunctionAST::codegen() {
    // First, check for an existing function from a previous 'extern' declaration.
    Function *TheFunction = TheModule->getFunction(Proto->getName());

    if (!TheFunction)
        TheFunction = Proto->codegen();

    if (!TheFunction)
        return nullptr;

    // Create a new basic block to start insertion into.
    BasicBlock *BB = BasicBlock::Create(TheContext, "entry", TheFunction);
    Builder.SetInsertPoint(BB);

    // Record the function arguments in the NamedValues map.
    NamedValues.clear();
    for (auto &Arg : TheFunction->args())
        NamedValues[Arg.getName()] = &Arg;

    if (Value *RetVal = Body->codegen()) {
        // Finish off the function.
        Builder.CreateRet(RetVal);

        // Validate the generated code, checking for consistency.
        verifyFunction(*TheFunction);

        return TheFunction;
    }

    // Error reading body, remove function.
    TheFunction->eraseFromParent();
    return nullptr;
}

//=====//
// Top-Level parsing and JIT Driver
//=====//

static void HandleDefinition() {
    if (auto FnAST = ParseDefinition()) {
        if (auto *FnIR = FnAST->codegen()) {
            fprintf(stderr, "Read function definition:");
            FnIR->print(errs());
            fprintf(stderr, "\n");
        }
    } else {
        // Skip token for error recovery.
        getNextToken();
    }
}

static void HandleExtern() {
    if (auto ProtoAST = ParseExtern()) {
        if (auto *FnIR = ProtoAST->codegen()) {

```

```

        fprintf(stderr, "Read extern: ");
        FnIR->print(errs());
        fprintf(stderr, "\n");
    }
} else {
    // Skip token for error recovery.
    getNextToken();
}
}

static void HandleTopLevelExpression() {
    // Evaluate a top-level expression into an anonymous function.
    if (auto FnAST = ParseTopLevelExpr()) {
        if (auto *FnIR = FnAST->codegen()) {
            fprintf(stderr, "Read top-level expression:");
            FnIR->print(errs());
            fprintf(stderr, "\n");
        }
    } else {
        // Skip token for error recovery.
        getNextToken();
    }
}

/// top ::= definition | external | expression | ';'
static void MainLoop() {
    while (true) {
        fprintf(stderr, "ready> ");
        switch (CurTok) {
            case tok_eof:
                return;
            case ';': // ignore top-level semicolons.
                getNextToken();
                break;
            case tok_def:
                HandleDefinition();
                break;
            case tok_extern:
                HandleExtern();
                break;
            default:
                HandleTopLevelExpression();
                break;
        }
    }
}

//===-----//
// Main driver code.
//===-----//

int main() {
    // Install standard binary operators.
    // 1 is lowest precedence.
    BinopPrecedence['<'] = 10;
    BinopPrecedence['+'] = 20;
    BinopPrecedence['-'] = 20;
    BinopPrecedence['*'] = 40; // highest.

```

```
// Prime the first token.
fprintf(stderr, "ready> ");
getNextToken();

// Make the module, which holds all the code.
TheModule = llvm::make_unique<Module>("my cool jit", TheContext);

// Run the main "interpreter loop" now.
MainLoop();

// Print out all of the generated code.
TheModule->print(errs(), nullptr);

return 0;
}
```

Next: Adding JIT and Optimizer Support

Kaleidoscope: Adding JIT and Optimizer Support

- *Chapter 4 Introduction*
- *Trivial Constant Folding*
- *LLVM Optimization Passes*
- *Adding a JIT Compiler*
- *Full Code Listing*

Chapter 4 Introduction

Welcome to Chapter 4 of the “Implementing a language with LLVM” tutorial. Chapters 1-3 described the implementation of a simple language and added support for generating LLVM IR. This chapter describes two new techniques: adding optimizer support to your language, and adding JIT compiler support. These additions will demonstrate how to get nice, efficient code for the Kaleidoscope language.

Trivial Constant Folding

Our demonstration for Chapter 3 is elegant and easy to extend. Unfortunately, it does not produce wonderful code. The IRBuilder, however, does give us obvious optimizations when compiling simple code:

```
ready> def test(x) 1+2+x;
Read function definition:
define double @test(double %x) {
entry:
    %addtmp = fadd double 3.000000e+00, %x
    ret double %addtmp
}
```

This code is not a literal transcription of the AST built by parsing the input. That would be:


```

ready> def test(x) 1+2+x;
Read function definition:
define double @test(double %x) {
entry:
    %addtmp = fadd double 2.000000e+00, 1.000000e+00
    %addtmp1 = fadd double %addtmp, %x
    ret double %addtmp1
}

```

Constant folding, as seen above, in particular, is a very common and very important optimization: so much so that many language implementors implement constant folding support in their AST representation.

With LLVM, you don't need this support in the AST. Since all calls to build LLVM IR go through the LLVM IR builder, the builder itself checked to see if there was a constant folding opportunity when you call it. If so, it just does the constant fold and return the constant instead of creating an instruction.

Well, that was easy :). In practice, we recommend always using `IRBuilder` when generating code like this. It has no “syntactic overhead” for its use (you don't have to uglify your compiler with constant checks everywhere) and it can dramatically reduce the amount of LLVM IR that is generated in some cases (particular for languages with a macro preprocessor or that use a lot of constants).

On the other hand, the `IRBuilder` is limited by the fact that it does all of its analysis inline with the code as it is built. If you take a slightly more complex example:

```

ready> def test(x) (1+2+x)*(x+(1+2));
ready> Read function definition:
define double @test(double %x) {
entry:
    %addtmp = fadd double 3.000000e+00, %x
    %addtmp1 = fadd double %x, 3.000000e+00
    %multtmp = fmul double %addtmp, %addtmp1
    ret double %multtmp
}

```

In this case, the LHS and RHS of the multiplication are the same value. We'd really like to see this generate “`tmp = x+3; result = tmp*tmp;`” instead of computing “`x+3`” twice.

Unfortunately, no amount of local analysis will be able to detect and correct this. This requires two transformations: reassociation of expressions (to make the add's lexically identical) and Common Subexpression Elimination (CSE) to delete the redundant add instruction. Fortunately, LLVM provides a broad range of optimizations that you can use, in the form of “passes”.

LLVM Optimization Passes

LLVM provides many optimization passes, which do many different sorts of things and have different tradeoffs. Unlike other systems, LLVM doesn't hold to the mistaken notion that one set of optimizations is right for all languages and for all situations. LLVM allows a compiler implementor to make complete decisions about what optimizations to use, in which order, and in what situation.

As a concrete example, LLVM supports both “whole module” passes, which look across as large of body of code as they can (often a whole file, but if run at link time, this can be a substantial portion of the whole program). It also supports and includes “per-function” passes which just operate on a single function at a time, without looking at other functions. For more information on passes and how they are run, see the [How to Write a Pass](#) document and the [List of LLVM Passes](#).

For Kaleidoscope, we are currently generating functions on the fly, one at a time, as the user types them in. We aren't shooting for the ultimate optimization experience in this setting, but we also want to catch the easy and quick stuff

where possible. As such, we will choose to run a few per-function optimizations as the user types the function in. If we wanted to make a “static Kaleidoscope compiler”, we would use exactly the code we have now, except that we would defer running the optimizer until the entire file has been parsed.

In order to get per-function optimizations going, we need to set up a `FunctionPassManager` to hold and organize the LLVM optimizations that we want to run. Once we have that, we can add a set of optimizations to run. We’ll need a new `FunctionPassManager` for each module that we want to optimize, so we’ll write a function to create and initialize both the module and pass manager for us:

```
void InitializeModuleAndPassManager(void) {
    // Open a new module.
    TheModule = llvm::make_unique<Module>("my cool jit", TheContext);

    // Create a new pass manager attached to it.
    TheFPM = llvm::make_unique<FunctionPassManager>(TheModule.get());

    // Do simple "peephole" optimizations and bit-twiddling optzns.
    TheFPM->add(createInstructionCombiningPass());
    // Reassociate expressions.
    TheFPM->add(createReassociatePass());
    // Eliminate Common SubExpressions.
    TheFPM->add(createGVNPass());
    // Simplify the control flow graph (deleting unreachable blocks, etc).
    TheFPM->add(createCFGSimplificationPass());

    TheFPM->doInitialization();
}
```

This code initializes the global module `TheModule`, and the function pass manager `TheFPM`, which is attached to `TheModule`. Once the pass manager is set up, we use a series of “add” calls to add a bunch of LLVM passes.

In this case, we choose to add four optimization passes. The passes we choose here are a pretty standard set of “cleanup” optimizations that are useful for a wide variety of code. I won’t delve into what they do but, believe me, they are a good starting place :).

Once the `PassManager` is set up, we need to make use of it. We do this by running it after our newly created function is constructed (in `FunctionAST::codegen()`), but before it is returned to the client:

```
if (Value *RetVal = Body->codegen()) {
    // Finish off the function.
    Builder.CreateRet(RetVal);

    // Validate the generated code, checking for consistency.
    verifyFunction(*TheFunction);

    // Optimize the function.
    TheFPM->run(*TheFunction);

    return TheFunction;
}
```

As you can see, this is pretty straightforward. The `FunctionPassManager` optimizes and updates the LLVM `Function*` in place, improving (hopefully) its body. With this in place, we can try our test above again:

```
ready> def test(x) (1+2+x)*(x+(1+2));
ready> Read function definition:
define double @test(double %x) {
entry:
    %addtmp = fadd double %x, 3.000000e+00
```

```

    %multmp = fmul double %addtmp, %addtmp
    ret double %multmp
}

```

As expected, we now get our nicely optimized code, saving a floating point add instruction from every execution of this function.

LLVM provides a wide variety of optimizations that can be used in certain circumstances. Some [documentation about the various passes](#) is available, but it isn't very complete. Another good source of ideas can come from looking at the passes that Clang runs to get started. The “opt” tool allows you to experiment with passes from the command line, so you can see if they do anything.

Now that we have reasonable code coming out of our front-end, let's talk about executing it!

Adding a JIT Compiler

Code that is available in LLVM IR can have a wide variety of tools applied to it. For example, you can run optimizations on it (as we did above), you can dump it out in textual or binary forms, you can compile the code to an assembly file (.s) for some target, or you can JIT compile it. The nice thing about the LLVM IR representation is that it is the “common currency” between many different parts of the compiler.

In this section, we'll add JIT compiler support to our interpreter. The basic idea that we want for Kaleidoscope is to have the user enter function bodies as they do now, but immediately evaluate the top-level expressions they type in. For example, if they type in “1 + 2;”, we should evaluate and print out 3. If they define a function, they should be able to call it from the command line.

In order to do this, we first prepare the environment to create code for the current native target and declare and initialize the JIT. This is done by calling some `InitializeNativeTarget*` functions and adding a global variable `TheJIT`, and initializing it in `main`:

```

static std::unique_ptr<KaleidoscopeJIT> TheJIT;
...
int main() {
    InitializeNativeTarget();
    InitializeNativeTargetAsmPrinter();
    InitializeNativeTargetAsmParser();

    // Install standard binary operators.
    // 1 is lowest precedence.
    BinopPrecedence['<'] = 10;
    BinopPrecedence['+'] = 20;
    BinopPrecedence['-'] = 20;
    BinopPrecedence['*'] = 40; // highest.

    // Prime the first token.
    fprintf(stderr, "ready> ");
    getNextToken();

    TheJIT = llvm::make_unique<KaleidoscopeJIT>();

    // Run the main "interpreter loop" now.
    MainLoop();

    return 0;
}

```

We also need to setup the data layout for the JIT:

```
void InitializeModuleAndPassManager(void) {  
    // Open a new module.  
    TheModule = llvm::make_unique<Module>("my cool jit", TheContext);  
    TheModule->setDataLayout(TheJIT->getTargetMachine().createDataLayout());  
  
    // Create a new pass manager attached to it.  
    TheFPM = llvm::make_unique<FunctionPassManager>(TheModule.get());  
    ...  
}
```

The `KaleidoscopeJIT` class is a simple JIT built specifically for these tutorials, available inside the LLVM source code at `llvm-src/examples/Kaleidoscope/include/KaleidoscopeJIT.h`. In later chapters we will look at how it works and extend it with new features, but for now we will take it as given. Its API is very simple: `addModule` adds an LLVM IR module to the JIT, making its functions available for execution; `removeModule` removes a module, freeing any memory associated with the code in that module; and `findSymbol` allows us to look up pointers to the compiled code.

We can take this simple API and change our code that parses top-level expressions to look like this:

```
static void HandleTopLevelExpression() {  
    // Evaluate a top-level expression into an anonymous function.  
    if (auto FnAST = ParseTopLevelExpr()) {  
        if (FnAST->codegen()) {  
  
            // JIT the module containing the anonymous expression, keeping a handle so  
            // we can free it later.  
            auto H = TheJIT->addModule(std::move(TheModule));  
            InitializeModuleAndPassManager();  
  
            // Search the JIT for the __anon_expr symbol.  
            auto ExprSymbol = TheJIT->findSymbol("__anon_expr");  
            assert(ExprSymbol && "Function not found");  
  
            // Get the symbol's address and cast it to the right type (takes no  
            // arguments, returns a double) so we can call it as a native function.  
            double (*FP)() = (double (*)(intptr_t))ExprSymbol.getAddress();  
            fprintf(stderr, "Evaluated to %f\n", FP());  
  
            // Delete the anonymous expression module from the JIT.  
            TheJIT->removeModule(H);  
        }  
    }  
}
```

If parsing and codegen succeed, the next step is to add the module containing the top-level expression to the JIT. We do this by calling `addModule`, which triggers code generation for all the functions in the module, and returns a handle that can be used to remove the module from the JIT later. Once the module has been added to the JIT it can no longer be modified, so we also open a new module to hold subsequent code by calling `InitializeModuleAndPassManager()`.

Once we've added the module to the JIT we need to get a pointer to the final generated code. We do this by calling the JIT's `findSymbol` method, and passing the name of the top-level expression function: `__anon_expr`. Since we just added this function, we assert that `findSymbol` returned a result.

Next, we get the in-memory address of the `__anon_expr` function by calling `getAddress()` on the symbol. Recall that we compile top-level expressions into a self-contained LLVM function that takes no arguments and returns the computed double. Because the LLVM JIT compiler matches the native platform ABI, this means that you can just cast the result pointer to a function pointer of that type and call it directly. This means, there is no difference between JIT compiled code and native machine code that is statically linked into your application.

Finally, since we don't support re-evaluation of top-level expressions, we remove the module from the JIT when

we're done to free the associated memory. Recall, however, that the module we created a few lines earlier (via `InitializeModuleAndPassManager`) is still open and waiting for new code to be added.

With just these two changes, let's see how Kaleidoscope works now!

```
ready> 4+5;
Read top-level expression:
define double @0() {
entry:
    ret double 9.000000e+00
}

Evaluated to 9.000000
```

Well this looks like it is basically working. The dump of the function shows the “no argument function that always returns double” that we synthesize for each top-level expression that is typed in. This demonstrates very basic functionality, but can we do more?

```
ready> def testfunc(x y) x + y*2;
Read function definition:
define double @testfunc(double %x, double %y) {
entry:
    %multmp = fmul double %y, 2.000000e+00
    %addtmp = fadd double %multmp, %x
    ret double %addtmp
}

ready> testfunc(4, 10);
Read top-level expression:
define double @1() {
entry:
    %calltmp = call double @testfunc(double 4.000000e+00, double 1.000000e+01)
    ret double %calltmp
}

Evaluated to 24.000000

ready> testfunc(5, 10);
ready> LLVM ERROR: Program used external function 'testfunc' which could not be
↳resolved!
```

Function definitions and calls also work, but something went very wrong on that last line. The call looks valid, so what happened? As you may have guessed from the the API a Module is a unit of allocation for the JIT, and `testfunc` was part of the same module that contained anonymous expression. When we removed that module from the JIT to free the memory for the anonymous expression, we deleted the definition of `testfunc` along with it. Then, when we tried to call `testfunc` a second time, the JIT could no longer find it.

The easiest way to fix this is to put the anonymous expression in a separate module from the rest of the function definitions. The JIT will happily resolve function calls across module boundaries, as long as each of the functions called has a prototype, and is added to the JIT before it is called. By putting the anonymous expression in a different module we can delete it without affecting the rest of the functions.

In fact, we're going to go a step further and put every function in its own module. Doing so allows us to exploit a useful property of the KaleidoscopeJIT that will make our environment more REPL-like: Functions can be added to the JIT more than once (unlike a module where every function must have a unique definition). When you look up a symbol in KaleidoscopeJIT it will always return the most recent definition:

```
ready> def foo(x) x + 1;
Read function definition:
define double @foo(double %x) {
entry:
    %addtmp = fadd double %x, 1.000000e+00
    ret double %addtmp
}

ready> foo(2);
Evaluated to 3.000000

ready> def foo(x) x + 2;
define double @foo(double %x) {
entry:
    %addtmp = fadd double %x, 2.000000e+00
    ret double %addtmp
}

ready> foo(2);
Evaluated to 4.000000
```

To allow each function to live in its own module we'll need a way to re-generate previous function declarations into each new module we open:

```
static std::unique_ptr<KaleidoscopeJIT> TheJIT;

...

Function *getFunction(std::string Name) {
    // First, see if the function has already been added to the current module.
    if (auto *F = TheModule->getFunction(Name))
        return F;

    // If not, check whether we can codegen the declaration from some existing
    // prototype.
    auto FI = FunctionProtos.find(Name);
    if (FI != FunctionProtos.end())
        return FI->second->codegen();

    // If no existing prototype exists, return null.
    return nullptr;
}

...

Value *CallExprAST::codegen() {
    // Look up the name in the global module table.
    Function *CalleeF = getFunction(Callee);

    ...

Function *FunctionAST::codegen() {
    // Transfer ownership of the prototype to the FunctionProtos map, but keep a
    // reference to it for use below.
    auto &P = *Proto;
    FunctionProtos[Proto->getName()] = std::move(Proto);
    Function *TheFunction = getFunction(P.getName());
    if (!TheFunction)
```

```
return nullptr;
```

To enable this, we'll start by adding a new global, `FunctionProtos`, that holds the most recent prototype for each function. We'll also add a convenience method, `getFunction()`, to replace calls to `TheModule->getFunction()`. Our convenience method searches `TheModule` for an existing function declaration, falling back to generating a new declaration from `FunctionProtos` if it doesn't find one. In `CallExprAST::codegen()` we just need to replace the call to `TheModule->getFunction()`. In `FunctionAST::codegen()` we need to update the `FunctionProtos` map first, then call `getFunction()`. With this done, we can always obtain a function declaration in the current module for any previously declared function.

We also need to update `HandleDefinition` and `HandleExtern`:

```
static void HandleDefinition() {
    if (auto FnAST = ParseDefinition()) {
        if (auto *FnIR = FnAST->codegen()) {
            fprintf(stderr, "Read function definition:");
            FnIR->print(errs());
            fprintf(stderr, "\n");
            TheJIT->addModule(std::move(TheModule));
            InitializeModuleAndPassManager();
        }
    } else {
        // Skip token for error recovery.
        getNextToken();
    }
}

static void HandleExtern() {
    if (auto ProtoAST = ParseExtern()) {
        if (auto *FnIR = ProtoAST->codegen()) {
            fprintf(stderr, "Read extern: ");
            FnIR->print(errs());
            fprintf(stderr, "\n");
            FunctionProtos[ProtoAST->getName()] = std::move(ProtoAST);
        }
    } else {
        // Skip token for error recovery.
        getNextToken();
    }
}
```

In `HandleDefinition`, we add two lines to transfer the newly defined function to the JIT and open a new module. In `HandleExtern`, we just need to add one line to add the prototype to `FunctionProtos`.

With these changes made, let's try our REPL again (I removed the dump of the anonymous functions this time, you should get the idea by now):

```
ready> def foo(x) x + 1;
ready> foo(2);
Evaluated to 3.000000

ready> def foo(x) x + 2;
ready> foo(2);
Evaluated to 4.000000
```

It works!

Even with this simple code, we get some surprisingly powerful capabilities - check this out:

```
ready> extern sin(x);
Read extern:
declare double @sin(double)

ready> extern cos(x);
Read extern:
declare double @cos(double)

ready> sin(1.0);
Read top-level expression:
define double @2() {
entry:
    ret double 0x3FEAED548F090CEE
}

Evaluated to 0.841471

ready> def foo(x) sin(x)*sin(x) + cos(x)*cos(x);
Read function definition:
define double @foo(double %x) {
entry:
    %calltmp = call double @sin(double %x)
    %multmp = fmul double %calltmp, %calltmp
    %calltmp2 = call double @cos(double %x)
    %multmp4 = fmul double %calltmp2, %calltmp2
    %addtmp = fadd double %multmp, %multmp4
    ret double %addtmp
}

ready> foo(4.0);
Read top-level expression:
define double @3() {
entry:
    %calltmp = call double @foo(double 4.000000e+00)
    ret double %calltmp
}

Evaluated to 1.000000
```

Whoa, how does the JIT know about sin and cos? The answer is surprisingly simple: The KaleidoscopeJIT has a straightforward symbol resolution rule that it uses to find symbols that aren't available in any given module: First it searches all the modules that have already been added to the JIT, from the most recent to the oldest, to find the newest definition. If no definition is found inside the JIT, it falls back to calling `dlsym("sin")` on the Kaleidoscope process itself. Since `sin` is defined within the JIT's address space, it simply patches up calls in the module to call the libm version of `sin` directly. But in some cases this even goes further: as `sin` and `cos` are names of standard math functions, the constant folder will directly evaluate the function calls to the correct result when called with constants like in the `sin(1.0)` above.

In the future we'll see how tweaking this symbol resolution rule can be used to enable all sorts of useful features, from security (restricting the set of symbols available to JIT'd code), to dynamic code generation based on symbol names, and even lazy compilation.

One immediate benefit of the symbol resolution rule is that we can now extend the language by writing arbitrary C++ code to implement operations. For example, if we add:

```
#ifdef LLVM_ON_WIN32
#define DLLEXPORT __declspec(dllexport)
#else
```



```

#define DLLEXPORT
#endif

/// putchar - putchar that takes a double and returns 0.
extern "C" DLLEXPORT double putchar(double X) {
    fputc((char)X, stderr);
    return 0;
}

```

Note, that for Windows we need to actually export the functions because the dynamic symbol loader will use `GetProcAddress` to find the symbols.

Now we can produce simple output to the console by using things like: `extern putchar(x); putchar(120);`, which prints a lowercase ‘x’ on the console (120 is the ASCII code for ‘x’). Similar code could be used to implement file I/O, console input, and many other capabilities in Kaleidoscope.

This completes the JIT and optimizer chapter of the Kaleidoscope tutorial. At this point, we can compile a non-Turing-complete programming language, optimize and JIT compile it in a user-driven way. Next up we’ll look into [extending the language with control flow constructs](#), tackling some interesting LLVM IR issues along the way.

Full Code Listing

Here is the complete code listing for our running example, enhanced with the LLVM JIT and optimizer. To build this example, use:

```

# Compile
clang++ -g toy.cpp `llvm-config --cxxflags --ldflags --system-libs --libs core mcjit_
↪native` -O3 -o toy
# Run
./toy

```

If you are compiling this on Linux, make sure to add the “-rdynamic” option as well. This makes sure that the external functions are resolved properly at runtime.

Here is the code:

```

#include "llvm/ADT/APFloat.h"
#include "llvm/ADT/STLExtras.h"
#include "llvm/IR/BasicBlock.h"
#include "llvm/IR/Constants.h"
#include "llvm/IR/DerivedTypes.h"
#include "llvm/IR/Function.h"
#include "llvm/IR/IRBuilder.h"
#include "llvm/IR/LLVMContext.h"
#include "llvm/IR/LegacyPassManager.h"
#include "llvm/IR/Module.h"
#include "llvm/IR/Type.h"
#include "llvm/IR/Verifier.h"
#include "llvm/Support/TargetSelect.h"
#include "llvm/Target/TargetMachine.h"
#include "llvm/Transforms/Scalar.h"
#include "llvm/Transforms/Scalar/GVN.h"
#include "../include/KaleidoscopeJIT.h"
#include <algorithm>
#include <cassert>
#include <cctype>
#include <cstdint>

```

```
#include <stdio>
#include <stdlib>
#include <map>
#include <memory>
#include <string>
#include <vector>

using namespace llvm;
using namespace llvm::orc;

//=====//
// Lexer
//=====//

// The lexer returns tokens [0-255] if it is an unknown character, otherwise one
// of these for known things.
enum Token {
    tok_eof = -1,

    // commands
    tok_def = -2,
    tok_extern = -3,

    // primary
    tok_identifier = -4,
    tok_number = -5
};

static std::string IdentifierStr; // Filled in if tok_identifier
static double NumVal;           // Filled in if tok_number

/// gettok - Return the next token from standard input.
static int gettok() {
    static int LastChar = ' ';

    // Skip any whitespace.
    while (isspace(LastChar))
        LastChar = getchar();

    if (isalpha(LastChar)) { // identifier: [a-zA-Z][a-zA-Z0-9]*
        IdentifierStr = LastChar;
        while (isalnum((LastChar = getchar())))
            IdentifierStr += LastChar;

        if (IdentifierStr == "def")
            return tok_def;
        if (IdentifierStr == "extern")
            return tok_extern;
        return tok_identifier;
    }

    if (isdigit(LastChar) || LastChar == '.') { // Number: [0-9.]+
        std::string NumStr;
        do {
            NumStr += LastChar;
            LastChar = getchar();
        } while (isdigit(LastChar) || LastChar == '.');
```

```

    NumVal = strtod(NumStr.c_str(), nullptr);
    return tok_number;
}

if (LastChar == '#') {
    // Comment until end of line.
    do
        LastChar = getchar();
    while (LastChar != EOF && LastChar != '\n' && LastChar != '\r');

    if (LastChar != EOF)
        return gettok();
}

// Check for end of file. Don't eat the EOF.
if (LastChar == EOF)
    return tok_eof;

// Otherwise, just return the character as its ascii value.
int ThisChar = LastChar;
LastChar = getchar();
return ThisChar;
}

//===-----
// Abstract Syntax Tree (aka Parse Tree)
//===-----

namespace {

/// ExprAST - Base class for all expression nodes.
class ExprAST {
public:
    virtual ~ExprAST() = default;

    virtual Value *codegen() = 0;
};

/// NumberExprAST - Expression class for numeric literals like "1.0".
class NumberExprAST : public ExprAST {
    double Val;

public:
    NumberExprAST(double Val) : Val(Val) {}

    Value *codegen() override;
};

/// VariableExprAST - Expression class for referencing a variable, like "a".
class VariableExprAST : public ExprAST {
    std::string Name;

public:
    VariableExprAST(const std::string &Name) : Name(Name) {}

    Value *codegen() override;
};

```

```

/// BinaryExprAST - Expression class for a binary operator.
class BinaryExprAST : public ExprAST {
    char Op;
    std::unique_ptr<ExprAST> LHS, RHS;

public:
    BinaryExprAST(char Op, std::unique_ptr<ExprAST> LHS,
                  std::unique_ptr<ExprAST> RHS)
        : Op(Op), LHS(std::move(LHS)), RHS(std::move(RHS)) {}

    Value *codegen() override;
};

/// CallExprAST - Expression class for function calls.
class CallExprAST : public ExprAST {
    std::string Callee;
    std::vector<std::unique_ptr<ExprAST>> Args;

public:
    CallExprAST(const std::string &Callee,
                std::vector<std::unique_ptr<ExprAST>> Args)
        : Callee(Callee), Args(std::move(Args)) {}

    Value *codegen() override;
};

/// PrototypeAST - This class represents the "prototype" for a function,
/// which captures its name, and its argument names (thus implicitly the number
/// of arguments the function takes).
class PrototypeAST {
    std::string Name;
    std::vector<std::string> Args;

public:
    PrototypeAST(const std::string &Name, std::vector<std::string> Args)
        : Name(Name), Args(std::move(Args)) {}

    Function *codegen();
    const std::string &getName() const { return Name; }
};

/// FunctionAST - This class represents a function definition itself.
class FunctionAST {
    std::unique_ptr<PrototypeAST> Proto;
    std::unique_ptr<ExprAST> Body;

public:
    FunctionAST(std::unique_ptr<PrototypeAST> Proto,
                std::unique_ptr<ExprAST> Body)
        : Proto(std::move(Proto)), Body(std::move(Body)) {}

    Function *codegen();
};

} // end anonymous namespace

//=====
// Parser

```

```
//=====//

// CurTok/getNextToken - Provide a simple token buffer. CurTok is the current
// token the parser is looking at. getNextToken reads another token from the
// lexer and updates CurTok with its results.
static int CurTok;
static int getNextToken() { return CurTok = gettok(); }

// BinopPrecedence - This holds the precedence for each binary operator that is
// defined.
static std::map<char, int> BinopPrecedence;

// GetTokPrecedence - Get the precedence of the pending binary operator token.
static int GetTokPrecedence() {
    if (!isascii(CurTok))
        return -1;

    // Make sure it's a declared binop.
    int TokPrec = BinopPrecedence[CurTok];
    if (TokPrec <= 0)
        return -1;
    return TokPrec;
}

// LogError* - These are little helper functions for error handling.
std::unique_ptr<ExprAST> LogError(const char *Str) {
    fprintf(stderr, "Error: %s\n", Str);
    return nullptr;
}

std::unique_ptr<PrototypeAST> LogErrorP(const char *Str) {
    LogError(Str);
    return nullptr;
}

static std::unique_ptr<ExprAST> ParseExpression();

// numberexpr ::= number
static std::unique_ptr<ExprAST> ParseNumberExpr() {
    auto Result = llvm::make_unique<NumberExprAST>(NumVal);
    getNextToken(); // consume the number
    return std::move(Result);
}

// parenexpr ::= '(' expression ')'
static std::unique_ptr<ExprAST> ParseParenExpr() {
    getNextToken(); // eat (.
    auto V = ParseExpression();
    if (!V)
        return nullptr;

    if (CurTok != ')')
        return LogError("expected ')'");
    getNextToken(); // eat ).
    return V;
}

// identifierexpr
```

```

/// ::= identifier
/// ::= identifier '(' expression* ')'
static std::unique_ptr<ExprAST> ParseIdentifierExpr() {
    std::string IdName = IdentifierStr;

    getNextToken(); // eat identifier.

    if (CurTok != '(') // Simple variable ref.
        return llvm::make_unique<VariableExprAST>(IdName);

    // Call.
    getNextToken(); // eat (
    std::vector<std::unique_ptr<ExprAST>> Args;
    if (CurTok != ')') {
        while (true) {
            if (auto Arg = ParseExpression())
                Args.push_back(std::move(Arg));
            else
                return nullptr;

            if (CurTok == ')')
                break;

            if (CurTok != ',')
                return LogError("Expected ')' or ',' in argument list");
            getNextToken();
        }
    }

    // Eat the ')'.
    getNextToken();

    return llvm::make_unique<CallExprAST>(IdName, std::move(Args));
}

/// primary
/// ::= identifierexpr
/// ::= numberexpr
/// ::= parenexpr
static std::unique_ptr<ExprAST> ParsePrimary() {
    switch (CurTok) {
    default:
        return LogError("unknown token when expecting an expression");
    case tok_identifier:
        return ParseIdentifierExpr();
    case tok_number:
        return ParseNumberExpr();
    case '(':
        return ParseParenExpr();
    }
}

/// binoprhs
/// ::= ('+' primary)*
static std::unique_ptr<ExprAST> ParseBinOpRHS(int ExprPrec,
                                              std::unique_ptr<ExprAST> LHS) {
    // If this is a binop, find its precedence.
    while (true) {

```

```

    int TokPrec = GetTokPrecedence();

    // If this is a binop that binds at least as tightly as the current binop,
    // consume it, otherwise we are done.
    if (TokPrec < ExprPrec)
        return LHS;

    // Okay, we know this is a binop.
    int BinOp = CurTok;
    getNextToken(); // eat binop

    // Parse the primary expression after the binary operator.
    auto RHS = ParsePrimary();
    if (!RHS)
        return nullptr;

    // If BinOp binds less tightly with RHS than the operator after RHS, let
    // the pending operator take RHS as its LHS.
    int NextPrec = GetTokPrecedence();
    if (TokPrec < NextPrec) {
        RHS = ParseBinOpRHS(TokPrec + 1, std::move(RHS));
        if (!RHS)
            return nullptr;
    }

    // Merge LHS/RHS.
    LHS =
        llvm::make_unique<BinaryExprAST>(BinOp, std::move(LHS), std::move(RHS));
}

/// expression
/// ::= primary binoprhs
///
static std::unique_ptr<ExprAST> ParseExpression() {
    auto LHS = ParsePrimary();
    if (!LHS)
        return nullptr;

    return ParseBinOpRHS(0, std::move(LHS));
}

/// prototype
/// ::= id '(' id* ')'
static std::unique_ptr<PrototypeAST> ParsePrototype() {
    if (CurTok != tok_identifier)
        return LogErrorP("Expected function name in prototype");

    std::string FnName = IdentifierStr;
    getNextToken();

    if (CurTok != '(')
        return LogErrorP("Expected '(' in prototype");

    std::vector<std::string> ArgNames;
    while (getNextToken() == tok_identifier)
        ArgNames.push_back(IdentifierStr);
    if (CurTok != ')')

```

```

    return LogErrorP("Expected ')' in prototype");

    // success.
    getNextToken(); // eat ')'.

    return llvm::make_unique<PrototypeAST>(FnName, std::move(ArgNames));
}

// definition ::= 'def' prototype expression
static std::unique_ptr<FunctionAST> ParseDefinition() {
    getNextToken(); // eat def.
    auto Proto = ParsePrototype();
    if (!Proto)
        return nullptr;

    if (auto E = ParseExpression())
        return llvm::make_unique<FunctionAST>(std::move(Proto), std::move(E));
    return nullptr;
}

// toplevelexpr ::= expression
static std::unique_ptr<FunctionAST> ParseTopLevelExpr() {
    if (auto E = ParseExpression()) {
        // Make an anonymous proto.
        auto Proto = llvm::make_unique<PrototypeAST>("__anon_expr",
                                                    std::vector<std::string>());
        return llvm::make_unique<FunctionAST>(std::move(Proto), std::move(E));
    }
    return nullptr;
}

// external ::= 'extern' prototype
static std::unique_ptr<PrototypeAST> ParseExtern() {
    getNextToken(); // eat extern.
    return ParsePrototype();
}

//===-----
// Code Generation
//===-----

static LLVMContext TheContext;
static IRBuilder<> Builder(TheContext);
static std::unique_ptr<Module> TheModule;
static std::map<std::string, Value*> NamedValues;
static std::unique_ptr<legacy::FunctionPassManager> TheFPM;
static std::unique_ptr<KaleidoscopeJIT> TheJIT;
static std::map<std::string, std::unique_ptr<PrototypeAST>> FunctionProtos;

Value *LogErrorV(const char *Str) {
    LogError(Str);
    return nullptr;
}

Function *getFunction(std::string Name) {
    // First, see if the function has already been added to the current module.
    if (auto *F = TheModule->getFunction(Name))
        return F;

```



```

// If not, check whether we can codegen the declaration from some existing
// prototype.
auto FI = FunctionProtos.find(Name);
if (FI != FunctionProtos.end())
    return FI->second->codegen();

// If no existing prototype exists, return null.
return nullptr;
}

Value *NumberExprAST::codegen() {
    return ConstantFP::get(TheContext, APFloat(Val));
}

Value *VariableExprAST::codegen() {
    // Look this variable up in the function.
    Value *V = NamedValues[Name];
    if (!V)
        return LogErrorV("Unknown variable name");
    return V;
}

Value *BinaryExprAST::codegen() {
    Value *L = LHS->codegen();
    Value *R = RHS->codegen();
    if (!L || !R)
        return nullptr;

    switch (Op) {
    case '+':
        return Builder.CreateFAdd(L, R, "addtmp");
    case '-':
        return Builder.CreateFSub(L, R, "subtmp");
    case '*':
        return Builder.CreateFMul(L, R, "multtmp");
    case '<':
        L = Builder.CreateFCmpULT(L, R, "cmptmp");
        // Convert bool 0/1 to double 0.0 or 1.0
        return Builder.CreateUIToFP(L, Type::getDoubleTy(TheContext), "booltmp");
    default:
        return LogErrorV("invalid binary operator");
    }
}

Value *CallExprAST::codegen() {
    // Look up the name in the global module table.
    Function *CalleeF = getFunction(Callee);
    if (!CalleeF)
        return LogErrorV("Unknown function referenced");

    // If argument mismatch error.
    if (CalleeF->arg_size() != Args.size())
        return LogErrorV("Incorrect # arguments passed");

    std::vector<Value *> ArgsV;
    for (unsigned i = 0, e = Args.size(); i != e; ++i) {
        ArgsV.push_back(Args[i]->codegen());
    }
}

```

```

    if (!ArgsV.back())
        return nullptr;
}

return Builder.CreateCall(CalleeF, ArgsV, "calltmp");
}

Function *PrototypeAST::codegen() {
    // Make the function type: double(double,double) etc.
    std::vector<Type *> Doubles(Args.size(), Type::getDoubleTy(TheContext));
    FunctionType *FT =
        FunctionType::get(Type::getDoubleTy(TheContext), Doubles, false);

    Function *F =
        Function::Create(FT, Function::ExternalLinkage, Name, TheModule.get());

    // Set names for all arguments.
    unsigned Idx = 0;
    for (auto &Arg : F->args())
        Arg.setName(Args[Idx++]);

    return F;
}

Function *FunctionAST::codegen() {
    // Transfer ownership of the prototype to the FunctionProtos map, but keep a
    // reference to it for use below.
    auto &P = *Proto;
    FunctionProtos[Proto->getName()] = std::move(Proto);
    Function *TheFunction = getFunction(P.getName());
    if (!TheFunction)
        return nullptr;

    // Create a new basic block to start insertion into.
    BasicBlock *BB = BasicBlock::Create(TheContext, "entry", TheFunction);
    Builder.SetInsertPoint(BB);

    // Record the function arguments in the NamedValues map.
    NamedValues.clear();
    for (auto &Arg : TheFunction->args())
        NamedValues[Arg.getName()] = &Arg;

    if (Value *RetVal = Body->codegen()) {
        // Finish off the function.
        Builder.CreateRet(RetVal);

        // Validate the generated code, checking for consistency.
        verifyFunction(*TheFunction);

        // Run the optimizer on the function.
        TheFPM->run(*TheFunction);

        return TheFunction;
    }

    // Error reading body, remove function.
    TheFunction->eraseFromParent();
    return nullptr;
}

```

```

}

//===-----//
// Top-Level parsing and JIT Driver
//===-----//

static void InitializeModuleAndPassManager() {
    // Open a new module.
    TheModule = llvm::make_unique<Module>("my cool jit", TheContext);
    TheModule->setDataLayout(TheJIT->getTargetMachine().createDataLayout());

    // Create a new pass manager attached to it.
    TheFPM = llvm::make_unique<legacy::FunctionPassManager>(TheModule.get());

    // Do simple "peephole" optimizations and bit-twiddling optzns.
    TheFPM->add(createInstructionCombiningPass());
    // Reassociate expressions.
    TheFPM->add(createReassociatePass());
    // Eliminate Common SubExpressions.
    TheFPM->add(createGVNPass());
    // Simplify the control flow graph (deleting unreachable blocks, etc).
    TheFPM->add(createCFGSimplificationPass());

    TheFPM->doInitialization();
}

static void HandleDefinition() {
    if (auto FnAST = ParseDefinition()) {
        if (auto *FnIR = FnAST->codegen()) {
            fprintf(stderr, "Read function definition:");
            FnIR->print(errs());
            fprintf(stderr, "\n");
            TheJIT->addModule(std::move(TheModule));
            InitializeModuleAndPassManager();
        }
    } else {
        // Skip token for error recovery.
        getNextToken();
    }
}

static void HandleExtern() {
    if (auto ProtoAST = ParseExtern()) {
        if (auto *FnIR = ProtoAST->codegen()) {
            fprintf(stderr, "Read extern: ");
            FnIR->print(errs());
            fprintf(stderr, "\n");
            FunctionProtos[ProtoAST->getName()] = std::move(ProtoAST);
        }
    } else {
        // Skip token for error recovery.
        getNextToken();
    }
}

static void HandleTopLevelExpression() {
    // Evaluate a top-level expression into an anonymous function.
    if (auto FnAST = ParseTopLevelExpr()) {

```

```

    if (FnAST->codegen()) {
        // JIT the module containing the anonymous expression, keeping a handle so
        // we can free it later.
        auto H = TheJIT->addModule(std::move(TheModule));
        InitializeModuleAndPassManager();

        // Search the JIT for the __anon_expr symbol.
        auto ExprSymbol = TheJIT->findSymbol("__anon_expr");
        assert(ExprSymbol && "Function not found");

        // Get the symbol's address and cast it to the right type (takes no
        // arguments, returns a double) so we can call it as a native function.
        double (*FP)() = (double (*)(intptr_t))cantFail(ExprSymbol.getAddress());
        fprintf(stderr, "Evaluated to %f\n", FP());

        // Delete the anonymous expression module from the JIT.
        TheJIT->removeModule(H);
    }
} else {
    // Skip token for error recovery.
    getNextToken();
}
}

/// top ::= definition | external | expression | ';'
static void MainLoop() {
    while (true) {
        fprintf(stderr, "ready> ");
        switch (CurTok) {
            case tok_eof:
                return;
            case ';': // ignore top-level semicolons.
                getNextToken();
                break;
            case tok_def:
                HandleDefinition();
                break;
            case tok_extern:
                HandleExtern();
                break;
            default:
                HandleTopLevelExpression();
                break;
        }
    }
}

//===-----
// "Library" functions that can be "extern'd" from user code.
//===-----

#ifdef LLVM_ON_WIN32
#define DLLEXPORT __declspec(dllexport)
#else
#define DLLEXPORT
#endif

/// putchar - putchar that takes a double and returns 0.

```

```

extern "C" DLLEXPORT double putchar(double X) {
    fputc((char)X, stderr);
    return 0;
}

/// printf - printf that takes a double prints it as "%f\n", returning 0.
extern "C" DLLEXPORT double printf(double X) {
    fprintf(stderr, "%f\n", X);
    return 0;
}

//===-----
// Main driver code.
//===-----

int main() {
    InitializeNativeTarget();
    InitializeNativeTargetAsmPrinter();
    InitializeNativeTargetAsmParser();

    // Install standard binary operators.
    // 1 is lowest precedence.
    BinopPrecedence['<'] = 10;
    BinopPrecedence['+'] = 20;
    BinopPrecedence['-'] = 20;
    BinopPrecedence['*'] = 40; // highest.

    // Prime the first token.
    fprintf(stderr, "ready> ");
    getNextToken();

    TheJIT = llvm::make_unique<KaleidoscopeJIT>();

    InitializeModuleAndPassManager();

    // Run the main "interpreter loop" now.
    MainLoop();

    return 0;
}

```

Next: Extending the language: control flow

Kaleidoscope: Extending the Language: Control Flow

- *Chapter 5 Introduction*
- *If/Then/Else*
 - *Lexer Extensions for If/Then/Else*
 - *AST Extensions for If/Then/Else*
 - *Parser Extensions for If/Then/Else*
 - *LLVM IR for If/Then/Else*

- *Code Generation for If/Then/Else*
- *‘for’ Loop Expression*
 - *Lexer Extensions for the ‘for’ Loop*
 - *AST Extensions for the ‘for’ Loop*
 - *Parser Extensions for the ‘for’ Loop*
 - *LLVM IR for the ‘for’ Loop*
 - *Code Generation for the ‘for’ Loop*
- *Full Code Listing*

Chapter 5 Introduction

Welcome to Chapter 5 of the “[Implementing a language with LLVM](#)” tutorial. Parts 1-4 described the implementation of the simple Kaleidoscope language and included support for generating LLVM IR, followed by optimizations and a JIT compiler. Unfortunately, as presented, Kaleidoscope is mostly useless: it has no control flow other than call and return. This means that you can’t have conditional branches in the code, significantly limiting its power. In this episode of “build that compiler”, we’ll extend Kaleidoscope to have an if/then/else expression plus a simple ‘for’ loop.

If/Then/Else

Extending Kaleidoscope to support if/then/else is quite straightforward. It basically requires adding support for this “new” concept to the lexer, parser, AST, and LLVM code emitter. This example is nice, because it shows how easy it is to “grow” a language over time, incrementally extending it as new ideas are discovered.

Before we get going on “how” we add this extension, let’s talk about “what” we want. The basic idea is that we want to be able to write this sort of thing:

```
def fib(x)
  if x < 3 then
    1
  else
    fib(x-1)+fib(x-2);
```

In Kaleidoscope, every construct is an expression: there are no statements. As such, the if/then/else expression needs to return a value like any other. Since we’re using a mostly functional form, we’ll have it evaluate its conditional, then return the ‘then’ or ‘else’ value based on how the condition was resolved. This is very similar to the C “?:” expression.

The semantics of the if/then/else expression is that it evaluates the condition to a boolean equality value: 0.0 is considered to be false and everything else is considered to be true. If the condition is true, the first subexpression is evaluated and returned, if the condition is false, the second subexpression is evaluated and returned. Since Kaleidoscope allows side-effects, this behavior is important to nail down.

Now that we know what we “want”, let’s break this down into its constituent pieces.

Lexer Extensions for If/Then/Else

The lexer extensions are straightforward. First we add new enum values for the relevant tokens:

```
// control
tok_if = -6,
tok_then = -7,
tok_else = -8,
```

Once we have that, we recognize the new keywords in the lexer. This is pretty simple stuff:

```
...
if (IdentifierStr == "def")
    return tok_def;
if (IdentifierStr == "extern")
    return tok_extern;
if (IdentifierStr == "if")
    return tok_if;
if (IdentifierStr == "then")
    return tok_then;
if (IdentifierStr == "else")
    return tok_else;
return tok_identifier;
```

AST Extensions for If/Then/Else

To represent the new expression we add a new AST node for it:

```
/// IfExprAST - Expression class for if/then/else.
class IfExprAST : public ExprAST {
    std::unique_ptr<ExprAST> Cond, Then, Else;

public:
    IfExprAST(std::unique_ptr<ExprAST> Cond, std::unique_ptr<ExprAST> Then,
              std::unique_ptr<ExprAST> Else)
        : Cond(std::move(Cond)), Then(std::move(Then)), Else(std::move(Else)) {}

    Value *codegen() override;
};
```

The AST node just has pointers to the various subexpressions.

Parser Extensions for If/Then/Else

Now that we have the relevant tokens coming from the lexer and we have the AST node to build, our parsing logic is relatively straightforward. First we define a new parsing function:

```
/// ifexpr ::= 'if' expression 'then' expression 'else' expression
static std::unique_ptr<ExprAST> ParseIfExpr() {
    getNextToken(); // eat the if.

    // condition.
    auto Cond = ParseExpression();
    if (!Cond)
        return nullptr;

    if (CurTok != tok_then)
        return LogError("expected then");
    getNextToken(); // eat the then
```

```
auto Then = ParseExpression();
if (!Then)
    return nullptr;

if (CurTok != tok_else)
    return LogError("expected else");

getNextToken();

auto Else = ParseExpression();
if (!Else)
    return nullptr;

return llvm::make_unique<IfExprAST>(std::move(Cond), std::move(Then),
                                     std::move(Else));
}
```

Next we hook it up as a primary expression:

```
static std::unique_ptr<ExprAST> ParsePrimary() {
    switch (CurTok) {
    default:
        return LogError("unknown token when expecting an expression");
    case tok_identifier:
        return ParseIdentifierExpr();
    case tok_number:
        return ParseNumberExpr();
    case '(':
        return ParseParenExpr();
    case tok_if:
        return ParseIfExpr();
    }
}
```

LLVM IR for If/Then/Else

Now that we have it parsing and building the AST, the final piece is adding LLVM code generation support. This is the most interesting part of the if/then/else example, because this is where it starts to introduce new concepts. All of the code above has been thoroughly described in previous chapters.

To motivate the code we want to produce, let's take a look at a simple example. Consider:

```
extern foo();
extern bar();
def baz(x) if x then foo() else bar();
```

If you disable optimizations, the code you'll (soon) get from Kaleidoscope looks like this:

```
declare double @foo()

declare double @bar()

define double @baz(double %x) {
entry:
    %ifcond = fcmp one double %x, 0.000000e+00
```



```

br i1 %ifcond, label %then, label %else

then:      ; preds = %entry
%calltmp = call double @foo()
br label %ifcont

else:      ; preds = %entry
%calltmp1 = call double @bar()
br label %ifcont

ifcont:    ; preds = %else, %then
%iftmp = phi double [ %calltmp, %then ], [ %calltmp1, %else ]
ret double %iftmp
}

```

To visualize the control flow graph, you can use a nifty feature of the LLVM ‘opt’ tool. If you put this LLVM IR into “t.ll” and run “llvm-as < t.ll | opt -analyze -view-cfg”, a window will pop up and you’ll see this graph:

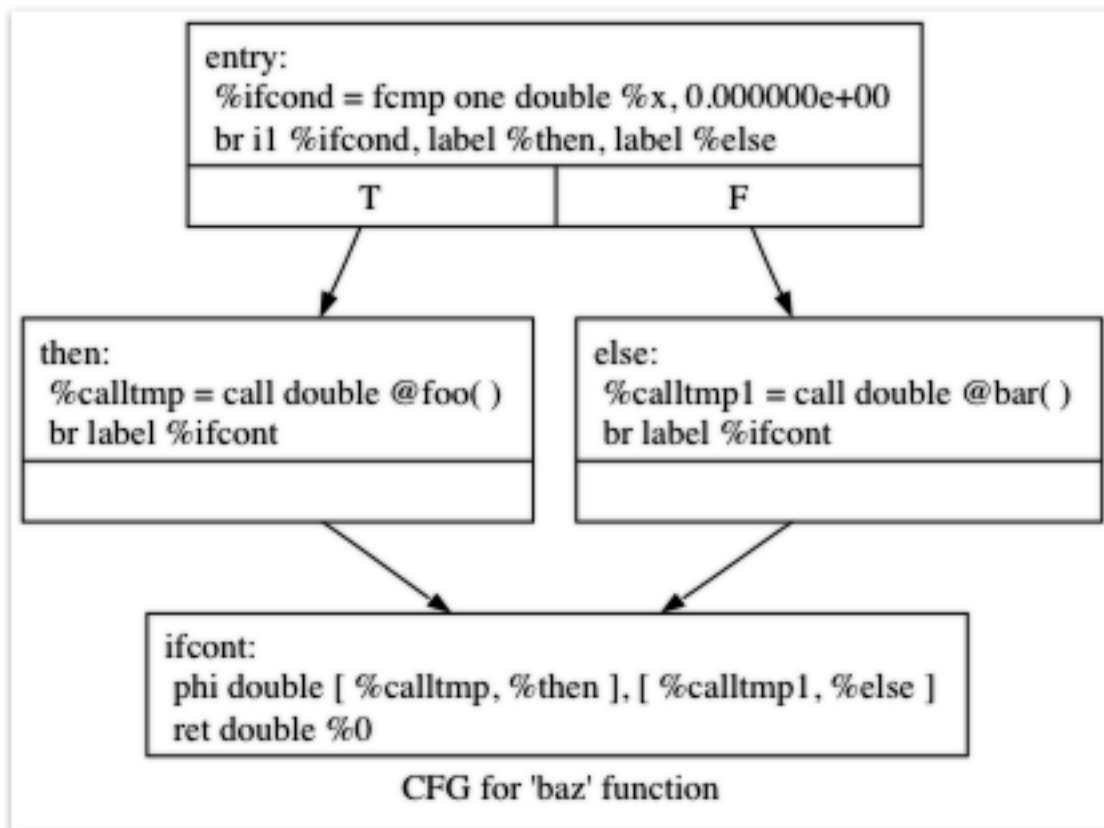


Fig. 2.1: Example CFG

Another way to get this is to call “F->viewCFG()” or “F->viewCFGOnly()” (where F is a “Function*”) either by inserting actual calls into the code and recompiling or by calling these in the debugger. LLVM has many nice features for visualizing various graphs.

Getting back to the generated code, it is fairly simple: the entry block evaluates the conditional expression (“x” in our case here) and compares the result to 0.0 with the “fcmp one” instruction (‘one’ is “Ordered and Not Equal”). Based on the result of this expression, the code jumps to either the “then” or “else” blocks, which contain the expressions for

the true/false cases.

Once the then/else blocks are finished executing, they both branch back to the ‘ifcont’ block to execute the code that happens after the if/then/else. In this case the only thing left to do is to return to the caller of the function. The question then becomes: how does the code know which expression to return?

The answer to this question involves an important SSA operation: the [Phi operation](#). If you’re not familiar with SSA, [the wikipedia article](#) is a good introduction and there are various other introductions to it available on your favorite search engine. The short version is that “execution” of the Phi operation requires “remembering” which block control came from. The Phi operation takes on the value corresponding to the input control block. In this case, if control comes in from the “then” block, it gets the value of “calltmp”. If control comes from the “else” block, it gets the value of “calltmp1”.

At this point, you are probably starting to think “Oh no! This means my simple and elegant front-end will have to start generating SSA form in order to use LLVM!”. Fortunately, this is not the case, and we strongly advise *not* implementing an SSA construction algorithm in your front-end unless there is an amazingly good reason to do so. In practice, there are two sorts of values that float around in code written for your average imperative programming language that might need Phi nodes:

1. Code that involves user variables: `x = 1; x = x + 1;`
2. Values that are implicit in the structure of your AST, such as the Phi node in this case.

In [Chapter 7](#) of this tutorial (“mutable variables”), we’ll talk about #1 in depth. For now, just believe me that you don’t need SSA construction to handle this case. For #2, you have the choice of using the techniques that we will describe for #1, or you can insert Phi nodes directly, if convenient. In this case, it is really easy to generate the Phi node, so we choose to do it directly.

Okay, enough of the motivation and overview, lets generate code!

Code Generation for If/Then/Else

In order to generate code for this, we implement the `codegen` method for `IfExprAST`:

```
Value *IfExprAST::codegen() {
    Value *CondV = Cond->codegen();
    if (!CondV)
        return nullptr;

    // Convert condition to a bool by comparing non-equal to 0.0.
    CondV = Builder.CreateFCmpONE(
        CondV, ConstantFP::get(TheContext, APFloat(0.0)), "ifcond");
```

This code is straightforward and similar to what we saw before. We emit the expression for the condition, then compare that value to zero to get a truth value as a 1-bit (bool) value.

```
Function *TheFunction = Builder.GetInsertBlock()->getParent();

// Create blocks for the then and else cases. Insert the 'then' block at the
// end of the function.
BasicBlock *ThenBB =
    BasicBlock::Create(TheContext, "then", TheFunction);
BasicBlock *ElseBB = BasicBlock::Create(TheContext, "else");
BasicBlock *MergeBB = BasicBlock::Create(TheContext, "ifcont");

Builder.CreateCondBr(CondV, ThenBB, ElseBB);
```

This code creates the basic blocks that are related to the if/then/else statement, and correspond directly to the blocks in the example above. The first line gets the current Function object that is being built. It gets this by asking the builder for the current BasicBlock, and asking that block for its “parent” (the function it is currently embedded into).

Once it has that, it creates three blocks. Note that it passes “TheFunction” into the constructor for the “then” block. This causes the constructor to automatically insert the new block into the end of the specified function. The other two blocks are created, but aren’t yet inserted into the function.

Once the blocks are created, we can emit the conditional branch that chooses between them. Note that creating new blocks does not implicitly affect the IRBuilder, so it is still inserting into the block that the condition went into. Also note that it is creating a branch to the “then” block and the “else” block, even though the “else” block isn’t inserted into the function yet. This is all ok: it is the standard way that LLVM supports forward references.

```
// Emit then value.
Builder.SetInsertPoint(ThenBB);

Value *ThenV = Then->codegen();
if (!ThenV)
    return nullptr;

Builder.CreateBr(MergeBB);
// Codegen of 'Then' can change the current block, update ThenBB for the PHI.
ThenBB = Builder.GetInsertBlock();
```

After the conditional branch is inserted, we move the builder to start inserting into the “then” block. Strictly speaking, this call moves the insertion point to be at the end of the specified block. However, since the “then” block is empty, it also starts out by inserting at the beginning of the block. :)

Once the insertion point is set, we recursively codegen the “then” expression from the AST. To finish off the “then” block, we create an unconditional branch to the merge block. One interesting (and very important) aspect of the LLVM IR is that it **requires all basic blocks to be “terminated”** with a **control flow instruction** such as return or branch. This means that all control flow, *including fall throughs* must be made explicit in the LLVM IR. If you violate this rule, the verifier will emit an error.

The final line here is quite subtle, but is very important. The basic issue is that when we create the Phi node in the merge block, we need to set up the block/value pairs that indicate how the Phi will work. Importantly, the Phi node expects to have an entry for each predecessor of the block in the CFG. Why then, are we getting the current block when we just set it to ThenBB 5 lines above? The problem is that the “Then” expression may actually itself change the block that the Builder is emitting into if, for example, it contains a nested “if/then/else” expression. Because calling `codegen()` recursively could arbitrarily change the notion of the current block, we are required to get an up-to-date value for code that will set up the Phi node.

```
// Emit else block.
TheFunction->getBasicBlockList().push_back(ElseBB);
Builder.SetInsertPoint(ElseBB);

Value *ElseV = Else->codegen();
if (!ElseV)
    return nullptr;

Builder.CreateBr(MergeBB);
// codegen of 'Else' can change the current block, update ElseBB for the PHI.
ElseBB = Builder.GetInsertBlock();
```

Code generation for the ‘else’ block is basically identical to codegen for the ‘then’ block. The only significant difference is the first line, which adds the ‘else’ block to the function. Recall previously that the ‘else’ block was created, but not added to the function. Now that the ‘then’ and ‘else’ blocks are emitted, we can finish up with the merge code:

```
// Emit merge block.
TheFunction->getBasicBlockList().push_back(MergeBB);
Builder.SetInsertPoint(MergeBB);
PHINode *PN =
    Builder.CreatePHI(Type::getDoubleTy(TheContext), 2, "iftmp");

PN->addIncoming(ThenV, ThenBB);
PN->addIncoming(ElseV, ElseBB);
return PN;
}
```

The first two lines here are now familiar: the first adds the “merge” block to the Function object (it was previously floating, like the else block above). The second changes the insertion point so that newly created code will go into the “merge” block. Once that is done, we need to create the PHI node and set up the block/value pairs for the PHI.

Finally, the CodeGen function returns the phi node as the value computed by the if/then/else expression. In our example above, this returned value will feed into the code for the top-level function, which will create the return instruction.

Overall, we now have the ability to execute conditional code in Kaleidoscope. With this extension, Kaleidoscope is a fairly complete language that can calculate a wide variety of numeric functions. Next up we’ll add another useful expression that is familiar from non-functional languages...

‘for’ Loop Expression

Now that we know how to add basic control flow constructs to the language, we have the tools to add more powerful things. Lets add something more aggressive, a ‘for’ expression:

```
extern putchar(char);
def printstar(n)
    for i = 1, i < n, 1.0 in
        putchar(42); # ascii 42 = '*'

# print 100 '*' characters
printstar(100);
```

This expression defines a new variable (“i” in this case) which iterates from a starting value, while the condition (“i < n” in this case) is true, incrementing by an optional step value (“1.0” in this case). If the step value is omitted, it defaults to 1.0. While the loop is true, it executes its body expression. Because we don’t have anything better to return, we’ll just define the loop as always returning 0.0. In the future when we have mutable variables, it will get more useful.

As before, lets talk about the changes that we need to Kaleidoscope to support this.

Lexer Extensions for the ‘for’ Loop

The lexer extensions are the same sort of thing as for if/then/else:

```
... in enum Token ...
// control
tok_if = -6, tok_then = -7, tok_else = -8,
tok_for = -9, tok_in = -10

... in gettok ...
if (IdentifierStr == "def")
```

```

    return tok_def;
if (IdentifierStr == "extern")
    return tok_extern;
if (IdentifierStr == "if")
    return tok_if;
if (IdentifierStr == "then")
    return tok_then;
if (IdentifierStr == "else")
    return tok_else;
if (IdentifierStr == "for")
    return tok_for;
if (IdentifierStr == "in")
    return tok_in;
return tok_identifier;

```

AST Extensions for the ‘for’ Loop

The AST node is just as simple. It basically boils down to capturing the variable name and the constituent expressions in the node.

```

/// ForExprAST - Expression class for for/in.
class ForExprAST : public ExprAST {
    std::string VarName;
    std::unique_ptr<ExprAST> Start, End, Step, Body;

public:
    ForExprAST(const std::string &VarName, std::unique_ptr<ExprAST> Start,
               std::unique_ptr<ExprAST> End, std::unique_ptr<ExprAST> Step,
               std::unique_ptr<ExprAST> Body)
        : VarName(VarName), Start(std::move(Start)), End(std::move(End)),
          Step(std::move(Step)), Body(std::move(Body)) {}

    Value *codegen() override;
};

```

Parser Extensions for the ‘for’ Loop

The parser code is also fairly standard. The only interesting thing here is handling of the optional step value. The parser code handles it by checking to see if the second comma is present. If not, it sets the step value to null in the AST node:

```

/// forexpr ::= 'for' identifier '=' expr ',' expr (',' expr)? 'in' expression
static std::unique_ptr<ExprAST> ParseForExpr() {
    getNextToken(); // eat the for.

    if (CurTok != tok_identifier)
        return LogError("expected identifier after for");

    std::string IdName = IdentifierStr;
    getNextToken(); // eat identifier.

    if (CurTok != '=')
        return LogError("expected '=' after for");
    getNextToken(); // eat '='.

```

```
auto Start = ParseExpression();
if (!Start)
    return nullptr;
if (CurTok != ',')
    return LogError("expected ',' after for start value");
getNextToken();

auto End = ParseExpression();
if (!End)
    return nullptr;

// The step value is optional.
std::unique_ptr<ExprAST> Step;
if (CurTok == ',') {
    getNextToken();
    Step = ParseExpression();
    if (!Step)
        return nullptr;
}

if (CurTok != tok_in)
    return LogError("expected 'in' after for");
getNextToken(); // eat 'in'.

auto Body = ParseExpression();
if (!Body)
    return nullptr;

return llvm::make_unique<ForExprAST>(IdName, std::move(Start),
                                     std::move(End), std::move(Step),
                                     std::move(Body));
}
```

And again we hook it up as a primary expression:

```
static std::unique_ptr<ExprAST> ParsePrimary() {
    switch (CurTok) {
    default:
        return LogError("unknown token when expecting an expression");
    case tok_identifier:
        return ParseIdentifierExpr();
    case tok_number:
        return ParseNumberExpr();
    case '(':
        return ParseParenExpr();
    case tok_if:
        return ParseIfExpr();
    case tok_for:
        return ParseForExpr();
    }
}
```

LLVM IR for the ‘for’ Loop

Now we get to the good part: the LLVM IR we want to generate for this thing. With the simple example above, we get this LLVM IR (note that this dump is generated with optimizations disabled for clarity):

```
declare double @putchar(double)

define double @printstar(double %n) {
entry:
    ; initial value = 1.0 (inlined into phi)
    br label %loop

loop:      ; preds = %loop, %entry
    %i = phi double [ 1.000000e+00, %entry ], [ %nextvar, %loop ]
    ; body
    %calltmp = call double @putchar(double 4.200000e+01)
    ; increment
    %nextvar = fadd double %i, 1.000000e+00

    ; termination test
    %cmptmp = fcmp ult double %i, %n
    %booltmp = uitoffp i1 %cmptmp to double
    %loopcond = fcmp one double %booltmp, 0.000000e+00
    br i1 %loopcond, label %loop, label %afterloop

afterloop: ; preds = %loop
    ; loop always returns 0.0
    ret double 0.000000e+00
}
```

This loop contains all the same constructs we saw before: a phi node, several expressions, and some basic blocks. Lets see how this fits together.

Code Generation for the ‘for’ Loop

The first part of codegen is very simple: we just output the start expression for the loop value:

```
Value *ForExprAST::codegen() {
    // Emit the start code first, without 'variable' in scope.
    Value *StartVal = Start->codegen();
    if (!StartVal)
        return nullptr;
}
```

With this out of the way, the next step is to set up the LLVM basic block for the start of the loop body. In the case above, the whole loop body is one block, but remember that the body code itself could consist of multiple blocks (e.g. if it contains an if/then/else or a for/in expression).

```
// Make the new basic block for the loop header, inserting after current
// block.
Function *TheFunction = Builder.GetInsertBlock()->getParent();
BasicBlock *PreheaderBB = Builder.GetInsertBlock();
BasicBlock *LoopBB =
    BasicBlock::Create(TheContext, "loop", TheFunction);

// Insert an explicit fall through from the current block to the LoopBB.
Builder.CreateBr(LoopBB);
```

This code is similar to what we saw for if/then/else. Because we will need it to create the Phi node, we remember the block that falls through into the loop. Once we have that, we create the actual block that starts the loop and create an unconditional branch for the fall-through between the two blocks.

```
// Start insertion in LoopBB.
Builder.SetInsertPoint(LoopBB);

// Start the PHI node with an entry for Start.
PHINode *Variable = Builder.CreatePHI(Type::getDoubleTy(TheContext),
                                     2, VarName.c_str());
Variable->addIncoming(StartVal, PreheaderBB);
```

Now that the “preheader” for the loop is set up, we switch to emitting code for the loop body. To begin with, we move the insertion point and create the PHI node for the loop induction variable. Since we already know the incoming value for the starting value, we add it to the Phi node. Note that the Phi will eventually get a second value for the backedge, but we can’t set it up yet (because it doesn’t exist!).

```
// Within the loop, the variable is defined equal to the PHI node. If it
// shadows an existing variable, we have to restore it, so save it now.
Value *OldVal = NamedValues[VarName];
NamedValues[VarName] = Variable;

// Emit the body of the loop. This, like any other expr, can change the
// current BB. Note that we ignore the value computed by the body, but don't
// allow an error.
if (!Body->codegen())
    return nullptr;
```

Now the code starts to get more interesting. Our ‘for’ loop introduces a new variable to the symbol table. This means that our symbol table can now contain either function arguments or loop variables. To handle this, before we codegen the body of the loop, we add the loop variable as the current value for its name. Note that it is possible that there is a variable of the same name in the outer scope. It would be easy to make this an error (emit an error and return null if there is already an entry for VarName) but we choose to allow shadowing of variables. In order to handle this correctly, we remember the Value that we are potentially shadowing in OldVal (which will be null if there is no shadowed variable).

Once the loop variable is set into the symbol table, the code recursively codegen’s the body. This allows the body to use the loop variable: any references to it will naturally find it in the symbol table.

```
// Emit the step value.
Value *StepVal = nullptr;
if (Step) {
    StepVal = Step->codegen();
    if (!StepVal)
        return nullptr;
} else {
    // If not specified, use 1.0.
    StepVal = ConstantFP::get(TheContext, APFloat(1.0));
}

Value *NextVar = Builder.CreateFAdd(Variable, StepVal, "nextvar");
```

Now that the body is emitted, we compute the next value of the iteration variable by adding the step value, or 1.0 if it isn’t present. ‘NextVar’ will be the value of the loop variable on the next iteration of the loop.

```
// Compute the end condition.
Value *EndCond = End->codegen();
if (!EndCond)
```



```

return nullptr;

// Convert condition to a bool by comparing non-equal to 0.0.
EndCond = Builder.CreateFCmpONE(
    EndCond, ConstantFP::get(TheContext, APFloat(0.0)), "loopcond");

```

Finally, we evaluate the exit value of the loop, to determine whether the loop should exit. This mirrors the condition evaluation for the if/then/else statement.

```

// Create the "after loop" block and insert it.
BasicBlock *LoopEndBB = Builder.GetInsertBlock();
BasicBlock *AfterBB =
    BasicBlock::Create(TheContext, "afterloop", TheFunction);

// Insert the conditional branch into the end of LoopEndBB.
Builder.CreateCondBr(EndCond, LoopBB, AfterBB);

// Any new code will be inserted in AfterBB.
Builder.SetInsertPoint(AfterBB);

```

With the code for the body of the loop complete, we just need to finish up the control flow for it. This code remembers the end block (for the phi node), then creates the block for the loop exit (“afterloop”). Based on the value of the exit condition, it creates a conditional branch that chooses between executing the loop again and exiting the loop. Any future code is emitted in the “afterloop” block, so it sets the insertion position to it.

```

// Add a new entry to the PHI node for the backedge.
Variable->addIncoming(NextVar, LoopEndBB);

// Restore the unshadowed variable.
if (OldVal)
    NamedValues[VarName] = OldVal;
else
    NamedValues.erase(VarName);

// for expr always returns 0.0.
return Constant::getNullValue(Type::getDoubleTy(TheContext));
}

```

The final code handles various cleanups: now that we have the “NextVar” value, we can add the incoming value to the loop PHI node. After that, we remove the loop variable from the symbol table, so that it isn’t in scope after the for loop. Finally, code generation of the for loop always returns 0.0, so that is what we return from `ForExprAST::codegen()`.

With this, we conclude the “adding control flow to Kaleidoscope” chapter of the tutorial. In this chapter we added two control flow constructs, and used them to motivate a couple of aspects of the LLVM IR that are important for front-end implementors to know. In the next chapter of our saga, we will get a bit crazier and add [user-defined operators](#) to our poor innocent language.

Full Code Listing

Here is the complete code listing for our running example, enhanced with the if/then/else and for expressions. To build this example, use:

```

# Compile
clang++ -g toy.cpp `llvm-config --cxxflags --ldflags --system-libs --libs core mcjit_
↪native` -O3 -o toy

```

```
# Run
./toy
```

Here is the code:

```
#include "llvm/ADT/APFloat.h"
#include "llvm/ADT/STLExtras.h"
#include "llvm/IR/BasicBlock.h"
#include "llvm/IR/Constants.h"
#include "llvm/IR/DerivedTypes.h"
#include "llvm/IR/Function.h"
#include "llvm/IR/Instructions.h"
#include "llvm/IR/IRBuilder.h"
#include "llvm/IR/LLVMContext.h"
#include "llvm/IR/LegacyPassManager.h"
#include "llvm/IR/Module.h"
#include "llvm/IR/Type.h"
#include "llvm/IR/Verifier.h"
#include "llvm/Support/TargetSelect.h"
#include "llvm/Target/TargetMachine.h"
#include "llvm/Transforms/Scalar.h"
#include "llvm/Transforms/Scalar/GVN.h"
#include "../include/KaleidoscopeJIT.h"
#include <algorithm>
#include <cassert>
#include <cctype>
#include <cstdint>
#include <cstdio>
#include <cstdlib>
#include <map>
#include <memory>
#include <string>
#include <vector>

using namespace llvm;
using namespace llvm::orc;

//=====//
// Lexer
//=====//

// The lexer returns tokens [0-255] if it is an unknown character, otherwise one
// of these for known things.
enum Token {
    tok_eof = -1,

    // commands
    tok_def = -2,
    tokExtern = -3,

    // primary
    tok_identifier = -4,
    tok_number = -5,

    // control
    tok_if = -6,
    tok_then = -7,
    tok_else = -8,
```

```

    tok_for = -9,
    tok_in = -10
};

static std::string IdentifierStr; // Filled in if tok_identifier
static double NumVal;           // Filled in if tok_number

/// gettok - Return the next token from standard input.
static int gettok() {
    static int LastChar = ' ';

    // Skip any whitespace.
    while (isspace(LastChar))
        LastChar = getchar();

    if (isalpha(LastChar)) { // identifier: [a-zA-Z][a-zA-Z0-9]*
        IdentifierStr = LastChar;
        while (isalnum((LastChar = getchar())))
            IdentifierStr += LastChar;

        if (IdentifierStr == "def")
            return tok_def;
        if (IdentifierStr == "extern")
            return tok_extern;
        if (IdentifierStr == "if")
            return tok_if;
        if (IdentifierStr == "then")
            return tok_then;
        if (IdentifierStr == "else")
            return tok_else;
        if (IdentifierStr == "for")
            return tok_for;
        if (IdentifierStr == "in")
            return tok_in;
        return tok_identifier;
    }

    if (isdigit(LastChar) || LastChar == '.') { // Number: [0-9.]+
        std::string NumStr;
        do {
            NumStr += LastChar;
            LastChar = getchar();
        } while (isdigit(LastChar) || LastChar == '.');

        NumVal = strtod(NumStr.c_str(), nullptr);
        return tok_number;
    }

    if (LastChar == '#') {
        // Comment until end of line.
        do
            LastChar = getchar();
        while (LastChar != EOF && LastChar != '\n' && LastChar != '\r');

        if (LastChar != EOF)
            return gettok();
    }
}

```

```
// Check for end of file. Don't eat the EOF.
if (LastChar == EOF)
    return tok_eof;

// Otherwise, just return the character as its ascii value.
int ThisChar = LastChar;
LastChar = getchar();
return ThisChar;
}

//===-----//
// Abstract Syntax Tree (aka Parse Tree)
//===-----//

namespace {

/// ExprAST - Base class for all expression nodes.
class ExprAST {
public:
    virtual ~ExprAST() = default;

    virtual Value *codegen() = 0;
};

/// NumberExprAST - Expression class for numeric literals like "1.0".
class NumberExprAST : public ExprAST {
    double Val;

public:
    NumberExprAST(double Val) : Val(Val) {}

    Value *codegen() override;
};

/// VariableExprAST - Expression class for referencing a variable, like "a".
class VariableExprAST : public ExprAST {
    std::string Name;

public:
    VariableExprAST(const std::string &Name) : Name(Name) {}

    Value *codegen() override;
};

/// BinaryExprAST - Expression class for a binary operator.
class BinaryExprAST : public ExprAST {
    char Op;
    std::unique_ptr<ExprAST> LHS, RHS;

public:
    BinaryExprAST(char Op, std::unique_ptr<ExprAST> LHS,
                  std::unique_ptr<ExprAST> RHS)
        : Op(Op), LHS(std::move(LHS)), RHS(std::move(RHS)) {}

    Value *codegen() override;
};

/// CallExprAST - Expression class for function calls.
```

```

class CallExprAST : public ExprAST {
    std::string Callee;
    std::vector<std::unique_ptr<ExprAST>> Args;

public:
    CallExprAST(const std::string &Callee,
                std::vector<std::unique_ptr<ExprAST>> Args)
        : Callee(Callee), Args(std::move(Args)) {}

    Value *codegen() override;
};

/// IfExprAST - Expression class for if/then/else.
class IfExprAST : public ExprAST {
    std::unique_ptr<ExprAST> Cond, Then, Else;

public:
    IfExprAST(std::unique_ptr<ExprAST> Cond, std::unique_ptr<ExprAST> Then,
              std::unique_ptr<ExprAST> Else)
        : Cond(std::move(Cond)), Then(std::move(Then)), Else(std::move(Else)) {}

    Value *codegen() override;
};

/// ForExprAST - Expression class for for/in.
class ForExprAST : public ExprAST {
    std::string VarName;
    std::unique_ptr<ExprAST> Start, End, Step, Body;

public:
    ForExprAST(const std::string &VarName, std::unique_ptr<ExprAST> Start,
              std::unique_ptr<ExprAST> End, std::unique_ptr<ExprAST> Step,
              std::unique_ptr<ExprAST> Body)
        : VarName(VarName), Start(std::move(Start)), End(std::move(End)),
          Step(std::move(Step)), Body(std::move(Body)) {}

    Value *codegen() override;
};

/// PrototypeAST - This class represents the "prototype" for a function,
/// which captures its name, and its argument names (thus implicitly the number
/// of arguments the function takes).
class PrototypeAST {
    std::string Name;
    std::vector<std::string> Args;

public:
    PrototypeAST(const std::string &Name, std::vector<std::string> Args)
        : Name(Name), Args(std::move(Args)) {}

    Function *codegen();
    const std::string &getName() const { return Name; }
};

/// FunctionAST - This class represents a function definition itself.
class FunctionAST {
    std::unique_ptr<PrototypeAST> Proto;
    std::unique_ptr<ExprAST> Body;

```

```

public:
    FunctionAST(std::unique_ptr<PrototypeAST> Proto,
                std::unique_ptr<ExprAST> Body)
        : Proto(std::move(Proto)), Body(std::move(Body)) {}

    Function *codegen();
};

} // end anonymous namespace

//===----- Parser
// Parser
//===-----

/// CurTok/getNextToken - Provide a simple token buffer. CurTok is the current
/// token the parser is looking at. getNextToken reads another token from the
/// lexer and updates CurTok with its results.
static int CurTok;
static int getNextToken() { return CurTok = gettok(); }

/// BinopPrecedence - This holds the precedence for each binary operator that is
/// defined.
static std::map<char, int> BinopPrecedence;

/// GetTokPrecedence - Get the precedence of the pending binary operator token.
static int GetTokPrecedence() {
    if (!isascii(CurTok))
        return -1;

    // Make sure it's a declared binop.
    int TokPrec = BinopPrecedence[CurTok];
    if (TokPrec <= 0)
        return -1;
    return TokPrec;
}

/// LogError* - These are little helper functions for error handling.
std::unique_ptr<ExprAST> LogError(const char *Str) {
    fprintf(stderr, "Error: %s\n", Str);
    return nullptr;
}

std::unique_ptr<PrototypeAST> LogErrorP(const char *Str) {
    LogError(Str);
    return nullptr;
}

static std::unique_ptr<ExprAST> ParseExpression();

/// numberexpr ::= number
static std::unique_ptr<ExprAST> ParseNumberExpr() {
    auto Result = llvm::make_unique<NumberExprAST>(NumVal);
    getNextToken(); // consume the number
    return std::move(Result);
}

/// parenexpr ::= '(' expression ')'

```

```

static std::unique_ptr<ExprAST> ParseParenExpr() {
    getNextToken(); // eat (.
    auto V = ParseExpression();
    if (!V)
        return nullptr;

    if (CurTok != ')')
        return LogError("expected ')'");
    getNextToken(); // eat ).
    return V;
}

/// identifierexpr
/// ::= identifier
/// ::= identifier '(' expression* ')'
static std::unique_ptr<ExprAST> ParseIdentifierExpr() {
    std::string IdName = IdentifierStr;

    getNextToken(); // eat identifier.

    if (CurTok != '(') // Simple variable ref.
        return llvm::make_unique<VariableExprAST>(IdName);

    // Call.
    getNextToken(); // eat (.
    std::vector<std::unique_ptr<ExprAST>> Args;
    if (CurTok != ')') {
        while (true) {
            if (auto Arg = ParseExpression())
                Args.push_back(std::move(Arg));
            else
                return nullptr;

            if (CurTok == ')')
                break;

            if (CurTok != ',')
                return LogError("Expected ')' or ',' in argument list");
            getNextToken();
        }
    }

    // Eat the ')'.
    getNextToken();

    return llvm::make_unique<CallExprAST>(IdName, std::move(Args));
}

/// ifexpr ::= 'if' expression 'then' expression 'else' expression
static std::unique_ptr<ExprAST> ParseIfExpr() {
    getNextToken(); // eat the if.

    // condition.
    auto Cond = ParseExpression();
    if (!Cond)
        return nullptr;

    if (CurTok != tok_then)

```

```

    return LogError("expected then");
getNextToken(); // eat the then

    auto Then = ParseExpression();
    if (!Then)
        return nullptr;

    if (CurTok != tok_else)
        return LogError("expected else");

    getNextToken();

    auto Else = ParseExpression();
    if (!Else)
        return nullptr;

    return llvm::make_unique<IfExprAST>(std::move(Cond), std::move(Then),
                                        std::move(Else));
}

/// forexpr ::= 'for' identifier '=' expr ',' expr (',' expr)? 'in' expression
static std::unique_ptr<ExprAST> ParseForExpr() {
    getNextToken(); // eat the for.

    if (CurTok != tok_identifier)
        return LogError("expected identifier after for");

    std::string IdName = IdentifierStr;
    getNextToken(); // eat identifier.

    if (CurTok != '=')
        return LogError("expected '=' after for");
    getNextToken(); // eat '='.

    auto Start = ParseExpression();
    if (!Start)
        return nullptr;
    if (CurTok != ',')
        return LogError("expected ',' after for start value");
    getNextToken();

    auto End = ParseExpression();
    if (!End)
        return nullptr;

    // The step value is optional.
    std::unique_ptr<ExprAST> Step;
    if (CurTok == ',') {
        getNextToken();
        Step = ParseExpression();
        if (!Step)
            return nullptr;
    }

    if (CurTok != tok_in)
        return LogError("expected 'in' after for");
    getNextToken(); // eat 'in'.

```



```

    auto Body = ParseExpression();
    if (!Body)
        return nullptr;

    return llvm::make_unique<ForExprAST>(IdName, std::move(Start), std::move(End),
                                         std::move(Step), std::move(Body));
}

/// primary
/// ::= identifierexpr
/// ::= numberexpr
/// ::= parenexpr
/// ::= ifexpr
/// ::= forexpr
static std::unique_ptr<ExprAST> ParsePrimary() {
    switch (CurTok) {
    default:
        return LogError("unknown token when expecting an expression");
    case tok_identifier:
        return ParseIdentifierExpr();
    case tok_number:
        return ParseNumberExpr();
    case '(':
        return ParseParenExpr();
    case tok_if:
        return ParseIfExpr();
    case tok_for:
        return ParseForExpr();
    }
}

/// binoprhs
/// ::= ('+' primary)*
static std::unique_ptr<ExprAST> ParseBinOpRHS(int ExprPrec,
                                              std::unique_ptr<ExprAST> LHS) {
    // If this is a binop, find its precedence.
    while (true) {
        int TokPrec = GetTokPrecedence();

        // If this is a binop that binds at least as tightly as the current binop,
        // consume it, otherwise we are done.
        if (TokPrec < ExprPrec)
            return LHS;

        // Okay, we know this is a binop.
        int BinOp = CurTok;
        getNextToken(); // eat binop

        // Parse the primary expression after the binary operator.
        auto RHS = ParsePrimary();
        if (!RHS)
            return nullptr;

        // If BinOp binds less tightly with RHS than the operator after RHS, let
        // the pending operator take RHS as its LHS.
        int NextPrec = GetTokPrecedence();
        if (TokPrec < NextPrec) {
            RHS = ParseBinOpRHS(TokPrec + 1, std::move(RHS));
        }
    }
}

```

```

        if (!RHS)
            return nullptr;
    }

    // Merge LHS/RHS.
    LHS =
        llvm::make_unique<BinaryExprAST>(BinOp, std::move(LHS), std::move(RHS));
}

/// expression
/// ::= primary binoprhs
///
static std::unique_ptr<ExprAST> ParseExpression() {
    auto LHS = ParsePrimary();
    if (!LHS)
        return nullptr;

    return ParseBinOpRHS(0, std::move(LHS));
}

/// prototype
/// ::= id '(' id* ')'
static std::unique_ptr<PrototypeAST> ParsePrototype() {
    if (CurTok != tok_identifier)
        return LogErrorP("Expected function name in prototype");

    std::string FnName = IdentifierStr;
    getNextToken();

    if (CurTok != '(')
        return LogErrorP("Expected '(' in prototype");

    std::vector<std::string> ArgNames;
    while (getNextToken() == tok_identifier)
        ArgNames.push_back(IdentifierStr);
    if (CurTok != ')')
        return LogErrorP("Expected ')' in prototype");

    // success.
    getNextToken(); // eat ')'.

    return llvm::make_unique<PrototypeAST>(FnName, std::move(ArgNames));
}

/// definition ::= 'def' prototype expression
static std::unique_ptr<FunctionAST> ParseDefinition() {
    getNextToken(); // eat def.
    auto Proto = ParsePrototype();
    if (!Proto)
        return nullptr;

    if (auto E = ParseExpression())
        return llvm::make_unique<FunctionAST>(std::move(Proto), std::move(E));
    return nullptr;
}

/// toplevelexpr ::= expression

```

```

static std::unique_ptr<FunctionAST> ParseTopLevelExpr() {
    if (auto E = ParseExpression()) {
        // Make an anonymous proto.
        auto Proto = llvm::make_unique<PrototypeAST>("__anon_expr",
                                                    std::vector<std::string>());
        return llvm::make_unique<FunctionAST>(std::move(Proto), std::move(E));
    }
    return nullptr;
}

/// external ::= 'extern' prototype
static std::unique_ptr<PrototypeAST> ParseExtern() {
    getNextToken(); // eat extern.
    return ParsePrototype();
}

//===-----
// Code Generation
//===-----

static LLVMContext TheContext;
static IRBuilder<> Builder(TheContext);
static std::unique_ptr<Module> TheModule;
static std::map<std::string, Value*> NamedValues;
static std::unique_ptr<legacy::FunctionPassManager> TheFPM;
static std::unique_ptr<KaleidoscopeJIT> TheJIT;
static std::map<std::string, std::unique_ptr<PrototypeAST>> FunctionProtos;

Value *LogErrorV(const char *Str) {
    LogError(Str);
    return nullptr;
}

Function *getFunction(std::string Name) {
    // First, see if the function has already been added to the current module.
    if (auto *F = TheModule->getFunction(Name))
        return F;

    // If not, check whether we can codegen the declaration from some existing
    // prototype.
    auto FI = FunctionProtos.find(Name);
    if (FI != FunctionProtos.end())
        return FI->second->codegen();

    // If no existing prototype exists, return null.
    return nullptr;
}

Value *NumberExprAST::codegen() {
    return ConstantFP::get(TheContext, APFloat(Val));
}

Value *VariableExprAST::codegen() {
    // Look this variable up in the function.
    Value *V = NamedValues[Name];
    if (!V)
        return LogErrorV("Unknown variable name");
    return V;
}

```

```

}

Value *BinaryExprAST::codegen() {
    Value *L = LHS->codegen();
    Value *R = RHS->codegen();
    if (!L || !R)
        return nullptr;

    switch (Op) {
    case '+':
        return Builder.CreateFAdd(L, R, "addtmp");
    case '-':
        return Builder.CreateFSub(L, R, "subtmp");
    case '*':
        return Builder.CreateFMul(L, R, "multmp");
    case '<':
        L = Builder.CreateFCmpULT(L, R, "cmptmp");
        // Convert bool 0/1 to double 0.0 or 1.0
        return Builder.CreateUIToFP(L, Type::getDoubleTy(TheContext), "booltmp");
    default:
        return LogErrorV("invalid binary operator");
    }
}

Value *CallExprAST::codegen() {
    // Look up the name in the global module table.
    Function *CalleeF = getFunction(Callee);
    if (!CalleeF)
        return LogErrorV("Unknown function referenced");

    // If argument mismatch error.
    if (CalleeF->arg_size() != Args.size())
        return LogErrorV("Incorrect # arguments passed");

    std::vector<Value *> ArgsV;
    for (unsigned i = 0, e = Args.size(); i != e; ++i) {
        ArgsV.push_back(Args[i]->codegen());
        if (!ArgsV.back())
            return nullptr;
    }

    return Builder.CreateCall(CalleeF, ArgsV, "calltmp");
}

Value *IfExprAST::codegen() {
    Value *CondV = Cond->codegen();
    if (!CondV)
        return nullptr;

    // Convert condition to a bool by comparing non-equal to 0.0.
    CondV = Builder.CreateFCmpONE(
        CondV, ConstantFP::get(TheContext, APFloat(0.0)), "ifcond");

    Function *TheFunction = Builder.GetInsertBlock()->getParent();

    // Create blocks for the then and else cases. Insert the 'then' block at the
    // end of the function.
    BasicBlock *ThenBB = BasicBlock::Create(TheContext, "then", TheFunction);

```

```

BasicBlock *ElseBB = BasicBlock::Create(TheContext, "else");
BasicBlock *MergeBB = BasicBlock::Create(TheContext, "ifcont");

Builder.CreateCondBr(CondV, ThenBB, ElseBB);

// Emit then value.
Builder.SetInsertPoint(ThenBB);

Value *ThenV = Then->codegen();
if (!ThenV)
    return nullptr;

Builder.CreateBr(MergeBB);
// Codegen of 'Then' can change the current block, update ThenBB for the PHI.
ThenBB = Builder.GetInsertBlock();

// Emit else block.
TheFunction->getBasicBlockList().push_back(ElseBB);
Builder.SetInsertPoint(ElseBB);

Value *ElseV = Else->codegen();
if (!ElseV)
    return nullptr;

Builder.CreateBr(MergeBB);
// Codegen of 'Else' can change the current block, update ElseBB for the PHI.
ElseBB = Builder.GetInsertBlock();

// Emit merge block.
TheFunction->getBasicBlockList().push_back(MergeBB);
Builder.SetInsertPoint(MergeBB);
PHINode *PN = Builder.CreatePHI(Type::getDoubleTy(TheContext), 2, "iftmp");

PN->addIncoming(ThenV, ThenBB);
PN->addIncoming(ElseV, ElseBB);
return PN;
}

// Output for-loop as:
// ...
// start = startexpr
// goto loop
// loop:
// variable = phi [start, loopheader], [nextvariable, loopend]
// ...
// bodyexpr
// ...
// loopend:
// step = stepexpr
// nextvariable = variable + step
// endcond = endexpr
// br endcond, loop, endloop
// outloop:
Value *ForExprAST::codegen() {
    // Emit the start code first, without 'variable' in scope.
    Value *StartVal = Start->codegen();
    if (!StartVal)
        return nullptr;

```

```

// Make the new basic block for the loop header, inserting after current
// block.
Function *TheFunction = Builder.GetInsertBlock()->getParent();
BasicBlock *PreheaderBB = Builder.GetInsertBlock();
BasicBlock *LoopBB = BasicBlock::Create(TheContext, "loop", TheFunction);

// Insert an explicit fall through from the current block to the LoopBB.
Builder.CreateBr(LoopBB);

// Start insertion in LoopBB.
Builder.SetInsertPoint(LoopBB);

// Start the PHI node with an entry for Start.
PHINode *Variable =
    Builder.CreatePHI(Type::getDoubleTy(TheContext), 2, VarName);
Variable->addIncoming(StartVal, PreheaderBB);

// Within the loop, the variable is defined equal to the PHI node. If it
// shadows an existing variable, we have to restore it, so save it now.
Value *OldVal = NamedValues[VarName];
NamedValues[VarName] = Variable;

// Emit the body of the loop. This, like any other expr, can change the
// current BB. Note that we ignore the value computed by the body, but don't
// allow an error.
if (!Body->codegen())
    return nullptr;

// Emit the step value.
Value *StepVal = nullptr;
if (Step) {
    StepVal = Step->codegen();
    if (!StepVal)
        return nullptr;
} else {
    // If not specified, use 1.0.
    StepVal = ConstantFP::get(TheContext, APFloat(1.0));
}

Value *NextVar = Builder.CreateFAdd(Variable, StepVal, "nextvar");

// Compute the end condition.
Value *EndCond = End->codegen();
if (!EndCond)
    return nullptr;

// Convert condition to a bool by comparing non-equal to 0.0.
EndCond = Builder.CreateFCmpONE(
    EndCond, ConstantFP::get(TheContext, APFloat(0.0)), "loopcond");

// Create the "after loop" block and insert it.
BasicBlock *LoopEndBB = Builder.GetInsertBlock();
BasicBlock *AfterBB =
    BasicBlock::Create(TheContext, "afterloop", TheFunction);

// Insert the conditional branch into the end of LoopEndBB.
Builder.CreateCondBr(EndCond, LoopBB, AfterBB);

```

```

    // Any new code will be inserted in AfterBB.
    Builder.SetInsertPoint(AfterBB);

    // Add a new entry to the PHI node for the backedge.
    Variable->addIncoming(NextVar, LoopEndBB);

    // Restore the unshadowed variable.
    if (OldVal)
        NamedValues[VarName] = OldVal;
    else
        NamedValues.erase(VarName);

    // for expr always returns 0.0.
    return Constant::getNullValue(Type::getDoubleTy(TheContext));
}

Function *PrototypeAST::codegen() {
    // Make the function type: double(double,double) etc.
    std::vector<Type *> Doubles(Args.size(), Type::getDoubleTy(TheContext));
    FunctionType *FT =
        FunctionType::get(Type::getDoubleTy(TheContext), Doubles, false);

    Function *F =
        Function::Create(FT, Function::ExternalLinkage, Name, TheModule.get());

    // Set names for all arguments.
    unsigned Idx = 0;
    for (auto &Arg : F->args())
        Arg.setName(Args[Idx++]);

    return F;
}

Function *FunctionAST::codegen() {
    // Transfer ownership of the prototype to the FunctionProtos map, but keep a
    // reference to it for use below.
    auto &P = *Proto;
    FunctionProtos[Proto->getName()] = std::move(Proto);
    Function *TheFunction = getFunction(P.getName());
    if (!TheFunction)
        return nullptr;

    // Create a new basic block to start insertion into.
    BasicBlock *BB = BasicBlock::Create(TheContext, "entry", TheFunction);
    Builder.SetInsertPoint(BB);

    // Record the function arguments in the NamedValues map.
    NamedValues.clear();
    for (auto &Arg : TheFunction->args())
        NamedValues[Arg.getName()] = &Arg;

    if (Value *RetVal = Body->codegen()) {
        // Finish off the function.
        Builder.CreateRet(RetVal);

        // Validate the generated code, checking for consistency.
        verifyFunction(*TheFunction);
    }
}

```

```

    // Run the optimizer on the function.
    TheFPM->run(*TheFunction);

    return TheFunction;
}

// Error reading body, remove function.
TheFunction->eraseFromParent();
return nullptr;
}

//===-----//
// Top-Level parsing and JIT Driver
//===-----//

static void InitializeModuleAndPassManager() {
    // Open a new module.
    TheModule = llvm::make_unique<Module>("my cool jit", TheContext);
    TheModule->setDataLayout(TheJIT->getTargetMachine().createDataLayout());

    // Create a new pass manager attached to it.
    TheFPM = llvm::make_unique<legacy::FunctionPassManager>(TheModule.get());

    // Do simple "peephole" optimizations and bit-twiddling optzns.
    TheFPM->add(createInstructionCombiningPass());
    // Reassociate expressions.
    TheFPM->add(createReassociatePass());
    // Eliminate Common SubExpressions.
    TheFPM->add(createGVNPass());
    // Simplify the control flow graph (deleting unreachable blocks, etc).
    TheFPM->add(createCFGSimplificationPass());

    TheFPM->doInitialization();
}

static void HandleDefinition() {
    if (auto FnAST = ParseDefinition()) {
        if (auto *FnIR = FnAST->codegen()) {
            fprintf(stderr, "Read function definition:");
            FnIR->print(errs());
            fprintf(stderr, "\n");
            TheJIT->addModule(std::move(TheModule));
            InitializeModuleAndPassManager();
        }
    } else {
        // Skip token for error recovery.
        getNextToken();
    }
}

static void HandleExtern() {
    if (auto ProtoAST = ParseExtern()) {
        if (auto *FnIR = ProtoAST->codegen()) {
            fprintf(stderr, "Read extern: ");
            FnIR->print(errs());
            fprintf(stderr, "\n");
            FunctionProtos[ProtoAST->getName()] = std::move(ProtoAST);
        }
    }
}

```



```

    }
  } else {
    // Skip token for error recovery.
    getNextToken();
  }
}

static void HandleTopLevelExpression() {
  // Evaluate a top-level expression into an anonymous function.
  if (auto FnAST = ParseTopLevelExpr()) {
    if (FnAST->codegen()) {
      // JIT the module containing the anonymous expression, keeping a handle so
      // we can free it later.
      auto H = TheJIT->addModule(std::move(TheModule));
      InitializeModuleAndPassManager();

      // Search the JIT for the __anon_expr symbol.
      auto ExprSymbol = TheJIT->findSymbol("__anon_expr");
      assert(ExprSymbol && "Function not found");

      // Get the symbol's address and cast it to the right type (takes no
      // arguments, returns a double) so we can call it as a native function.
      double (*FP)() = (double (*)())(intptr_t)cantFail(ExprSymbol.getAddress());
      fprintf(stderr, "Evaluated to %f\n", FP());

      // Delete the anonymous expression module from the JIT.
      TheJIT->removeModule(H);
    }
  } else {
    // Skip token for error recovery.
    getNextToken();
  }
}

/// top ::= definition | external | expression | ';'
static void MainLoop() {
  while (true) {
    fprintf(stderr, "ready> ");
    switch (CurTok) {
    case tok_eof:
      return;
    case ';': // ignore top-level semicolons.
      getNextToken();
      break;
    case tok_def:
      HandleDefinition();
      break;
    case tok_extern:
      HandleExtern();
      break;
    default:
      HandleTopLevelExpression();
      break;
    }
  }
}

//=====//

```

```
// "Library" functions that can be "extern'd" from user code.
//=====//

#ifdef LLVM_ON_WIN32
#define DLLEXPORT __declspec(dllexport)
#else
#define DLLEXPORT
#endif

/// putchar - putchar that takes a double and returns 0.
extern "C" DLLEXPORT double putchar(double X) {
    fputc((char)X, stderr);
    return 0;
}

/// printfd - printf that takes a double prints it as "%f\n", returning 0.
extern "C" DLLEXPORT double printfd(double X) {
    fprintf(stderr, "%f\n", X);
    return 0;
}

//=====//
// Main driver code.
//=====//

int main() {
    InitializeNativeTarget();
    InitializeNativeTargetAsmPrinter();
    InitializeNativeTargetAsmParser();

    // Install standard binary operators.
    // 1 is lowest precedence.
    BinopPrecedence['<'] = 10;
    BinopPrecedence['+'] = 20;
    BinopPrecedence['-'] = 20;
    BinopPrecedence['*'] = 40; // highest.

    // Prime the first token.
    fprintf(stderr, "ready> ");
    getNextToken();

    TheJIT = llvm::make_unique<KaleidoscopeJIT>();

    InitializeModuleAndPassManager();

    // Run the main "interpreter loop" now.
    MainLoop();

    return 0;
}
```

Next: Extending the language: user-defined operators

Kaleidoscope: Extending the Language: User-defined Operators

- [Chapter 6 Introduction](#)
- [User-defined Operators: the Idea](#)
- [User-defined Binary Operators](#)
- [User-defined Unary Operators](#)
- [Kicking the Tires](#)
- [Full Code Listing](#)

Chapter 6 Introduction

Welcome to Chapter 6 of the “[Implementing a language with LLVM](#)” tutorial. At this point in our tutorial, we now have a fully functional language that is fairly minimal, but also useful. There is still one big problem with it, however. Our language doesn’t have many useful operators (like division, logical negation, or even any comparisons besides less-than).

This chapter of the tutorial takes a wild digression into adding user-defined operators to the simple and beautiful Kaleidoscope language. This digression now gives us a simple and ugly language in some ways, but also a powerful one at the same time. One of the great things about creating your own language is that you get to decide what is good or bad. In this tutorial we’ll assume that it is okay to use this as a way to show some interesting parsing techniques.

At the end of this tutorial, we’ll run through an example Kaleidoscope application that *renders the Mandelbrot set*. This gives an example of what you can build with Kaleidoscope and its feature set.

User-defined Operators: the Idea

The “operator overloading” that we will add to Kaleidoscope is more general than in languages like C++. In C++, you are only allowed to redefine existing operators: you can’t programmatically change the grammar, introduce new operators, change precedence levels, etc. In this chapter, we will add this capability to Kaleidoscope, which will let the user round out the set of operators that are supported.

The point of going into user-defined operators in a tutorial like this is to show the power and flexibility of using a hand-written parser. Thus far, the parser we have been implementing uses recursive descent for most parts of the grammar and operator precedence parsing for the expressions. See [Chapter 2](#) for details. By using operator precedence parsing, it is very easy to allow the programmer to introduce new operators into the grammar: the grammar is dynamically extensible as the JIT runs.

The two specific features we’ll add are programmable unary operators (right now, Kaleidoscope has no unary operators at all) as well as binary operators. An example of this is:

```
# Logical unary not.
def unary!(v)
  if v then
    0
  else
    1;

# Define > with the same precedence as <.
def binary> 10 (LHS RHS)
  RHS < LHS;

# Binary "logical or", (note that it does not "short circuit")
def binary| 5 (LHS RHS)
```

```
if LHS then
  1
else if RHS then
  1
else
  0;

# Define = with slightly lower precedence than relationals.
def binary= 9 (LHS RHS)
  !(LHS < RHS | LHS > RHS);
```

Many languages aspire to being able to implement their standard runtime library in the language itself. In Kaleidoscope, we can implement significant parts of the language in the library!

We will break down implementation of these features into two parts: implementing support for user-defined binary operators and adding unary operators.

User-defined Binary Operators

Adding support for user-defined binary operators is pretty simple with our current framework. We'll first add support for the unary/binary keywords:

```
enum Token {
  ...
  // operators
  tok_binary = -11,
  tok_unary = -12
};
...
static int gettok() {
  ...
  if (IdentifierStr == "for")
    return tok_for;
  if (IdentifierStr == "in")
    return tok_in;
  if (IdentifierStr == "binary")
    return tok_binary;
  if (IdentifierStr == "unary")
    return tok_unary;
  return tok_identifier;
```

This just adds lexer support for the unary and binary keywords, like we did in [previous chapters](#). One nice thing about our current AST, is that we represent binary operators with full generalisation by using their ASCII code as the opcode. For our extended operators, we'll use this same representation, so we don't need any new AST or parser support.

On the other hand, we have to be able to represent the definitions of these new operators, in the “def binary| 5” part of the function definition. In our grammar so far, the “name” for the function definition is parsed as the “prototype” production and into the `PrototypeAST` AST node. To represent our new user-defined operators as prototypes, we have to extend the `PrototypeAST` AST node like this:

```
/// PrototypeAST - This class represents the "prototype" for a function,
/// which captures its argument names as well as if it is an operator.
class PrototypeAST {
  std::string Name;
  std::vector<std::string> Args;
  bool IsOperator;
```

```

unsigned Precedence;  // Precedence if a binary op.

public:
  PrototypeAST(const std::string &name, std::vector<std::string> Args,
              bool IsOperator = false, unsigned Prec = 0)
    : Name(name), Args(std::move(Args)), IsOperator(IsOperator),
      Precedence(Prec) {}

  Function *codegen();
  const std::string &getName() const { return Name; }

  bool isUnaryOp() const { return IsOperator && Args.size() == 1; }
  bool isBinaryOp() const { return IsOperator && Args.size() == 2; }

  char getOperatorName() const {
    assert(isUnaryOp() || isBinaryOp());
    return Name[Name.size() - 1];
  }

  unsigned getBinaryPrecedence() const { return Precedence; }
};

```

Basically, in addition to knowing a name for the prototype, we now keep track of whether it was an operator, and if it was, what precedence level the operator is at. The precedence is only used for binary operators (as you'll see below, it just doesn't apply for unary operators). Now that we have a way to represent the prototype for a user-defined operator, we need to parse it:

```

/// prototype
/// ::= id '(' id* ')'
/// ::= binary LETTER number? (id, id)
static std::unique_ptr<PrototypeAST> ParsePrototype() {
  std::string FnName;

  unsigned Kind = 0;  // 0 = identifier, 1 = unary, 2 = binary.
  unsigned BinaryPrecedence = 30;

  switch (CurTok) {
  default:
    return LogErrorP("Expected function name in prototype");
  case tok_identifier:
    FnName = IdentifierStr;
    Kind = 0;
    getNextToken();
    break;
  case tok_binary:
    getNextToken();
    if (!isascii(CurTok))
      return LogErrorP("Expected binary operator");
    FnName = "binary";
    FnName += (char)CurTok;
    Kind = 2;
    getNextToken();

    // Read the precedence if present.
    if (CurTok == tok_number) {
      if (NumVal < 1 || NumVal > 100)
        return LogErrorP("Invalid precedence: must be 1..100");
      BinaryPrecedence = (unsigned)NumVal;
    }
  }
}

```

```

    getNextToken();
}
break;
}

if (CurTok != '(')
    return LogErrorP("Expected '(' in prototype");

std::vector<std::string> ArgNames;
while (getNextToken() == tok_identifier)
    ArgNames.push_back(IdentifierStr);
if (CurTok != ')')
    return LogErrorP("Expected ')' in prototype");

// success.
getNextToken(); // eat ')'.

// Verify right number of names for operator.
if (Kind && ArgNames.size() != Kind)
    return LogErrorP("Invalid number of operands for operator");

return llvm::make_unique<PrototypeAST>(FnName, std::move(ArgNames), Kind != 0,
                                       BinaryPrecedence);
}

```

This is all fairly straightforward parsing code, and we have already seen a lot of similar code in the past. One interesting part about the code above is the couple lines that set up `FnName` for binary operators. This builds names like “binary@” for a newly defined “@” operator. It then takes advantage of the fact that symbol names in the LLVM symbol table are allowed to have any character in them, including embedded nul characters.

The next interesting thing to add, is codegen support for these binary operators. Given our current structure, this is a simple addition of a default case for our existing binary operator node:

```

Value *BinaryExprAST::codegen() {
    Value *L = LHS->codegen();
    Value *R = RHS->codegen();
    if (!L || !R)
        return nullptr;

    switch (Op) {
    case '+':
        return Builder.CreateFAdd(L, R, "addtmp");
    case '-':
        return Builder.CreateFSub(L, R, "subtmp");
    case '*':
        return Builder.CreateFMul(L, R, "multmp");
    case '<':
        L = Builder.CreateFCmpULT(L, R, "cmptmp");
        // Convert bool 0/1 to double 0.0 or 1.0
        return Builder.CreateUIToFP(L, Type::getDoubleTy(TheContext),
                                   "booltmp");
    default:
        break;
    }

    // If it wasn't a builtin binary operator, it must be a user defined one. Emit
    // a call to it.
    Function *F = getFunction(std::string("binary") + Op);
}

```

```

assert(F && "binary operator not found!");

Value *Ops[2] = { L, R };
return Builder.CreateCall(F, Ops, "binop");
}

```

As you can see above, the new code is actually really simple. It just does a lookup for the appropriate operator in the symbol table and generates a function call to it. Since user-defined operators are just built as normal functions (because the “prototype” boils down to a function with the right name) everything falls into place.

The final piece of code we are missing, is a bit of top-level magic:

```

Function *FunctionAST::codegen() {
    // Transfer ownership of the prototype to the FunctionProtos map, but keep a
    // reference to it for use below.
    auto &P = *Proto;
    FunctionProtos[Proto->getName()] = std::move(Proto);
    Function *TheFunction = getFunction(P.getName());
    if (!TheFunction)
        return nullptr;

    // If this is an operator, install it.
    if (P.isBinaryOp())
        BinopPrecedence[P.getOperatorName()] = P.getBinaryPrecedence();

    // Create a new basic block to start insertion into.
    BasicBlock *BB = BasicBlock::Create(TheContext, "entry", TheFunction);
    ...
}

```

Basically, before codegening a function, if it is a user-defined operator, we register it in the precedence table. This allows the binary operator parsing logic we already have in place to handle it. Since we are working on a fully-general operator precedence parser, this is all we need to do to “extend the grammar”.

Now we have useful user-defined binary operators. This builds a lot on the previous framework we built for other operators. Adding unary operators is a bit more challenging, because we don’t have any framework for it yet - lets see what it takes.

User-defined Unary Operators

Since we don’t currently support unary operators in the Kaleidoscope language, we’ll need to add everything to support them. Above, we added simple support for the ‘unary’ keyword to the lexer. In addition to that, we need an AST node:

```

// UnaryExprAST - Expression class for a unary operator.
class UnaryExprAST : public ExprAST {
    char Opcode;
    std::unique_ptr<ExprAST> Operand;

public:
    UnaryExprAST(char Opcode, std::unique_ptr<ExprAST> Operand)
        : Opcode(Opcode), Operand(std::move(Operand)) {}

    Value *codegen() override;
};

```

This AST node is very simple and obvious by now. It directly mirrors the binary operator AST node, except that it only has one child. With this, we need to add the parsing logic. Parsing a unary operator is pretty simple: we’ll add a new function to do it:

```

/// unary
/// ::= primary
/// ::= '!' unary
static std::unique_ptr<ExprAST> ParseUnary() {
    // If the current token is not an operator, it must be a primary expr.
    if (!isascii(CurTok) || CurTok == '(' || CurTok == ',')
        return ParsePrimary();

    // If this is a unary operator, read it.
    int Opc = CurTok;
    getNextToken();
    if (auto Operand = ParseUnary())
        return llvm::make_unique<UnaryExprAST>(Opc, std::move(Operand));
    return nullptr;
}

```

The grammar we add is pretty straightforward here. If we see a unary operator when parsing a primary operator, we eat the operator as a prefix and parse the remaining piece as another unary operator. This allows us to handle multiple unary operators (e.g. “!!x”). Note that unary operators can’t have ambiguous parses like binary operators can, so there is no need for precedence information.

The problem with this function, is that we need to call `ParseUnary` from somewhere. To do this, we change previous callers of `ParsePrimary` to call `ParseUnary` instead:

```

/// binoprhs
/// ::= ('+' unary)*
static std::unique_ptr<ExprAST> ParseBinOpRHS(int ExprPrec,
                                              std::unique_ptr<ExprAST> LHS) {
    ...
    // Parse the unary expression after the binary operator.
    auto RHS = ParseUnary();
    if (!RHS)
        return nullptr;
    ...
}
/// expression
/// ::= unary binoprhs
///
static std::unique_ptr<ExprAST> ParseExpression() {
    auto LHS = ParseUnary();
    if (!LHS)
        return nullptr;

    return ParseBinOpRHS(0, std::move(LHS));
}

```

With these two simple changes, we are now able to parse unary operators and build the AST for them. Next up, we need to add parser support for prototypes, to parse the unary operator prototype. We extend the binary operator code above with:

```

/// prototype
/// ::= id '(' id* ')'
/// ::= binary LETTER number? (id, id)
/// ::= unary LETTER (id)
static std::unique_ptr<PrototypeAST> ParsePrototype() {
    std::string FnName;

    unsigned Kind = 0; // 0 = identifier, 1 = unary, 2 = binary.

```



```

unsigned BinaryPrecedence = 30;

switch (CurTok) {
default:
    return LogErrorP("Expected function name in prototype");
case tok_identifier:
    FnName = IdentifierStr;
    Kind = 0;
    getNextToken();
    break;
case tok_unary:
    getNextToken();
    if (!isascii(CurTok))
        return LogErrorP("Expected unary operator");
    FnName = "unary";
    FnName += (char)CurTok;
    Kind = 1;
    getNextToken();
    break;
case tok_binary:
    ...

```

As with binary operators, we name unary operators with a name that includes the operator character. This assists us at code generation time. Speaking of, the final piece we need to add is codegen support for unary operators. It looks like this:

```

Value *UnaryExprAST::codegen() {
    Value *OperandV = Operand->codegen();
    if (!OperandV)
        return nullptr;

    Function *F = getFunction(std::string("unary") + Opcode);
    if (!F)
        return LogErrorV("Unknown unary operator");

    return Builder.CreateCall(F, OperandV, "unop");
}

```

This code is similar to, but simpler than, the code for binary operators. It is simpler primarily because it doesn't need to handle any predefined operators.

Kicking the Tires

It is somewhat hard to believe, but with a few simple extensions we've covered in the last chapters, we have grown a real-ish language. With this, we can do a lot of interesting things, including I/O, math, and a bunch of other things. For example, we can now add a nice sequencing operator (printf is defined to print out the specified value and a newline):

```

ready> extern printf(x);
Read extern:
declare double @printf(double)

ready> def binary : 1 (x y) 0; # Low-precedence operator that ignores operands.
...
ready> printf(123) : printf(456) : printf(789);
123.000000
456.000000

```

```
789.000000
Evaluated to 0.000000
```

We can also define a bunch of other “primitive” operations, such as:

```
# Logical unary not.
def unary!(v)
  if v then
    0
  else
    1;

# Unary negate.
def unary-(v)
  0-v;

# Define > with the same precedence as <.
def binary> 10 (LHS RHS)
  RHS < LHS;

# Binary logical or, which does not short circuit.
def binary| 5 (LHS RHS)
  if LHS then
    1
  else if RHS then
    1
  else
    0;

# Binary logical and, which does not short circuit.
def binary& 6 (LHS RHS)
  if !LHS then
    0
  else
    !!RHS;

# Define = with slightly lower precedence than relationals.
def binary = 9 (LHS RHS)
  !(LHS < RHS | LHS > RHS);

# Define ':' for sequencing: as a low-precedence operator that ignores operands
# and just returns the RHS.
def binary : 1 (x y) y;
```

Given the previous if/then/else support, we can also define interesting functions for I/O. For example, the following prints out a character whose “density” reflects the value passed in: the lower the value, the denser the character:

```
ready> extern putchar(char);
...
ready> def printdensity(d)
  if d > 8 then
    putchar(32) # ' '
  else if d > 4 then
    putchar(46) # '.'
  else if d > 2 then
    putchar(43) # '+'
  else
    putchar(42); # '*'
```

```
...
ready> printdensity(1): printdensity(2): printdensity(3):
      printdensity(4): printdensity(5): printdensity(9):
      putchar(10);
****.
Evaluated to 0.000000
```

Based on these simple primitive operations, we can start to define more interesting things. For example, here's a little function that determines the number of iterations it takes for a certain function in the complex plane to diverge:

```
# Determine whether the specific location diverges.
# Solve for  $z = z^2 + c$  in the complex plane.
def mandelconverger(real imag iters creal cimag)
    if iters > 255 | (real*real + imag*imag > 4) then
        iters
    else
        mandelconverger(real*real - imag*imag + creal,
                        2*real*imag + cimag,
                        iters+1, creal, cimag);

# Return the number of iterations required for the iteration to escape
def mandelconverge(real imag)
    mandelconverger(real, imag, 0, real, imag);
```

This “ $z = z_2 + c$ ” function is a beautiful little creature that is the basis for computation of the [Mandelbrot Set](#). Our `mandelconverge` function returns the number of iterations that it takes for a complex orbit to escape, saturating to 255. This is not a very useful function by itself, but if you plot its value over a two-dimensional plane, you can see the Mandelbrot set. Given that we are limited to using `putchard` here, our amazing graphical output is limited, but we can whip together something using the density plotter above:

```
# Compute and plot the mandelbrot set with the specified 2 dimensional range
# info.
def mandelhelp(xmin xmax xstep  ymin ymax ystep)
    for y = ymin, y < ymax, ystep in (
        (for x = xmin, x < xmax, xstep in
            printdensity(mandelconverge(x,y)))
        : putchar(10)
    )

# mandel - This is a convenient helper function for plotting the mandelbrot set
# from the specified position with the specified Magnification.
def mandel(realstart imagstart realmag imagmag)
    mandelhelp(realstart, realstart+realmag*78, realmag,
                imagstart, imagstart+imagmag*40, imagmag);
```

Given this, we can try plotting out the mandelbrot set! Lets try it out:

```
ready> mandel(-2.3, -1.3, 0.05, 0.07);  
*****+++++*****  
*****+++++*****  
*****+++++*****  
*****+++++*****  
*****+++++. . . +++++*****  
*****+++++. . . +++++*****  
*****+++++. . . +++++*****  
*****+++++. . . +++++*****  
*****+++++. . . +++++*****  
*****+++++. . . +++++*****
```


[illegible]

```

Evaluated to 0.000000
ready> ^D

```

```

...+++++++
...+++++++

```

At this point, you may be starting to realize that Kaleidoscope is a real and powerful language. It may not be self-similar :), but it can be used to plot things that are!

With this, we conclude the “adding user-defined operators” chapter of the tutorial. We have successfully augmented our language, adding the ability to extend the language in the library, and we have shown how this can be used to build a simple but interesting end-user application in Kaleidoscope. At this point, Kaleidoscope can build a variety of applications that are functional and can call functions with side-effects, but it can’t actually define and mutate a variable itself.

Strikingly, variable mutation is an important feature of some languages, and it is not at all obvious how to [add support for mutable variables](#) without having to add an “SSA construction” phase to your front-end. In the next chapter, we will describe how you can add variable mutation without building SSA in your front-end.

Full Code Listing

Here is the complete code listing for our running example, enhanced with the support for user-defined operators. To build this example, use:

```

# Compile
clang++ -g toy.cpp `llvm-config --cxxflags --ldflags --system-libs --libs core mcjit_
↪native` -O3 -o toy
# Run
./toy

```

On some platforms, you will need to specify `-rdynamic` or `-Wl,--export-dynamic` when linking. This ensures that symbols defined in the main executable are exported to the dynamic linker and so are available for symbol resolution at run time. This is not needed if you compile your support code into a shared library, although doing that will cause problems on Windows.

Here is the code:

```

#include "llvm/ADT/APFloat.h"
#include "llvm/ADT/STLExtras.h"
#include "llvm/IR/BasicBlock.h"
#include "llvm/IR/Constants.h"
#include "llvm/IR/DerivedTypes.h"
#include "llvm/IR/Function.h"
#include "llvm/IR/Instructions.h"
#include "llvm/IR/IRBuilder.h"
#include "llvm/IR/LLVMContext.h"
#include "llvm/IR/LegacyPassManager.h"
#include "llvm/IR/Module.h"
#include "llvm/IR/Type.h"
#include "llvm/IR/Verifier.h"
#include "llvm/Support/TargetSelect.h"
#include "llvm/Target/TargetMachine.h"
#include "llvm/Transforms/Scalar.h"
#include "llvm/Transforms/Scalar/GVN.h"
#include "../include/KaleidoscopeJIT.h"
#include <algorithm>
#include <cassert>
#include <cctype>

```

```

#include <cstdint>
#include <cstdio>
#include <cstdlib>
#include <map>
#include <memory>
#include <string>
#include <vector>

using namespace llvm;
using namespace llvm::orc;

//===----- Lexer
// Lexer
//===-----

// The lexer returns tokens [0-255] if it is an unknown character, otherwise one
// of these for known things.
enum Token {
    tok_eof = -1,

    // commands
    tok_def = -2,
    tokExtern = -3,

    // primary
    tok_identifier = -4,
    tok_number = -5,

    // control
    tok_if = -6,
    tok_then = -7,
    tok_else = -8,
    tok_for = -9,
    tok_in = -10,

    // operators
    tok_binary = -11,
    tok_unary = -12
};

static std::string IdentifierStr; // Filled in if tok_identifier
static double NumVal;           // Filled in if tok_number

/// gettok - Return the next token from standard input.
static int gettok() {
    static int LastChar = ' ';

    // Skip any whitespace.
    while (isspace(LastChar))
        LastChar = getchar();

    if (isalpha(LastChar)) { // identifier: [a-zA-Z][a-zA-Z0-9]*
        IdentifierStr = LastChar;
        while (isalnum((LastChar = getchar())))
            IdentifierStr += LastChar;

        if (IdentifierStr == "def")
            return tok_def;
    }

```

```

    if (IdentifierStr == "extern")
        return tok_extern;
    if (IdentifierStr == "if")
        return tok_if;
    if (IdentifierStr == "then")
        return tok_then;
    if (IdentifierStr == "else")
        return tok_else;
    if (IdentifierStr == "for")
        return tok_for;
    if (IdentifierStr == "in")
        return tok_in;
    if (IdentifierStr == "binary")
        return tok_binary;
    if (IdentifierStr == "unary")
        return tok_unary;
    return tok_identifier;
}

if (isdigit>LastChar) || LastChar == '.') { // Number: [0-9.]+
    std::string NumStr;
    do {
        NumStr += LastChar;
        LastChar = getchar();
    } while (isdigit>LastChar) || LastChar == '.');

    NumVal = strtod(NumStr.c_str(), nullptr);
    return tok_number;
}

if (LastChar == '#') {
    // Comment until end of line.
    do
        LastChar = getchar();
    while (LastChar != EOF && LastChar != '\n' && LastChar != '\r');

    if (LastChar != EOF)
        return gettok();
}

// Check for end of file. Don't eat the EOF.
if (LastChar == EOF)
    return tok_eof;

// Otherwise, just return the character as its ascii value.
int ThisChar = LastChar;
LastChar = getchar();
return ThisChar;
}

//===-----
// Abstract Syntax Tree (aka Parse Tree)
//===-----

namespace {

/// ExprAST - Base class for all expression nodes.
class ExprAST {

```



```

public:
    virtual ~ExprAST() = default;

    virtual Value *codegen() = 0;
};

/// NumberExprAST - Expression class for numeric literals like "1.0".
class NumberExprAST : public ExprAST {
    double Val;

public:
    NumberExprAST(double Val) : Val(Val) {}

    Value *codegen() override;
};

/// VariableExprAST - Expression class for referencing a variable, like "a".
class VariableExprAST : public ExprAST {
    std::string Name;

public:
    VariableExprAST(const std::string &Name) : Name(Name) {}

    Value *codegen() override;
};

/// UnaryExprAST - Expression class for a unary operator.
class UnaryExprAST : public ExprAST {
    char Opcode;
    std::unique_ptr<ExprAST> Operand;

public:
    UnaryExprAST(char Opcode, std::unique_ptr<ExprAST> Operand)
        : Opcode(Opcode), Operand(std::move(Operand)) {}

    Value *codegen() override;
};

/// BinaryExprAST - Expression class for a binary operator.
class BinaryExprAST : public ExprAST {
    char Op;
    std::unique_ptr<ExprAST> LHS, RHS;

public:
    BinaryExprAST(char Op, std::unique_ptr<ExprAST> LHS,
                  std::unique_ptr<ExprAST> RHS)
        : Op(Op), LHS(std::move(LHS)), RHS(std::move(RHS)) {}

    Value *codegen() override;
};

/// CallExprAST - Expression class for function calls.
class CallExprAST : public ExprAST {
    std::string Callee;
    std::vector<std::unique_ptr<ExprAST>> Args;

public:
    CallExprAST(const std::string &Callee,

```

```

        std::vector<std::unique_ptr<ExprAST>> Args)
        : Callee(Callee), Args(std::move(Args)) {}

    Value *codegen() override;
};

/// IfExprAST - Expression class for if/then/else.
class IfExprAST : public ExprAST {
    std::unique_ptr<ExprAST> Cond, Then, Else;

public:
    IfExprAST(std::unique_ptr<ExprAST> Cond, std::unique_ptr<ExprAST> Then,
               std::unique_ptr<ExprAST> Else)
        : Cond(std::move(Cond)), Then(std::move(Then)), Else(std::move(Else)) {}

    Value *codegen() override;
};

/// ForExprAST - Expression class for for/in.
class ForExprAST : public ExprAST {
    std::string VarName;
    std::unique_ptr<ExprAST> Start, End, Step, Body;

public:
    ForExprAST(const std::string &VarName, std::unique_ptr<ExprAST> Start,
               std::unique_ptr<ExprAST> End, std::unique_ptr<ExprAST> Step,
               std::unique_ptr<ExprAST> Body)
        : VarName(VarName), Start(std::move(Start)), End(std::move(End)),
          Step(std::move(Step)), Body(std::move(Body)) {}

    Value *codegen() override;
};

/// PrototypeAST - This class represents the "prototype" for a function,
/// which captures its name, and its argument names (thus implicitly the number
/// of arguments the function takes), as well as if it is an operator.
class PrototypeAST {
    std::string Name;
    std::vector<std::string> Args;
    bool IsOperator;
    unsigned Precedence; // Precedence if a binary op.

public:
    PrototypeAST(const std::string &Name, std::vector<std::string> Args,
                 bool IsOperator = false, unsigned Prec = 0)
        : Name(Name), Args(std::move(Args)), IsOperator(IsOperator),
          Precedence(Prec) {}

    Function *codegen();
    const std::string &getName() const { return Name; }

    bool isUnaryOp() const { return IsOperator && Args.size() == 1; }
    bool isBinaryOp() const { return IsOperator && Args.size() == 2; }

    char getOperatorName() const {
        assert(isUnaryOp() || isBinaryOp());
        return Name[Name.size() - 1];
    }
}

```

```

    unsigned getBinaryPrecedence() const { return Precedence; }
};

/// FunctionAST - This class represents a function definition itself.
class FunctionAST {
    std::unique_ptr<PrototypeAST> Proto;
    std::unique_ptr<ExprAST> Body;

public:
    FunctionAST(std::unique_ptr<PrototypeAST> Proto,
                std::unique_ptr<ExprAST> Body)
        : Proto(std::move(Proto)), Body(std::move(Body)) {}

    Function *codegen();
};

} // end anonymous namespace

//===-----
// Parser
//===-----

/// CurTok/getNextToken - Provide a simple token buffer. CurTok is the current
/// token the parser is looking at. getNextToken reads another token from the
/// lexer and updates CurTok with its results.
static int CurTok;
static int getNextToken() { return CurTok = gettok(); }

/// BinopPrecedence - This holds the precedence for each binary operator that is
/// defined.
static std::map<char, int> BinopPrecedence;

/// GetTokPrecedence - Get the precedence of the pending binary operator token.
static int GetTokPrecedence() {
    if (!isascii(CurTok))
        return -1;

    // Make sure it's a declared binop.
    int TokPrec = BinopPrecedence[CurTok];
    if (TokPrec <= 0)
        return -1;
    return TokPrec;
}

/// Error* - These are little helper functions for error handling.
std::unique_ptr<ExprAST> LogError(const char *Str) {
    fprintf(stderr, "Error: %s\n", Str);
    return nullptr;
}

std::unique_ptr<PrototypeAST> LogErrorP(const char *Str) {
    LogError(Str);
    return nullptr;
}

static std::unique_ptr<ExprAST> ParseExpression();

```

```

/// numberexpr ::= number
static std::unique_ptr<ExprAST> ParseNumberExpr() {
    auto Result = llvm::make_unique<NumberExprAST>(NumVal);
    getNextToken(); // consume the number
    return std::move(Result);
}

/// parenexpr ::= '(' expression ')'
static std::unique_ptr<ExprAST> ParseParenExpr() {
    getNextToken(); // eat '('
    auto V = ParseExpression();
    if (!V)
        return nullptr;

    if (CurTok != ')')
        return LogError("expected ')'");
    getNextToken(); // eat ')'
    return V;
}

/// identifierexpr
/// ::= identifier
/// ::= identifier '(' expression* ')'
static std::unique_ptr<ExprAST> ParseIdentifierExpr() {
    std::string IdName = IdentifierStr;

    getNextToken(); // eat identifier.

    if (CurTok != '(') // Simple variable ref.
        return llvm::make_unique<VariableExprAST>(IdName);

    // Call.
    getNextToken(); // eat '('
    std::vector<std::unique_ptr<ExprAST>> Args;
    if (CurTok != ')') {
        while (true) {
            if (auto Arg = ParseExpression())
                Args.push_back(std::move(Arg));
            else
                return nullptr;

            if (CurTok == ')')
                break;

            if (CurTok != ',')
                return LogError("Expected ')' or ',' in argument list");
            getNextToken();
        }
    }

    // Eat the ')'
    getNextToken();

    return llvm::make_unique<CallExprAST>(IdName, std::move(Args));
}

/// ifexpr ::= 'if' expression 'then' expression 'else' expression
static std::unique_ptr<ExprAST> ParseIfExpr() {

```

```

getNextToken(); // eat the if.

// condition.
auto Cond = ParseExpression();
if (!Cond)
    return nullptr;

if (CurTok != tok_then)
    return LogError("expected then");
getNextToken(); // eat the then

auto Then = ParseExpression();
if (!Then)
    return nullptr;

if (CurTok != tok_else)
    return LogError("expected else");

getNextToken();

auto Else = ParseExpression();
if (!Else)
    return nullptr;

return llvm::make_unique<IfExprAST>(std::move(Cond), std::move(Then),
                                   std::move(Else));
}

/// forexpr ::= 'for' identifier '=' expr ',' expr (',' expr)? 'in' expression
static std::unique_ptr<ExprAST> ParseForExpr() {
    getNextToken(); // eat the for.

    if (CurTok != tok_identifier)
        return LogError("expected identifier after for");

    std::string IdName = IdentifierStr;
    getNextToken(); // eat identifier.

    if (CurTok != '=')
        return LogError("expected '=' after for");
    getNextToken(); // eat '='.

    auto Start = ParseExpression();
    if (!Start)
        return nullptr;
    if (CurTok != ',')
        return LogError("expected ',' after for start value");
    getNextToken();

    auto End = ParseExpression();
    if (!End)
        return nullptr;

    // The step value is optional.
    std::unique_ptr<ExprAST> Step;
    if (CurTok == ',') {
        getNextToken();
        Step = ParseExpression();
    }

```

```

    if (!Step)
        return nullptr;
}

if (CurTok != tok_in)
    return LogError("expected 'in' after for");
getNextToken(); // eat 'in'.

auto Body = ParseExpression();
if (!Body)
    return nullptr;

return llvm::make_unique<ForExprAST>(IdName, std::move(Start), std::move(End),
                                     std::move(Step), std::move(Body));
}

/// primary
/// ::= identifierexpr
/// ::= numberexpr
/// ::= parenexpr
/// ::= ifexpr
/// ::= forexpr
static std::unique_ptr<ExprAST> ParsePrimary() {
    switch (CurTok) {
    default:
        return LogError("unknown token when expecting an expression");
    case tok_identifier:
        return ParseIdentifierExpr();
    case tok_number:
        return ParseNumberExpr();
    case '(':
        return ParseParenExpr();
    case tok_if:
        return ParseIfExpr();
    case tok_for:
        return ParseForExpr();
    }
}

/// unary
/// ::= primary
/// ::= '!' unary
static std::unique_ptr<ExprAST> ParseUnary() {
    // If the current token is not an operator, it must be a primary expr.
    if (!isascii(CurTok) || CurTok == '(' || CurTok == ',')
        return ParsePrimary();

    // If this is a unary operator, read it.
    int Opc = CurTok;
    getNextToken();
    if (auto Operand = ParseUnary())
        return llvm::make_unique<UnaryExprAST>(Opc, std::move(Operand));
    return nullptr;
}

/// binoprhs
/// ::= ('+' unary)*
static std::unique_ptr<ExprAST> ParseBinOpRHS(int ExprPrec,

```

```

std::unique_ptr<ExprAST> LHS) {
    // If this is a binop, find its precedence.
    while (true) {
        int TokPrec = GetTokPrecedence();

        // If this is a binop that binds at least as tightly as the current binop,
        // consume it, otherwise we are done.
        if (TokPrec < ExprPrec)
            return LHS;

        // Okay, we know this is a binop.
        int BinOp = CurTok;
        getNextToken(); // eat binop

        // Parse the unary expression after the binary operator.
        auto RHS = ParseUnary();
        if (!RHS)
            return nullptr;

        // If BinOp binds less tightly with RHS than the operator after RHS, let
        // the pending operator take RHS as its LHS.
        int NextPrec = GetTokPrecedence();
        if (TokPrec < NextPrec) {
            RHS = ParseBinOpRHS(TokPrec + 1, std::move(RHS));
            if (!RHS)
                return nullptr;
        }

        // Merge LHS/RHS.
        LHS =
            llvm::make_unique<BinaryExprAST>(BinOp, std::move(LHS), std::move(RHS));
    }
}

/// expression
/// ::= unary binoprhs
///
static std::unique_ptr<ExprAST> ParseExpression() {
    auto LHS = ParseUnary();
    if (!LHS)
        return nullptr;

    return ParseBinOpRHS(0, std::move(LHS));
}

/// prototype
/// ::= id '(' id* ')'
/// ::= binary LETTER number? (id, id)
/// ::= unary LETTER (id)
static std::unique_ptr<PrototypeAST> ParsePrototype() {
    std::string FnName;

    unsigned Kind = 0; // 0 = identifier, 1 = unary, 2 = binary.
    unsigned BinaryPrecedence = 30;

    switch (CurTok) {
    default:
        return LogErrorP("Expected function name in prototype");

```

```

case tok_identifier:
    FnName = IdentifierStr;
    Kind = 0;
    getNextToken();
    break;
case tok_unary:
    getNextToken();
    if (!isascii(CurTok))
        return LogErrorP("Expected unary operator");
    FnName = "unary";
    FnName += (char)CurTok;
    Kind = 1;
    getNextToken();
    break;
case tok_binary:
    getNextToken();
    if (!isascii(CurTok))
        return LogErrorP("Expected binary operator");
    FnName = "binary";
    FnName += (char)CurTok;
    Kind = 2;
    getNextToken();

    // Read the precedence if present.
    if (CurTok == tok_number) {
        if (NumVal < 1 || NumVal > 100)
            return LogErrorP("Invalid precedence: must be 1..100");
        BinaryPrecedence = (unsigned)NumVal;
        getNextToken();
    }
    break;
}

if (CurTok != '(')
    return LogErrorP("Expected '(' in prototype");

std::vector<std::string> ArgNames;
while (getNextToken() == tok_identifier)
    ArgNames.push_back(IdentifierStr);
if (CurTok != ')')
    return LogErrorP("Expected ')' in prototype");

// success.
getNextToken(); // eat ')'.

// Verify right number of names for operator.
if (Kind && ArgNames.size() != Kind)
    return LogErrorP("Invalid number of operands for operator");

return llvm::make_unique<PrototypeAST>(FnName, ArgNames, Kind != 0,
                                       BinaryPrecedence);
}

// definition ::= 'def' prototype expression
static std::unique_ptr<FunctionAST> ParseDefinition() {
    getNextToken(); // eat def.
    auto Proto = ParsePrototype();
    if (!Proto)

```



```

    return nullptr;

    if (auto E = ParseExpression())
        return llvm::make_unique<FunctionAST>(std::move(Proto), std::move(E));
    return nullptr;
}

/// toplevelexpr ::= expression
static std::unique_ptr<FunctionAST> ParseTopLevelExpr() {
    if (auto E = ParseExpression()) {
        // Make an anonymous proto.
        auto Proto = llvm::make_unique<PrototypeAST>("__anon_expr",
                                                    std::vector<std::string>());
        return llvm::make_unique<FunctionAST>(std::move(Proto), std::move(E));
    }
    return nullptr;
}

/// external ::= 'extern' prototype
static std::unique_ptr<PrototypeAST> ParseExtern() {
    getNextToken(); // eat extern.
    return ParsePrototype();
}

//===-----
// Code Generation
//===-----

static LLVMContext TheContext;
static IRBuilder<> Builder(TheContext);
static std::unique_ptr<Module> TheModule;
static std::map<std::string, Value*> NamedValues;
static std::unique_ptr<legacy::FunctionPassManager> TheFPM;
static std::unique_ptr<KaleidoscopeJIT> TheJIT;
static std::map<std::string, std::unique_ptr<PrototypeAST>> FunctionProtos;

Value *LogErrorV(const char *Str) {
    LogError(Str);
    return nullptr;
}

Function *getFunction(std::string Name) {
    // First, see if the function has already been added to the current module.
    if (auto *F = TheModule->getFunction(Name))
        return F;

    // If not, check whether we can codegen the declaration from some existing
    // prototype.
    auto FI = FunctionProtos.find(Name);
    if (FI != FunctionProtos.end())
        return FI->second->codegen();

    // If no existing prototype exists, return null.
    return nullptr;
}

Value *NumberExprAST::codegen() {
    return ConstantFP::get(TheContext, APFloat(Val));
}

```

```
}

Value *VariableExprAST::codegen() {
    // Look this variable up in the function.
    Value *V = NamedValues[Name];
    if (!V)
        return LogErrorV("Unknown variable name");
    return V;
}

Value *UnaryExprAST::codegen() {
    Value *OperandV = Operand->codegen();
    if (!OperandV)
        return nullptr;

    Function *F = getFunction(std::string("unary") + Opcode);
    if (!F)
        return LogErrorV("Unknown unary operator");

    return Builder.CreateCall(F, OperandV, "unop");
}

Value *BinaryExprAST::codegen() {
    Value *L = LHS->codegen();
    Value *R = RHS->codegen();
    if (!L || !R)
        return nullptr;

    switch (Op) {
    case '+':
        return Builder.CreateFAdd(L, R, "addtmp");
    case '-':
        return Builder.CreateFSub(L, R, "subtmp");
    case '*':
        return Builder.CreateFMul(L, R, "multmp");
    case '<':
        L = Builder.CreateFCmpULT(L, R, "cmptmp");
        // Convert bool 0/1 to double 0.0 or 1.0
        return Builder.CreateUIToFP(L, Type::getDoubleTy(TheContext), "booltmp");
    default:
        break;
    }

    // If it wasn't a builtin binary operator, it must be a user defined one. Emit
    // a call to it.
    Function *F = getFunction(std::string("binary") + Op);
    assert(F && "binary operator not found!");

    Value *Ops[] = {L, R};
    return Builder.CreateCall(F, Ops, "binop");
}

Value *CallExprAST::codegen() {
    // Look up the name in the global module table.
    Function *CalleeF = getFunction(Callee);
    if (!CalleeF)
        return LogErrorV("Unknown function referenced");
}
```

```

// If argument mismatch error.
if (CalleeF->arg_size() != Args.size())
    return LogErrorV("Incorrect # arguments passed");

std::vector<Value *> ArgsV;
for (unsigned i = 0, e = Args.size(); i != e; ++i) {
    ArgsV.push_back(Args[i]->codegen());
    if (!ArgsV.back())
        return nullptr;
}

return Builder.CreateCall(CalleeF, ArgsV, "calltmp");
}

Value *IfExprAST::codegen() {
    Value *CondV = Cond->codegen();
    if (!CondV)
        return nullptr;

    // Convert condition to a bool by comparing non-equal to 0.0.
    CondV = Builder.CreateFCmpONE(
        CondV, ConstantFP::get(TheContext, APFloat(0.0)), "ifcond");

    Function *TheFunction = Builder.GetInsertBlock()->getParent();

    // Create blocks for the then and else cases. Insert the 'then' block at the
// end of the function.
    BasicBlock *ThenBB = BasicBlock::Create(TheContext, "then", TheFunction);
    BasicBlock *ElseBB = BasicBlock::Create(TheContext, "else");
    BasicBlock *MergeBB = BasicBlock::Create(TheContext, "ifcont");

    Builder.CreateCondBr(CondV, ThenBB, ElseBB);

    // Emit then value.
    Builder.SetInsertPoint(ThenBB);

    Value *ThenV = Then->codegen();
    if (!ThenV)
        return nullptr;

    Builder.CreateBr(MergeBB);
    // Codegen of 'Then' can change the current block, update ThenBB for the PHI.
    ThenBB = Builder.GetInsertBlock();

    // Emit else block.
    TheFunction->getBasicBlockList().push_back(ElseBB);
    Builder.SetInsertPoint(ElseBB);

    Value *ElseV = Else->codegen();
    if (!ElseV)
        return nullptr;

    Builder.CreateBr(MergeBB);
    // Codegen of 'Else' can change the current block, update ElseBB for the PHI.
    ElseBB = Builder.GetInsertBlock();

    // Emit merge block.
    TheFunction->getBasicBlockList().push_back(MergeBB);

```

```

    Builder.SetInsertPoint(MergeBB);
    PHINode *PN = Builder.CreatePHI(Type::getDoubleTy(TheContext), 2, "iftmp");

    PN->addIncoming(ThenV, ThenBB);
    PN->addIncoming(ElseV, ElseBB);
    return PN;
}

// Output for-loop as:
// ...
// start = startexpr
// goto loop
// loop:
//   variable = phi [start, loopheader], [nextvariable, loopend]
//   ...
//   bodyexpr
//   ...
// loopend:
//   step = stepexpr
//   nextvariable = variable + step
//   endcond = endexpr
//   br endcond, loop, endloop
// outloop:
Value *ForExprAST::codegen() {
    // Emit the start code first, without 'variable' in scope.
    Value *StartVal = Start->codegen();
    if (!StartVal)
        return nullptr;

    // Make the new basic block for the loop header, inserting after current
    // block.
    Function *TheFunction = Builder.GetInsertBlock()->getParent();
    BasicBlock *PreheaderBB = Builder.GetInsertBlock();
    BasicBlock *LoopBB = BasicBlock::Create(TheContext, "loop", TheFunction);

    // Insert an explicit fall through from the current block to the LoopBB.
    Builder.CreateBr(LoopBB);

    // Start insertion in LoopBB.
    Builder.SetInsertPoint(LoopBB);

    // Start the PHI node with an entry for Start.
    PHINode *Variable =
        Builder.CreatePHI(Type::getDoubleTy(TheContext), 2, VarName);
    Variable->addIncoming(StartVal, PreheaderBB);

    // Within the loop, the variable is defined equal to the PHI node. If it
    // shadows an existing variable, we have to restore it, so save it now.
    Value *OldVal = NamedValues[VarName];
    NamedValues[VarName] = Variable;

    // Emit the body of the loop. This, like any other expr, can change the
    // current BB. Note that we ignore the value computed by the body, but don't
    // allow an error.
    if (!Body->codegen())
        return nullptr;

    // Emit the step value.

```

```

Value *StepVal = nullptr;
if (Step) {
    StepVal = Step->codegen();
    if (!StepVal)
        return nullptr;
} else {
    // If not specified, use 1.0.
    StepVal = ConstantFP::get(TheContext, APFloat(1.0));
}

Value *NextVar = Builder.CreateFAdd(Variable, StepVal, "nextvar");

// Compute the end condition.
Value *EndCond = End->codegen();
if (!EndCond)
    return nullptr;

// Convert condition to a bool by comparing non-equal to 0.0.
EndCond = Builder.CreateFCmpONE(
    EndCond, ConstantFP::get(TheContext, APFloat(0.0)), "loopcond");

// Create the "after loop" block and insert it.
BasicBlock *LoopEndBB = Builder.GetInsertBlock();
BasicBlock *AfterBB =
    BasicBlock::Create(TheContext, "afterloop", TheFunction);

// Insert the conditional branch into the end of LoopEndBB.
Builder.CreateCondBr(EndCond, LoopBB, AfterBB);

// Any new code will be inserted in AfterBB.
Builder.SetInsertPoint(AfterBB);

// Add a new entry to the PHI node for the backedge.
Variable->addIncoming(NextVar, LoopEndBB);

// Restore the unshadowed variable.
if (OldVal)
    NamedValues[VarName] = OldVal;
else
    NamedValues.erase(VarName);

// for expr always returns 0.0.
return Constant::getNullValue(Type::getDoubleTy(TheContext));
}

Function *PrototypeAST::codegen() {
    // Make the function type: double(double,double) etc.
    std::vector<Type *> Doubles(Args.size(), Type::getDoubleTy(TheContext));
    FunctionType *FT =
        FunctionType::get(Type::getDoubleTy(TheContext), Doubles, false);

    Function *F =
        Function::Create(FT, Function::ExternalLinkage, Name, TheModule.get());

    // Set names for all arguments.
    unsigned Idx = 0;
    for (auto &Arg : F->args())
        Arg.setName(Args[Idx++]);
}

```

```

    return F;
}

Function *FunctionAST::codegen() {
    // Transfer ownership of the prototype to the FunctionProtos map, but keep a
    // reference to it for use below.
    auto &P = *Proto;
    FunctionProtos[Proto->getName()] = std::move(Proto);
    Function *TheFunction = getFunction(P.getName());
    if (!TheFunction)
        return nullptr;

    // If this is an operator, install it.
    if (P.isBinaryOp())
        BinopPrecedence[P.getOperatorName()] = P.getBinaryPrecedence();

    // Create a new basic block to start insertion into.
    BasicBlock *BB = BasicBlock::Create(TheContext, "entry", TheFunction);
    Builder.SetInsertPoint(BB);

    // Record the function arguments in the NamedValues map.
    NamedValues.clear();
    for (auto &Arg : TheFunction->args())
        NamedValues[Arg.getName()] = &Arg;

    if (Value *RetVal = Body->codegen()) {
        // Finish off the function.
        Builder.CreateRet(RetVal);

        // Validate the generated code, checking for consistency.
        verifyFunction(*TheFunction);

        // Run the optimizer on the function.
        TheFPM->run(*TheFunction);

        return TheFunction;
    }

    // Error reading body, remove function.
    TheFunction->eraseFromParent();

    if (P.isBinaryOp())
        BinopPrecedence.erase(P.getOperatorName());
    return nullptr;
}

//===-----//
// Top-Level parsing and JIT Driver
//===-----//

static void InitializeModuleAndPassManager() {
    // Open a new module.
    TheModule = llvm::make_unique<Module>("my cool jit", TheContext);
    TheModule->setDataLayout(TheJIT->getTargetMachine().createDataLayout());

    // Create a new pass manager attached to it.
    TheFPM = llvm::make_unique<legacy::FunctionPassManager>(TheModule.get());
}

```

```

// Do simple "peephole" optimizations and bit-twiddling optzns.
TheFPM->add(createInstructionCombiningPass());
// Reassociate expressions.
TheFPM->add(createReassociatePass());
// Eliminate Common SubExpressions.
TheFPM->add(createGVNPass());
// Simplify the control flow graph (deleting unreachable blocks, etc).
TheFPM->add(createCFGSimplificationPass());

TheFPM->doInitialization();
}

static void HandleDefinition() {
    if (auto FnAST = ParseDefinition()) {
        if (auto *FnIR = FnAST->codegen()) {
            fprintf(stderr, "Read function definition:");
            FnIR->print(errs());
            fprintf(stderr, "\n");
            TheJIT->addModule(std::move(TheModule));
            InitializeModuleAndPassManager();
        }
    } else {
        // Skip token for error recovery.
        getNextToken();
    }
}

static void HandleExtern() {
    if (auto ProtoAST = ParseExtern()) {
        if (auto *FnIR = ProtoAST->codegen()) {
            fprintf(stderr, "Read extern: ");
            FnIR->print(errs());
            fprintf(stderr, "\n");
            FunctionProtos[ProtoAST->getName()] = std::move(ProtoAST);
        }
    } else {
        // Skip token for error recovery.
        getNextToken();
    }
}

static void HandleTopLevelExpression() {
    // Evaluate a top-level expression into an anonymous function.
    if (auto FnAST = ParseTopLevelExpr()) {
        if (FnAST->codegen()) {
            // JIT the module containing the anonymous expression, keeping a handle so
            // we can free it later.
            auto H = TheJIT->addModule(std::move(TheModule));
            InitializeModuleAndPassManager();

            // Search the JIT for the __anon_expr symbol.
            auto ExprSymbol = TheJIT->findSymbol("__anon_expr");
            assert(ExprSymbol && "Function not found");

            // Get the symbol's address and cast it to the right type (takes no
            // arguments, returns a double) so we can call it as a native function.
            double (*FP)() = (double (*)())(intptr_t)cantFail(ExprSymbol.getAddress());

```

```
        fprintf(stderr, "Evaluated to %f\n", FP());

        // Delete the anonymous expression module from the JIT.
        TheJIT->removeModule(H);
    }
} else {
    // Skip token for error recovery.
    getNextToken();
}
}

/// top ::= definition | external | expression | ';'
static void MainLoop() {
    while (true) {
        fprintf(stderr, "ready> ");
        switch (CurTok) {
            case tok_eof:
                return;
            case ';': // ignore top-level semicolons.
                getNextToken();
                break;
            case tok_def:
                HandleDefinition();
                break;
            case tokExtern:
                HandleExtern();
                break;
            default:
                HandleTopLevelExpression();
                break;
        }
    }
}

//===-----
// "Library" functions that can be "extern'd" from user code.
//===-----

#ifdef LLVM_ON_WIN32
#define DLLEXPORT __declspec(dllexport)
#else
#define DLLEXPORT
#endif

/// putchar - putchar that takes a double and returns 0.
extern "C" DLLEXPORT double putchar(double X) {
    fputc((char)X, stderr);
    return 0;
}

/// printfd - printf that takes a double prints it as "%f\n", returning 0.
extern "C" DLLEXPORT double printfd(double X) {
    fprintf(stderr, "%f\n", X);
    return 0;
}

//===-----
// Main driver code.
```



```
//=====//

int main() {
    InitializeNativeTarget();
    InitializeNativeTargetAsmPrinter();
    InitializeNativeTargetAsmParser();

    // Install standard binary operators.
    // 1 is lowest precedence.
    BinopPrecedence['<'] = 10;
    BinopPrecedence['+'] = 20;
    BinopPrecedence['-'] = 20;
    BinopPrecedence['*'] = 40; // highest.

    // Prime the first token.
    fprintf(stderr, "ready> ");
    getNextToken();

    TheJIT = llvm::make_unique<KaleidoscopeJIT>();

    InitializeModuleAndPassManager();

    // Run the main "interpreter loop" now.
    MainLoop();

    return 0;
}
```

Next: Extending the language: mutable variables / SSA construction

Kaleidoscope: Extending the Language: Mutable Variables

- *Chapter 7 Introduction*
- *Why is this a hard problem?*
- *Memory in LLVM*
- *Mutable Variables in Kaleidoscope*
- *Adjusting Existing Variables for Mutation*
- *New Assignment Operator*
- *User-defined Local Variables*
- *Full Code Listing*

Chapter 7 Introduction

Welcome to Chapter 7 of the “Implementing a language with LLVM” tutorial. In chapters 1 through 6, we’ve built a very respectable, albeit simple, **functional programming language**. In our journey, we learned some parsing techniques, how to build and represent an AST, how to build LLVM IR, and how to optimize the resultant code as well as JIT compile it.

While Kaleidoscope is interesting as a functional language, the fact that it is functional makes it “too easy” to generate LLVM IR for it. In particular, a functional language makes it very easy to build LLVM IR directly in [SSA form](#). Since LLVM requires that the input code be in SSA form, this is a very nice property and it is often unclear to newcomers how to generate code for an imperative language with mutable variables.

The short (and happy) summary of this chapter is that there is no need for your front-end to build SSA form: LLVM provides highly tuned and well tested support for this, though the way it works is a bit unexpected for some.

Why is this a hard problem?

To understand why mutable variables cause complexities in SSA construction, consider this extremely simple C example:

```
int G, H;
int test(_Bool Condition) {
    int X;
    if (Condition)
        X = G;
    else
        X = H;
    return X;
}
```

In this case, we have the variable “X”, whose value depends on the path executed in the program. Because there are two different possible values for X before the return instruction, a PHI node is inserted to merge the two values. The LLVM IR that we want for this example looks like this:

```
@G = weak global i32 0 ; type of @G is i32*
@H = weak global i32 0 ; type of @H is i32*

define i32 @test(i1 %Condition) {
entry:
    br i1 %Condition, label %cond_true, label %cond_false

cond_true:
    %X.0 = load i32* @G
    br label %cond_next

cond_false:
    %X.1 = load i32* @H
    br label %cond_next

cond_next:
    %X.2 = phi i32 [ %X.1, %cond_false ], [ %X.0, %cond_true ]
    ret i32 %X.2
}
```

In this example, the loads from the G and H global variables are explicit in the LLVM IR, and they live in the then/else branches of the if statement (cond_true/cond_false). In order to merge the incoming values, the X.2 phi node in the cond_next block selects the right value to use based on where control flow is coming from: if control flow comes from the cond_false block, X.2 gets the value of X.1. Alternatively, if control flow comes from cond_true, it gets the value of X.0. The intent of this chapter is not to explain the details of SSA form. For more information, see one of the many [online references](#).

The question for this article is “who places the phi nodes when lowering assignments to mutable variables?”. The issue here is that LLVM *requires* that its IR be in SSA form: there is no “non-ssa” mode for it. However, SSA construction

requires non-trivial algorithms and data structures, so it is inconvenient and wasteful for every front-end to have to reproduce this logic.

Memory in LLVM

The ‘trick’ here is that while LLVM does require all register values to be in SSA form, it does not require (or permit) memory objects to be in SSA form. In the example above, note that the loads from G and H are direct accesses to G and H: they are not renamed or versioned. This differs from some other compiler systems, which do try to version memory objects. In LLVM, instead of encoding dataflow analysis of memory into the LLVM IR, it is handled with [Analysis Passes](#) which are computed on demand.

With this in mind, the high-level idea is that we want to make a stack variable (which lives in memory, because it is on the stack) for each mutable object in a function. To take advantage of this trick, we need to talk about how LLVM represents stack variables.

In LLVM, all memory accesses are explicit with load/store instructions, and it is carefully designed not to have (or need) an “address-of” operator. Notice how the type of the @G/@H global variables is actually “i32*” even though the variable is defined as “i32”. What this means is that @G defines *space* for an i32 in the global data area, but its *name* actually refers to the address for that space. Stack variables work the same way, except that instead of being declared with global variable definitions, they are declared with the [LLVM alloca instruction](#):

```
define i32 @example() {
entry:
  %X = alloca i32           ; type of %X is i32*.
  ...
  %tmp = load i32* %X       ; load the stack value %X from the stack.
  %tmp2 = add i32 %tmp, 1    ; increment it
  store i32 %tmp2, i32* %X  ; store it back
  ...
}
```

This code shows an example of how you can declare and manipulate a stack variable in the LLVM IR. Stack memory allocated with the alloca instruction is fully general: you can pass the address of the stack slot to functions, you can store it in other variables, etc. In our example above, we could rewrite the example to use the alloca technique to avoid using a PHI node:

```
@G = weak global i32 0      ; type of @G is i32*
@H = weak global i32 0      ; type of @H is i32*

define i32 @test(i1 %Condition) {
entry:
  %X = alloca i32           ; type of %X is i32*.
  br i1 %Condition, label %cond_true, label %cond_false

cond_true:
  %X.0 = load i32* @G
  store i32 %X.0, i32* %X    ; Update X
  br label %cond_next

cond_false:
  %X.1 = load i32* @H
  store i32 %X.1, i32* %X    ; Update X
  br label %cond_next

cond_next:
  %X.2 = load i32* %X        ; Read X
  ret i32 %X.2
}
```

With this, we have discovered a way to handle arbitrary mutable variables without the need to create Phi nodes at all:

1. Each mutable variable becomes a stack allocation.
2. Each read of the variable becomes a load from the stack.
3. Each update of the variable becomes a store to the stack.
4. Taking the address of a variable just uses the stack address directly.

While this solution has solved our immediate problem, it introduced another one: we have now apparently introduced a lot of stack traffic for very simple and common operations, a major performance problem. Fortunately for us, the LLVM optimizer has a highly-tuned optimization pass named “mem2reg” that handles this case, promoting allocas like this into SSA registers, inserting Phi nodes as appropriate. If you run this example through the pass, for example, you’ll get:

```
$ llvm-as < example.ll | opt -mem2reg | llvm-dis
@G = weak global i32 0
@H = weak global i32 0

define i32 @test(i1 %Condition) {
entry:
    br i1 %Condition, label %cond_true, label %cond_false

cond_true:
    %X.0 = load i32* @G
    br label %cond_next

cond_false:
    %X.1 = load i32* @H
    br label %cond_next

cond_next:
    %X.01 = phi i32 [ %X.1, %cond_false ], [ %X.0, %cond_true ]
    ret i32 %X.01
}
```

The mem2reg pass implements the standard “iterated dominance frontier” algorithm for constructing SSA form and has a number of optimizations that speed up (very common) degenerate cases. The mem2reg optimization pass is the answer to dealing with mutable variables, and we highly recommend that you depend on it. Note that mem2reg only works on variables in certain circumstances:

1. mem2reg is alloca-driven: it looks for allocas and if it can handle them, it promotes them. It does not apply to global variables or heap allocations.
2. mem2reg only looks for alloca instructions in the entry block of the function. Being in the entry block guarantees that the alloca is only executed once, which makes analysis simpler.
3. mem2reg only promotes allocas whose uses are direct loads and stores. If the address of the stack object is passed to a function, or if any funny pointer arithmetic is involved, the alloca will not be promoted.
4. mem2reg only works on allocas of [first class](#) values (such as pointers, scalars and vectors), and only if the array size of the allocation is 1 (or missing in the .ll file). mem2reg is not capable of promoting structs or arrays to registers. Note that the “sroa” pass is more powerful and can promote structs, “unions”, and arrays in many cases.

All of these properties are easy to satisfy for most imperative languages, and we’ll illustrate it below with Kaleidoscope. The final question you may be asking is: should I bother with this nonsense for my front-end? Wouldn’t it be better if I just did SSA construction directly, avoiding use of the mem2reg optimization pass? In short, we strongly

recommend that you use this technique for building SSA form, unless there is an extremely good reason not to. Using this technique is:

- Proven and well tested: clang uses this technique for local mutable variables. As such, the most common clients of LLVM are using this to handle a bulk of their variables. You can be sure that bugs are found fast and fixed early.
- Extremely Fast: mem2reg has a number of special cases that make it fast in common cases as well as fully general. For example, it has fast-paths for variables that are only used in a single block, variables that only have one assignment point, good heuristics to avoid insertion of unneeded phi nodes, etc.
- Needed for debug info generation: [Debug information in LLVM](#) relies on having the address of the variable exposed so that debug info can be attached to it. This technique dovetails very naturally with this style of debug info.

If nothing else, this makes it much easier to get your front-end up and running, and is very simple to implement. Let's extend Kaleidoscope with mutable variables now!

Mutable Variables in Kaleidoscope

Now that we know the sort of problem we want to tackle, let's see what this looks like in the context of our little Kaleidoscope language. We're going to add two features:

1. The ability to mutate variables with the '=' operator.
2. The ability to define new variables.

While the first item is really what this is about, we only have variables for incoming arguments as well as for induction variables, and redefining those only goes so far :). Also, the ability to define new variables is a useful thing regardless of whether you will be mutating them. Here's a motivating example that shows how we could use these:

```
# Define ':' for sequencing: as a low-precedence operator that ignores operands
# and just returns the RHS.
def binary : 1 (x y) y;

# Recursive fib, we could do this before.
def fib(x)
  if (x < 3) then
    1
  else
    fib(x-1)+fib(x-2);

# Iterative fib.
def fibi(x)
  var a = 1, b = 1, c in
  (for i = 3, i < x in
    c = a + b :
    a = b :
    b = c) :
  b;

# Call it.
fibi(10);
```

In order to mutate variables, we have to change our existing variables to use the “alloca trick”. Once we have that, we'll add our new operator, then extend Kaleidoscope to support new variable definitions.

Adjusting Existing Variables for Mutation

The symbol table in Kaleidoscope is managed at code generation time by the ‘NamedValues’ map. This map currently keeps track of the LLVM “Value*” that holds the double value for the named variable. In order to support mutation, we need to change this slightly, so that NamedValues holds the *memory location* of the variable in question. Note that this change is a refactoring: it changes the structure of the code, but does not (by itself) change the behavior of the compiler. All of these changes are isolated in the Kaleidoscope code generator.

At this point in Kaleidoscope’s development, it only supports variables for two things: incoming arguments to functions and the induction variable of ‘for’ loops. For consistency, we’ll allow mutation of these variables in addition to other user-defined variables. This means that these will both need memory locations.

To start our transformation of Kaleidoscope, we’ll change the NamedValues map so that it maps to AllocaInst* instead of Value*. Once we do this, the C++ compiler will tell us what parts of the code we need to update:

```
static std::map<std::string, AllocaInst*> NamedValues;
```

Also, since we will need to create these allocas, we’ll use a helper function that ensures that the allocas are created in the entry block of the function:

```
/// CreateEntryBlockAlloca - Create an alloca instruction in the entry block of
/// the function. This is used for mutable variables etc.
static AllocaInst *CreateEntryBlockAlloca(Function *TheFunction,
                                          const std::string &VarName) {
    IRBuilder<> TmpB(&TheFunction->getEntryBlock(),
                  TheFunction->getEntryBlock().begin());
    return TmpB.CreateAlloca(Type::getDoubleTy(TheContext), 0,
                           VarName.c_str());
}
```

This funny looking code creates an IRBuilder object that is pointing at the first instruction (.begin()) of the entry block. It then creates an alloca with the expected name and returns it. Because all values in Kaleidoscope are doubles, there is no need to pass in a type to use.

With this in place, the first functionality change we want to make belongs to variable references. In our new scheme, variables live on the stack, so code generating a reference to them actually needs to produce a load from the stack slot:

```
Value *VariableExprAST::codegen() {
    // Look this variable up in the function.
    Value *V = NamedValues[Name];
    if (!V)
        return LogErrorV("Unknown variable name");

    // Load the value.
    return Builder.CreateLoad(V, Name.c_str());
}
```

As you can see, this is pretty straightforward. Now we need to update the things that define the variables to set up the alloca. We’ll start with ForExprAST::codegen() (see the [full code listing](#) for the unabridged code):

```
Function *TheFunction = Builder.GetInsertBlock()->getParent();

// Create an alloca for the variable in the entry block.
AllocaInst *Alloca = CreateEntryBlockAlloca(TheFunction, VarName);

// Emit the start code first, without 'variable' in scope.
Value *StartVal = Start->codegen();
if (!StartVal)
```

```

    return nullptr;

    // Store the value into the alloca.
    Builder.CreateStore(StartVal, Alloca);
    ...

    // Compute the end condition.
    Value *EndCond = End->codegen();
    if (!EndCond)
        return nullptr;

    // Reload, increment, and restore the alloca. This handles the case where
    // the body of the loop mutates the variable.
    Value *CurVar = Builder.CreateLoad(Alloca);
    Value *NextVar = Builder.CreateFAdd(CurVar, StepVal, "nextvar");
    Builder.CreateStore(NextVar, Alloca);
    ...

```

This code is virtually identical to the code before we allowed mutable variables. The big difference is that we no longer have to construct a PHI node, and we use load/store to access the variable as needed.

To support mutable argument variables, we need to also make allocas for them. The code for this is also pretty simple:

```

Function *FunctionAST::codegen() {
    ...
    Builder.SetInsertPoint(BB);

    // Record the function arguments in the NamedValues map.
    NamedValues.clear();
    for (auto &Arg : TheFunction->args()) {
        // Create an alloca for this variable.
        AllocaInst *Alloca = CreateEntryBlockAlloca(TheFunction, Arg.getName());

        // Store the initial value into the alloca.
        Builder.CreateStore(&Arg, Alloca);

        // Add arguments to variable symbol table.
        NamedValues[Arg.getName()] = Alloca;
    }

    if (Value *RetVal = Body->codegen()) {
        ...
    }
}

```

For each argument, we make an alloca, store the input value to the function into the alloca, and register the alloca as the memory location for the argument. This method gets invoked by `FunctionAST::codegen()` right after it sets up the entry block for the function.

The final missing piece is adding the `mem2reg` pass, which allows us to get good codegen once again:

```

// Promote allocas to registers.
TheFPM->add(createPromoteMemoryToRegisterPass());
// Do simple "peephole" optimizations and bit-twiddling optzns.
TheFPM->add(createInstructionCombiningPass());
// Reassociate expressions.
TheFPM->add(createReassociatePass());
...

```

It is interesting to see what the code looks like before and after the `mem2reg` optimization runs. For example, this is

the before/after code for our recursive fib function. Before the optimization:

```
define double @fib(double %x) {
entry:
    %x1 = alloca double
    store double %x, double* %x1
    %x2 = load double, double* %x1
    %cmptmp = fcmp ult double %x2, 3.000000e+00
    %booltmp = uitofp i1 %cmptmp to double
    %ifcond = fcmp one double %booltmp, 0.000000e+00
    br i1 %ifcond, label %then, label %else

then:    ; preds = %entry
    br label %ifcont

else:    ; preds = %entry
    %x3 = load double, double* %x1
    %subtmp = fsub double %x3, 1.000000e+00
    %calltmp = call double @fib(double %subtmp)
    %x4 = load double, double* %x1
    %subtmp5 = fsub double %x4, 2.000000e+00
    %calltmp6 = call double @fib(double %subtmp5)
    %addtmp = fadd double %calltmp, %calltmp6
    br label %ifcont

ifcont:  ; preds = %else, %then
    %iftmp = phi double [ 1.000000e+00, %then ], [ %addtmp, %else ]
    ret double %iftmp
}
```

Here there is only one variable (x, the input argument) but you can still see the extremely simple-minded code generation strategy we are using. In the entry block, an alloca is created, and the initial input value is stored into it. Each reference to the variable does a reload from the stack. Also, note that we didn't modify the if/then/else expression, so it still inserts a PHI node. While we could make an alloca for it, it is actually easier to create a PHI node for it, so we still just make the PHI.

Here is the code after the mem2reg pass runs:

```
define double @fib(double %x) {
entry:
    %cmptmp = fcmp ult double %x, 3.000000e+00
    %booltmp = uitofp i1 %cmptmp to double
    %ifcond = fcmp one double %booltmp, 0.000000e+00
    br i1 %ifcond, label %then, label %else

then:
    br label %ifcont

else:
    %subtmp = fsub double %x, 1.000000e+00
    %calltmp = call double @fib(double %subtmp)
    %subtmp5 = fsub double %x, 2.000000e+00
    %calltmp6 = call double @fib(double %subtmp5)
    %addtmp = fadd double %calltmp, %calltmp6
    br label %ifcont

ifcont:  ; preds = %else, %then
    %iftmp = phi double [ 1.000000e+00, %then ], [ %addtmp, %else ]
    ret double %iftmp
}
```



```
}

```

This is a trivial case for mem2reg, since there are no redefinitions of the variable. The point of showing this is to calm your tension about inserting such blatant inefficiencies :).

After the rest of the optimizers run, we get:

```
define double @fib(double %x) {
entry:
  %cmptmp = fcmp ult double %x, 3.000000e+00
  %booltmp = uitofp i1 %cmptmp to double
  %ifcond = fcmp ueq double %booltmp, 0.000000e+00
  br i1 %ifcond, label %else, label %ifcont

else:
  %subtmp = fsub double %x, 1.000000e+00
  %calltmp = call double @fib(double %subtmp)
  %subtmp5 = fsub double %x, 2.000000e+00
  %calltmp6 = call double @fib(double %subtmp5)
  %addtmp = fadd double %calltmp, %calltmp6
  ret double %addtmp

ifcont:
  ret double 1.000000e+00
}
```

Here we see that the simplifcfg pass decided to clone the return instruction into the end of the ‘else’ block. This allowed it to eliminate some branches and the PHI node.

Now that all symbol table references are updated to use stack variables, we’ll add the assignment operator.

New Assignment Operator

With our current framework, adding a new assignment operator is really simple. We will parse it just like any other binary operator, but handle it internally (instead of allowing the user to define it). The first step is to set a precedence:

```
int main() {
  // Install standard binary operators.
  // 1 is lowest precedence.
  BinopPrecedence['='] = 2;
  BinopPrecedence['<'] = 10;
  BinopPrecedence['+'] = 20;
  BinopPrecedence['-'] = 20;
}
```

Now that the parser knows the precedence of the binary operator, it takes care of all the parsing and AST generation. We just need to implement codegen for the assignment operator. This looks like:

```
Value *BinaryExprAST::codegen() {
  // Special case '=' because we don't want to emit the LHS as an expression.
  if (Op == '=') {
    // Assignment requires the LHS to be an identifier.
    VariableExprAST *LHSE = dynamic_cast<VariableExprAST*>(LHS.get());
    if (!LHSE)
      return LogErrorV("destination of '=' must be a variable");
  }
}
```

Unlike the rest of the binary operators, our assignment operator doesn’t follow the “emit LHS, emit RHS, do computation” model. As such, it is handled as a special case before the other binary operators are handled. The other strange

thing is that it requires the LHS to be a variable. It is invalid to have “(x+1) = expr” - only things like “x = expr” are allowed.

```
// Codegen the RHS.
Value *Val = RHS->codegen();
if (!Val)
    return nullptr;

// Look up the name.
Value *Variable = NamedValues[LHSE->getName()];
if (!Variable)
    return LogErrorV("Unknown variable name");

Builder.CreateStore(Val, Variable);
return Val;
}
...
```

Once we have the variable, codegen’ing the assignment is straightforward: we emit the RHS of the assignment, create a store, and return the computed value. Returning a value allows for chained assignments like “X = (Y = Z)”.

Now that we have an assignment operator, we can mutate loop variables and arguments. For example, we can now run code like this:

```
# Function to print a double.
extern printf(x);

# Define ':' for sequencing: as a low-precedence operator that ignores operands
# and just returns the RHS.
def binary : 1 (x y) y;

def test(x)
  printf(x) :
  x = 4 :
  printf(x);

test(123);
```

When run, this example prints “123” and then “4”, showing that we did actually mutate the value! Okay, we have now officially implemented our goal: getting this to work requires SSA construction in the general case. However, to be really useful, we want the ability to define our own local variables, let’s add this next!

User-defined Local Variables

Adding var/in is just like any other extension we made to Kaleidoscope: we extend the lexer, the parser, the AST and the code generator. The first step for adding our new ‘var/in’ construct is to extend the lexer. As before, this is pretty trivial, the code looks like this:

```
enum Token {
    ...
    // var definition
    tok_var = -13
    ...
}
...
static int gettok() {
    ...
}
```

```

    if (IdentifierStr == "in")
        return tok_in;
    if (IdentifierStr == "binary")
        return tok_binary;
    if (IdentifierStr == "unary")
        return tok_unary;
    if (IdentifierStr == "var")
        return tok_var;
    return tok_identifier;
...

```

The next step is to define the AST node that we will construct. For var/in, it looks like this:

```

/// VarExprAST - Expression class for var/in
class VarExprAST : public ExprAST {
    std::vector<std::pair<std::string, std::unique_ptr<ExprAST>>> VarNames;
    std::unique_ptr<ExprAST> Body;

public:
    VarExprAST(std::vector<std::pair<std::string, std::unique_ptr<ExprAST>>> VarNames,
               std::unique_ptr<ExprAST> Body)
        : VarNames(std::move(VarNames)), Body(std::move(Body)) {}

    Value *codegen() override;
};

```

var/in allows a list of names to be defined all at once, and each name can optionally have an initializer value. As such, we capture this information in the VarNames vector. Also, var/in has a body, this body is allowed to access the variables defined by the var/in.

With this in place, we can define the parser pieces. The first thing we do is add it as a primary expression:

```

/// primary
/// ::= identifierexpr
/// ::= numberexpr
/// ::= parenexpr
/// ::= ifexpr
/// ::= forexpr
/// ::= varexpr
static std::unique_ptr<ExprAST> ParsePrimary() {
    switch (CurTok) {
    default:
        return LogError("unknown token when expecting an expression");
    case tok_identifier:
        return ParseIdentifierExpr();
    case tok_number:
        return ParseNumberExpr();
    case '(':
        return ParseParenExpr();
    case tok_if:
        return ParseIfExpr();
    case tok_for:
        return ParseForExpr();
    case tok_var:
        return ParseVarExpr();
    }
}

```

Next we define ParseVarExpr:

```

/// varexpr ::= 'var' identifier ('=' expression)?
///           (',' identifier ('=' expression))* 'in' expression
static std::unique_ptr<ExprAST> ParseVarExpr() {
    getNextToken(); // eat the var.

    std::vector<std::pair<std::string, std::unique_ptr<ExprAST>>> VarNames;

    // At least one variable name is required.
    if (CurTok != tok_identifier)
        return LogError("expected identifier after var");

```

The first part of this code parses the list of identifier/expr pairs into the local VarNames vector.

```

while (1) {
    std::string Name = IdentifierStr;
    getNextToken(); // eat identifier.

    // Read the optional initializer.
    std::unique_ptr<ExprAST> Init;
    if (CurTok == '=') {
        getNextToken(); // eat the '='.

        Init = ParseExpression();
        if (!Init) return nullptr;
    }

    VarNames.push_back(std::make_pair(Name, std::move(Init)));

    // End of var list, exit loop.
    if (CurTok != ',') break;
    getNextToken(); // eat the ','.

    if (CurTok != tok_identifier)
        return LogError("expected identifier list after var");
}

```

Once all the variables are parsed, we then parse the body and create the AST node:

```

// At this point, we have to have 'in'.
if (CurTok != tok_in)
    return LogError("expected 'in' keyword after 'var'");
getNextToken(); // eat 'in'.

auto Body = ParseExpression();
if (!Body)
    return nullptr;

return llvm::make_unique<VarExprAST>(std::move(VarNames),
                                     std::move(Body));
}

```

Now that we can parse and represent the code, we need to support emission of LLVM IR for it. This code starts out with:

```

Value *VarExprAST::codegen() {
    std::vector<AllocaInst *> OldBindings;

```

```
Function *TheFunction = Builder.GetInsertBlock()->getParent();

// Register all variables and emit their initializer.
for (unsigned i = 0, e = VarNames.size(); i != e; ++i) {
    const std::string &VarName = VarNames[i].first;
    ExprAST *Init = VarNames[i].second.get();
```

Basically it loops over all the variables, installing them one at a time. For each variable we put into the symbol table, we remember the previous value that we replace in OldBindings.

```
// Emit the initializer before adding the variable to scope, this prevents
// the initializer from referencing the variable itself, and permits stuff
// like this:
//   var a = 1 in
//   var a = a in ... # refers to outer 'a'.
Value *InitVal;
if (Init) {
    InitVal = Init->codegen();
    if (!InitVal)
        return nullptr;
} else { // If not specified, use 0.0.
    InitVal = ConstantFP::get(TheContext, APFloat(0.0));
}

AllocaInst *Alloca = CreateEntryBlockAlloca(TheFunction, VarName);
Builder.CreateStore(InitVal, Alloca);

// Remember the old variable binding so that we can restore the binding when
// we unrecurse.
OldBindings.push_back(NamedValues[VarName]);

// Remember this binding.
NamedValues[VarName] = Alloca;
}
```

There are more comments here than code. The basic idea is that we emit the initializer, create the alloca, then update the symbol table to point to it. Once all the variables are installed in the symbol table, we evaluate the body of the var/in expression:

```
// Codegen the body, now that all vars are in scope.
Value *BodyVal = Body->codegen();
if (!BodyVal)
    return nullptr;
```

Finally, before returning, we restore the previous variable bindings:

```
// Pop all our variables from scope.
for (unsigned i = 0, e = VarNames.size(); i != e; ++i)
    NamedValues[VarNames[i].first] = OldBindings[i];

// Return the body computation.
return BodyVal;
}
```

The end result of all of this is that we get properly scoped variable definitions, and we even (trivially) allow mutation of them :).

With this, we completed what we set out to do. Our nice iterative fib example from the intro compiles and runs just

fine. The mem2reg pass optimizes all of our stack variables into SSA registers, inserting PHI nodes where needed, and our front-end remains simple: no “iterated dominance frontier” computation anywhere in sight.

Full Code Listing

Here is the complete code listing for our running example, enhanced with mutable variables and var/in support. To build this example, use:

```
# Compile
clang++ -g toy.cpp `llvm-config --cxxflags --ldflags --system-libs --libs core mcjit_
↪native` -O3 -o toy
# Run
./toy
```

Here is the code:

```
#include "llvm/ADT/APFloat.h"
#include "llvm/ADT/STLExtras.h"
#include "llvm/IR/BasicBlock.h"
#include "llvm/IR/Constants.h"
#include "llvm/IR/DerivedTypes.h"
#include "llvm/IR/Function.h"
#include "llvm/IR/Instructions.h"
#include "llvm/IR/IRBuilder.h"
#include "llvm/IR/LLVMContext.h"
#include "llvm/IR/LegacyPassManager.h"
#include "llvm/IR/Module.h"
#include "llvm/IR/Type.h"
#include "llvm/IR/Verifier.h"
#include "llvm/Support/TargetSelect.h"
#include "llvm/Target/TargetMachine.h"
#include "llvm/Transforms/Scalar.h"
#include "llvm/Transforms/Scalar/GVN.h"
#include "../include/KaleidoscopeJIT.h"
#include <algorithm>
#include <cassert>
#include <cctype>
#include <cstdint>
#include <cstdio>
#include <cstdlib>
#include <map>
#include <memory>
#include <string>
#include <utility>
#include <vector>

using namespace llvm;
using namespace llvm::orc;

//=====//
// Lexer
//=====//

// The lexer returns tokens [0-255] if it is an unknown character, otherwise one
// of these for known things.
enum Token {
    tok_eof = -1,
```

```

// commands
tok_def = -2,
tok_extern = -3,

// primary
tok_identifier = -4,
tok_number = -5,

// control
tok_if = -6,
tok_then = -7,
tok_else = -8,
tok_for = -9,
tok_in = -10,

// operators
tok_binary = -11,
tok_unary = -12,

// var definition
tok_var = -13
};

static std::string IdentifierStr; // Filled in if tok_identifier
static double NumVal;           // Filled in if tok_number

///gettok - Return the next token from standard input.
static int gettok() {
    static int LastChar = ' ';

    // Skip any whitespace.
    while (isspace(LastChar))
        LastChar = getchar();

    if (isalpha(LastChar)) { // identifier: [a-zA-Z][a-zA-Z0-9]*
        IdentifierStr = LastChar;
        while (isalnum((LastChar = getchar())))
            IdentifierStr += LastChar;

        if (IdentifierStr == "def")
            return tok_def;
        if (IdentifierStr == "extern")
            return tok_extern;
        if (IdentifierStr == "if")
            return tok_if;
        if (IdentifierStr == "then")
            return tok_then;
        if (IdentifierStr == "else")
            return tok_else;
        if (IdentifierStr == "for")
            return tok_for;
        if (IdentifierStr == "in")
            return tok_in;
        if (IdentifierStr == "binary")
            return tok_binary;
        if (IdentifierStr == "unary")
            return tok_unary;
    }

```

```

    if (IdentifierStr == "var")
        return tok_var;
    return tok_identifier;
}

if (isdigit>LastChar) || LastChar == '.') { // Number: [0-9.]+
    std::string NumStr;
    do {
        NumStr += LastChar;
        LastChar = getchar();
    } while (isdigit>LastChar) || LastChar == '.');

    NumVal = strtod(NumStr.c_str(), nullptr);
    return tok_number;
}

if (LastChar == '#') {
    // Comment until end of line.
    do
        LastChar = getchar();
    while (LastChar != EOF && LastChar != '\n' && LastChar != '\r');

    if (LastChar != EOF)
        return gettok();
}

// Check for end of file. Don't eat the EOF.
if (LastChar == EOF)
    return tok_eof;

// Otherwise, just return the character as its ascii value.
int ThisChar = LastChar;
LastChar = getchar();
return ThisChar;
}

//=====
// Abstract Syntax Tree (aka Parse Tree)
//=====

namespace {

/// ExprAST - Base class for all expression nodes.
class ExprAST {
public:
    virtual ~ExprAST() = default;

    virtual Value *codegen() = 0;
};

/// NumberExprAST - Expression class for numeric literals like "1.0".
class NumberExprAST : public ExprAST {
    double Val;

public:
    NumberExprAST(double Val) : Val(Val) {}

    Value *codegen() override;
}

```



```

};

/// VariableExprAST - Expression class for referencing a variable, like "a".
class VariableExprAST : public ExprAST {
    std::string Name;

public:
    VariableExprAST(const std::string &Name) : Name(Name) {}

    Value *codegen() override;
    const std::string &getName() const { return Name; }
};

/// UnaryExprAST - Expression class for a unary operator.
class UnaryExprAST : public ExprAST {
    char Opcode;
    std::unique_ptr<ExprAST> Operand;

public:
    UnaryExprAST(char Opcode, std::unique_ptr<ExprAST> Operand)
        : Opcode(Opcode), Operand(std::move(Operand)) {}

    Value *codegen() override;
};

/// BinaryExprAST - Expression class for a binary operator.
class BinaryExprAST : public ExprAST {
    char Op;
    std::unique_ptr<ExprAST> LHS, RHS;

public:
    BinaryExprAST(char Op, std::unique_ptr<ExprAST> LHS,
                  std::unique_ptr<ExprAST> RHS)
        : Op(Op), LHS(std::move(LHS)), RHS(std::move(RHS)) {}

    Value *codegen() override;
};

/// CallExprAST - Expression class for function calls.
class CallExprAST : public ExprAST {
    std::string Callee;
    std::vector<std::unique_ptr<ExprAST>> Args;

public:
    CallExprAST(const std::string &Callee,
                std::vector<std::unique_ptr<ExprAST>> Args)
        : Callee(Callee), Args(std::move(Args)) {}

    Value *codegen() override;
};

/// IfExprAST - Expression class for if/then/else.
class IfExprAST : public ExprAST {
    std::unique_ptr<ExprAST> Cond, Then, Else;

public:
    IfExprAST(std::unique_ptr<ExprAST> Cond, std::unique_ptr<ExprAST> Then,
              std::unique_ptr<ExprAST> Else)

```

```

        : Cond(std::move(Cond)), Then(std::move(Then)), Else(std::move(Else)) {}

Value *codegen() override;
};

/// ForExprAST - Expression class for for/in.
class ForExprAST : public ExprAST {
    std::string VarName;
    std::unique_ptr<ExprAST> Start, End, Step, Body;

public:
    ForExprAST(const std::string &VarName, std::unique_ptr<ExprAST> Start,
               std::unique_ptr<ExprAST> End, std::unique_ptr<ExprAST> Step,
               std::unique_ptr<ExprAST> Body)
        : VarName(VarName), Start(std::move(Start)), End(std::move(End)),
          Step(std::move(Step)), Body(std::move(Body)) {}

Value *codegen() override;
};

/// VarExprAST - Expression class for var/in
class VarExprAST : public ExprAST {
    std::vector<std::pair<std::string, std::unique_ptr<ExprAST>>> VarNames;
    std::unique_ptr<ExprAST> Body;

public:
    VarExprAST(
        std::vector<std::pair<std::string, std::unique_ptr<ExprAST>>> VarNames,
        std::unique_ptr<ExprAST> Body)
        : VarNames(std::move(VarNames)), Body(std::move(Body)) {}

Value *codegen() override;
};

/// PrototypeAST - This class represents the "prototype" for a function,
/// which captures its name, and its argument names (thus implicitly the number
/// of arguments the function takes), as well as if it is an operator.
class PrototypeAST {
    std::string Name;
    std::vector<std::string> Args;
    bool IsOperator;
    unsigned Precedence; // Precedence if a binary op.

public:
    PrototypeAST(const std::string &Name, std::vector<std::string> Args,
                 bool IsOperator = false, unsigned Prec = 0)
        : Name(Name), Args(std::move(Args)), IsOperator(IsOperator),
          Precedence(Prec) {}

Function *codegen();
const std::string &getName() const { return Name; }

bool isUnaryOp() const { return IsOperator && Args.size() == 1; }
bool isBinaryOp() const { return IsOperator && Args.size() == 2; }

char getOperatorName() const {
    assert(isUnaryOp() || isBinaryOp());
    return Name[Name.size() - 1];
}

```

```

    }

    unsigned getBinaryPrecedence() const { return Precedence; }
};

/// FunctionAST - This class represents a function definition itself.
class FunctionAST {
    std::unique_ptr<PrototypeAST> Proto;
    std::unique_ptr<ExprAST> Body;

public:
    FunctionAST(std::unique_ptr<PrototypeAST> Proto,
                std::unique_ptr<ExprAST> Body)
        : Proto(std::move(Proto)), Body(std::move(Body)) {}

    Function *codegen();
};

} // end anonymous namespace

//===-----
// Parser
//===-----

/// CurTok/getNextToken - Provide a simple token buffer. CurTok is the current
/// token the parser is looking at. getNextToken reads another token from the
/// lexer and updates CurTok with its results.
static int CurTok;
static int getNextToken() { return CurTok = gettok(); }

/// BinopPrecedence - This holds the precedence for each binary operator that is
/// defined.
static std::map<char, int> BinopPrecedence;

/// GetTokPrecedence - Get the precedence of the pending binary operator token.
static int GetTokPrecedence() {
    if (!isascii(CurTok))
        return -1;

    // Make sure it's a declared binop.
    int TokPrec = BinopPrecedence[CurTok];
    if (TokPrec <= 0)
        return -1;
    return TokPrec;
}

/// LogError* - These are little helper functions for error handling.
std::unique_ptr<ExprAST> LogError(const char *Str) {
    fprintf(stderr, "Error: %s\n", Str);
    return nullptr;
}

std::unique_ptr<PrototypeAST> LogErrorP(const char *Str) {
    LogError(Str);
    return nullptr;
}

static std::unique_ptr<ExprAST> ParseExpression();

```

```

/// numberexpr ::= number
static std::unique_ptr<ExprAST> ParseNumberExpr() {
    auto Result = llvm::make_unique<NumberExprAST>(NumVal);
    getNextToken(); // consume the number
    return std::move(Result);
}

/// parenexpr ::= '(' expression ')'
static std::unique_ptr<ExprAST> ParseParenExpr() {
    getNextToken(); // eat '('
    auto V = ParseExpression();
    if (!V)
        return nullptr;

    if (CurTok != ')')
        return LogError("expected ')'");
    getNextToken(); // eat ')'
    return V;
}

/// identifierexpr
/// ::= identifier
/// ::= identifier '(' expression* ')'
static std::unique_ptr<ExprAST> ParseIdentifierExpr() {
    std::string IdName = IdentifierStr;

    getNextToken(); // eat identifier.

    if (CurTok != '(') // Simple variable ref.
        return llvm::make_unique<VariableExprAST>(IdName);

    // Call.
    getNextToken(); // eat '('
    std::vector<std::unique_ptr<ExprAST>> Args;
    if (CurTok != ')') {
        while (true) {
            if (auto Arg = ParseExpression())
                Args.push_back(std::move(Arg));
            else
                return nullptr;

            if (CurTok == ')')
                break;

            if (CurTok != ',')
                return LogError("Expected ')' or ',' in argument list");
            getNextToken();
        }
    }

    // Eat the ')'
    getNextToken();

    return llvm::make_unique<CallExprAST>(IdName, std::move(Args));
}

/// ifexpr ::= 'if' expression 'then' expression 'else' expression

```

```

static std::unique_ptr<ExprAST> ParseIfExpr() {
    getNextToken(); // eat the if.

    // condition.
    auto Cond = ParseExpression();
    if (!Cond)
        return nullptr;

    if (CurTok != tok_then)
        return LogError("expected then");
    getNextToken(); // eat the then

    auto Then = ParseExpression();
    if (!Then)
        return nullptr;

    if (CurTok != tok_else)
        return LogError("expected else");

    getNextToken();

    auto Else = ParseExpression();
    if (!Else)
        return nullptr;

    return llvm::make_unique<IfExprAST>(std::move(Cond), std::move(Then),
                                       std::move(Else));
}

// forexpr ::= 'for' identifier '=' expr ',' expr (',' expr)? 'in' expression
static std::unique_ptr<ExprAST> ParseForExpr() {
    getNextToken(); // eat the for.

    if (CurTok != tok_identifier)
        return LogError("expected identifier after for");

    std::string IdName = IdentifierStr;
    getNextToken(); // eat identifier.

    if (CurTok != '=')
        return LogError("expected '=' after for");
    getNextToken(); // eat '='.

    auto Start = ParseExpression();
    if (!Start)
        return nullptr;
    if (CurTok != ',')
        return LogError("expected ',' after for start value");
    getNextToken();

    auto End = ParseExpression();
    if (!End)
        return nullptr;

    // The step value is optional.
    std::unique_ptr<ExprAST> Step;
    if (CurTok == ',') {
        getNextToken();

```

```

    Step = ParseExpression();
    if (!Step)
        return nullptr;
}

if (CurTok != tok_in)
    return LogError("expected 'in' after for");
getNextToken(); // eat 'in'.

auto Body = ParseExpression();
if (!Body)
    return nullptr;

return llvm::make_unique<ForExprAST>(IdName, std::move(Start), std::move(End),
                                     std::move(Step), std::move(Body));
}

/// varexpr ::= 'var' identifier ('=' expression)?
///           (',' identifier ('=' expression)?)* 'in' expression
static std::unique_ptr<ExprAST> ParseVarExpr() {
    getNextToken(); // eat the var.

    std::vector<std::pair<std::string, std::unique_ptr<ExprAST>>> VarNames;

    // At least one variable name is required.
    if (CurTok != tok_identifier)
        return LogError("expected identifier after var");

    while (true) {
        std::string Name = IdentifierStr;
        getNextToken(); // eat identifier.

        // Read the optional initializer.
        std::unique_ptr<ExprAST> Init = nullptr;
        if (CurTok == '=') {
            getNextToken(); // eat the '='.

            Init = ParseExpression();
            if (!Init)
                return nullptr;
        }

        VarNames.push_back(std::make_pair(Name, std::move(Init)));

        // End of var list, exit loop.
        if (CurTok != ',')
            break;
        getNextToken(); // eat the ','.

        if (CurTok != tok_identifier)
            return LogError("expected identifier list after var");
    }

    // At this point, we have to have 'in'.
    if (CurTok != tok_in)
        return LogError("expected 'in' keyword after 'var'");
    getNextToken(); // eat 'in'.

```

```

    auto Body = ParseExpression();
    if (!Body)
        return nullptr;

    return llvm::make_unique<VarExprAST>(std::move(VarNames), std::move(Body));
}

/// primary
/// ::= identifierexpr
/// ::= numberexpr
/// ::= parenexpr
/// ::= ifexpr
/// ::= forexpr
/// ::= varexpr
static std::unique_ptr<ExprAST> ParsePrimary() {
    switch (CurTok) {
    default:
        return LogError("unknown token when expecting an expression");
    case tok_identifier:
        return ParseIdentifierExpr();
    case tok_number:
        return ParseNumberExpr();
    case '(':
        return ParseParenExpr();
    case tok_if:
        return ParseIfExpr();
    case tok_for:
        return ParseForExpr();
    case tok_var:
        return ParseVarExpr();
    }
}

/// unary
/// ::= primary
/// ::= '!' unary
static std::unique_ptr<ExprAST> ParseUnary() {
    // If the current token is not an operator, it must be a primary expr.
    if (!isascii(CurTok) || CurTok == '(' || CurTok == ',')
        return ParsePrimary();

    // If this is a unary operator, read it.
    int Opc = CurTok;
    getNextToken();
    if (auto Operand = ParseUnary())
        return llvm::make_unique<UnaryExprAST>(Opc, std::move(Operand));
    return nullptr;
}

/// binoprhs
/// ::= ('+' unary)*
static std::unique_ptr<ExprAST> ParseBinOpRHS(int ExprPrec,
                                              std::unique_ptr<ExprAST> LHS) {
    // If this is a binop, find its precedence.
    while (true) {
        int TokPrec = GetTokPrecedence();

        // If this is a binop that binds at least as tightly as the current binop,

```

```

    // consume it, otherwise we are done.
    if (TokPrec < ExprPrec)
        return LHS;

    // Okay, we know this is a binop.
    int BinOp = CurTok;
    getNextToken(); // eat binop

    // Parse the unary expression after the binary operator.
    auto RHS = ParseUnary();
    if (!RHS)
        return nullptr;

    // If BinOp binds less tightly with RHS than the operator after RHS, let
    // the pending operator take RHS as its LHS.
    int NextPrec = GetTokPrecedence();
    if (TokPrec < NextPrec) {
        RHS = ParseBinOpRHS(TokPrec + 1, std::move(RHS));
        if (!RHS)
            return nullptr;
    }

    // Merge LHS/RHS.
    LHS =
        llvm::make_unique<BinaryExprAST>(BinOp, std::move(LHS), std::move(RHS));
}

/// expression
/// ::= unary binoprhs
///
static std::unique_ptr<ExprAST> ParseExpression() {
    auto LHS = ParseUnary();
    if (!LHS)
        return nullptr;

    return ParseBinOpRHS(0, std::move(LHS));
}

/// prototype
/// ::= id '(' id* ')'
/// ::= binary LETTER number? (id, id)
/// ::= unary LETTER (id)
static std::unique_ptr<PrototypeAST> ParsePrototype() {
    std::string FnName;

    unsigned Kind = 0; // 0 = identifier, 1 = unary, 2 = binary.
    unsigned BinaryPrecedence = 30;

    switch (CurTok) {
    default:
        return LogErrorP("Expected function name in prototype");
    case tok_identifier:
        FnName = IdentifierStr;
        Kind = 0;
        getNextToken();
        break;
    case tok_unary:

```



```

getNextToken();
if (!isascii(CurTok))
    return LogErrorP("Expected unary operator");
FnName = "unary";
FnName += (char)CurTok;
Kind = 1;
getNextToken();
break;
case tok_binary:
    getNextToken();
    if (!isascii(CurTok))
        return LogErrorP("Expected binary operator");
    FnName = "binary";
    FnName += (char)CurTok;
    Kind = 2;
    getNextToken();

    // Read the precedence if present.
    if (CurTok == tok_number) {
        if (NumVal < 1 || NumVal > 100)
            return LogErrorP("Invalid precedence: must be 1..100");
        BinaryPrecedence = (unsigned)NumVal;
        getNextToken();
    }
    break;
}

if (CurTok != '(')
    return LogErrorP("Expected '(' in prototype");

std::vector<std::string> ArgNames;
while (getNextToken() == tok_identifier)
    ArgNames.push_back(IdentifierStr);
if (CurTok != ')')
    return LogErrorP("Expected ')' in prototype");

// success.
getNextToken(); // eat ')'.

// Verify right number of names for operator.
if (Kind && ArgNames.size() != Kind)
    return LogErrorP("Invalid number of operands for operator");

return llvm::make_unique<PrototypeAST>(FnName, ArgNames, Kind != 0,
                                       BinaryPrecedence);
}

/// definition ::= 'def' prototype expression
static std::unique_ptr<FunctionAST> ParseDefinition() {
    getNextToken(); // eat def.
    auto Proto = ParsePrototype();
    if (!Proto)
        return nullptr;

    if (auto E = ParseExpression())
        return llvm::make_unique<FunctionAST>(std::move(Proto), std::move(E));
    return nullptr;
}

```

```

/// toplevelexpr ::= expression
static std::unique_ptr<FunctionAST> ParseTopLevelExpr() {
    if (auto E = ParseExpression()) {
        // Make an anonymous proto.
        auto Proto = llvm::make_unique<PrototypeAST>("__anon_expr",
                                                    std::vector<std::string>());
        return llvm::make_unique<FunctionAST>(std::move(Proto), std::move(E));
    }
    return nullptr;
}

/// external ::= 'extern' prototype
static std::unique_ptr<PrototypeAST> ParseExtern() {
    getNextToken(); // eat extern.
    return ParsePrototype();
}

//===-----
// Code Generation
//===-----

static LLVMContext TheContext;
static IRBuilder<> Builder(TheContext);
static std::unique_ptr<Module> TheModule;
static std::map<std::string, AllocaInst *> NamedValues;
static std::unique_ptr<legacy::FunctionPassManager> TheFPM;
static std::unique_ptr<KaleidoscopeJIT> TheJIT;
static std::map<std::string, std::unique_ptr<PrototypeAST>> FunctionProtos;

Value *LogErrorV(const char *Str) {
    LogError(Str);
    return nullptr;
}

Function *getFunction(std::string Name) {
    // First, see if the function has already been added to the current module.
    if (auto *F = TheModule->getFunction(Name))
        return F;

    // If not, check whether we can codegen the declaration from some existing
// prototype.
    auto FI = FunctionProtos.find(Name);
    if (FI != FunctionProtos.end())
        return FI->second->codegen();

    // If no existing prototype exists, return null.
    return nullptr;
}

/// CreateEntryBlockAlloca - Create an alloca instruction in the entry block of
/// the function. This is used for mutable variables etc.
static AllocaInst *CreateEntryBlockAlloca(Function *TheFunction,
                                           const std::string &VarName) {
    IRBuilder<> TmpB(&TheFunction->getEntryBlock(),
                  TheFunction->getEntryBlock().begin());
    return TmpB.CreateAlloca(Type::getDoubleTy(TheContext), nullptr, VarName);
}

```

```

Value *NumberExprAST::codegen() {
    return ConstantFP::get(TheContext, APFloat(Val));
}

Value *VariableExprAST::codegen() {
    // Look this variable up in the function.
    Value *V = NamedValues[Name];
    if (!V)
        return LogErrorV("Unknown variable name");

    // Load the value.
    return Builder.CreateLoad(V, Name.c_str());
}

Value *UnaryExprAST::codegen() {
    Value *OperandV = Operand->codegen();
    if (!OperandV)
        return nullptr;

    Function *F = getFunction(std::string("unary") + Opcode);
    if (!F)
        return LogErrorV("Unknown unary operator");

    return Builder.CreateCall(F, OperandV, "unop");
}

Value *BinaryExprAST::codegen() {
    // Special case '=' because we don't want to emit the LHS as an expression.
    if (Op == '=') {
        // Assignment requires the LHS to be an identifier.
        // This assume we're building without RTTI because LLVM builds that way by
        // default. If you build LLVM with RTTI this can be changed to a
        // dynamic_cast for automatic error checking.
        VariableExprAST *LHSE = static_cast<VariableExprAST *>(LHS.get());
        if (!LHSE)
            return LogErrorV("destination of '=' must be a variable");
        // Codegen the RHS.
        Value *Val = RHS->codegen();
        if (!Val)
            return nullptr;

        // Look up the name.
        Value *Variable = NamedValues[LHSE->getName()];
        if (!Variable)
            return LogErrorV("Unknown variable name");

        Builder.CreateStore(Val, Variable);
        return Val;
    }

    Value *L = LHS->codegen();
    Value *R = RHS->codegen();
    if (!L || !R)
        return nullptr;

    switch (Op) {
    case '+':

```

```

    return Builder.CreateFAdd(L, R, "addtmp");
case '-':
    return Builder.CreateFSub(L, R, "subtmp");
case '*':
    return Builder.CreateFMul(L, R, "multmp");
case '<':
    L = Builder.CreateFCmpULT(L, R, "cmptmp");
    // Convert bool 0/1 to double 0.0 or 1.0
    return Builder.CreateUIToFP(L, Type::getDoubleTy(TheContext), "booltmp");
default:
    break;
}

// If it wasn't a builtin binary operator, it must be a user defined one. Emit
// a call to it.
Function *F = getFunction(std::string("binary") + Op);
assert(F && "binary operator not found!");

Value *Ops[] = {L, R};
return Builder.CreateCall(F, Ops, "binop");
}

Value *CallExprAST::codegen() {
    // Look up the name in the global module table.
    Function *CalleeF = getFunction(Callee);
    if (!CalleeF)
        return LogErrorV("Unknown function referenced");

    // If argument mismatch error.
    if (CalleeF->arg_size() != Args.size())
        return LogErrorV("Incorrect # arguments passed");

    std::vector<Value *> ArgsV;
    for (unsigned i = 0, e = Args.size(); i != e; ++i) {
        ArgsV.push_back(Args[i]->codegen());
        if (!ArgsV.back())
            return nullptr;
    }

    return Builder.CreateCall(CalleeF, ArgsV, "calltmp");
}

Value *IfExprAST::codegen() {
    Value *CondV = Cond->codegen();
    if (!CondV)
        return nullptr;

    // Convert condition to a bool by comparing non-equal to 0.0.
    CondV = Builder.CreateFCmpONE(
        CondV, ConstantFP::get(TheContext, APFloat(0.0)), "ifcond");

    Function *TheFunction = Builder.GetInsertBlock()->getParent();

    // Create blocks for the then and else cases. Insert the 'then' block at the
    // end of the function.
    BasicBlock *ThenBB = BasicBlock::Create(TheContext, "then", TheFunction);
    BasicBlock *ElseBB = BasicBlock::Create(TheContext, "else");
    BasicBlock *MergeBB = BasicBlock::Create(TheContext, "ifcont");

```

```

Builder.CreateCondBr(CondV, ThenBB, ElseBB);

// Emit then value.
Builder.SetInsertPoint(ThenBB);

Value *ThenV = Then->codegen();
if (!ThenV)
    return nullptr;

Builder.CreateBr(MergeBB);
// Codegen of 'Then' can change the current block, update ThenBB for the PHI.
ThenBB = Builder.GetInsertBlock();

// Emit else block.
TheFunction->getBasicBlockList().push_back(ElseBB);
Builder.SetInsertPoint(ElseBB);

Value *ElseV = Else->codegen();
if (!ElseV)
    return nullptr;

Builder.CreateBr(MergeBB);
// Codegen of 'Else' can change the current block, update ElseBB for the PHI.
ElseBB = Builder.GetInsertBlock();

// Emit merge block.
TheFunction->getBasicBlockList().push_back(MergeBB);
Builder.SetInsertPoint(MergeBB);
PHINode *PN = Builder.CreatePHI(Type::getDoubleTy(TheContext), 2, "iftmp");

PN->addIncoming(ThenV, ThenBB);
PN->addIncoming(ElseV, ElseBB);
return PN;
}

// Output for-loop as:
//  var = alloca double
//  ...
//  start = startexpr
//  store start -> var
//  goto loop
// loop:
//  ...
//  bodyexpr
//  ...
// loopend:
//  step = stepexpr
//  endcond = endexpr
//
//  curvar = load var
//  nextvar = curvar + step
//  store nextvar -> var
//  br endcond, loop, endloop
// outloop:
Value *ForExprAST::codegen() {
    Function *TheFunction = Builder.GetInsertBlock()->getParent();

```

```
// Create an alloca for the variable in the entry block.
AllocaInst *Alloca = CreateEntryBlockAlloca(TheFunction, VarName);

// Emit the start code first, without 'variable' in scope.
Value *StartVal = Start->codegen();
if (!StartVal)
    return nullptr;

// Store the value into the alloca.
Builder.CreateStore(StartVal, Alloca);

// Make the new basic block for the loop header, inserting after current
// block.
BasicBlock *LoopBB = BasicBlock::Create(TheContext, "loop", TheFunction);

// Insert an explicit fall through from the current block to the LoopBB.
Builder.CreateBr(LoopBB);

// Start insertion in LoopBB.
Builder.SetInsertPoint(LoopBB);

// Within the loop, the variable is defined equal to the PHI node. If it
// shadows an existing variable, we have to restore it, so save it now.
AllocaInst *OldVal = NamedValues[VarName];
NamedValues[VarName] = Alloca;

// Emit the body of the loop. This, like any other expr, can change the
// current BB. Note that we ignore the value computed by the body, but don't
// allow an error.
if (!Body->codegen())
    return nullptr;

// Emit the step value.
Value *StepVal = nullptr;
if (Step) {
    StepVal = Step->codegen();
    if (!StepVal)
        return nullptr;
} else {
    // If not specified, use 1.0.
    StepVal = ConstantFP::get(TheContext, APFloat(1.0));
}

// Compute the end condition.
Value *EndCond = End->codegen();
if (!EndCond)
    return nullptr;

// Reload, increment, and restore the alloca. This handles the case where
// the body of the loop mutates the variable.
Value *CurVar = Builder.CreateLoad(Alloca, VarName.c_str());
Value *NextVar = Builder.CreateFAdd(CurVar, StepVal, "nextvar");
Builder.CreateStore(NextVar, Alloca);

// Convert condition to a bool by comparing non-equal to 0.0.
EndCond = Builder.CreateFCmpONE(
    EndCond, ConstantFP::get(TheContext, APFloat(0.0)), "loopcond");
```

```

// Create the "after loop" block and insert it.
BasicBlock *AfterBB =
    BasicBlock::Create(TheContext, "afterloop", TheFunction);

// Insert the conditional branch into the end of LoopEndBB.
Builder.CreateCondBr(EndCond, LoopBB, AfterBB);

// Any new code will be inserted in AfterBB.
Builder.SetInsertPoint(AfterBB);

// Restore the unshadowed variable.
if (OldVal)
    NamedValues[VarName] = OldVal;
else
    NamedValues.erase(VarName);

// for expr always returns 0.0.
return Constant::getNullValue(Type::getDoubleTy(TheContext));
}

Value *VarExprAST::codegen() {
    std::vector<AllocaInst *> OldBindings;

    Function *TheFunction = Builder.GetInsertBlock()->getParent();

    // Register all variables and emit their initializer.
    for (unsigned i = 0, e = VarNames.size(); i != e; ++i) {
        const std::string &VarName = VarNames[i].first;
        ExprAST *Init = VarNames[i].second.get();

        // Emit the initializer before adding the variable to scope, this prevents
        // the initializer from referencing the variable itself, and permits stuff
        // like this:
        //   var a = 1 in
        //     var a = a in ...   # refers to outer 'a'.
        Value *InitVal;
        if (Init) {
            InitVal = Init->codegen();
            if (!InitVal)
                return nullptr;
        } else { // If not specified, use 0.0.
            InitVal = ConstantFP::get(TheContext, APFloat(0.0));
        }

        AllocaInst *Alloca = CreateEntryBlockAlloca(TheFunction, VarName);
        Builder.CreateStore(InitVal, Alloca);

        // Remember the old variable binding so that we can restore the binding when
        // we unrecurse.
        OldBindings.push_back(NamedValues[VarName]);

        // Remember this binding.
        NamedValues[VarName] = Alloca;
    }

    // Codegen the body, now that all vars are in scope.
    Value *BodyVal = Body->codegen();
    if (!BodyVal)

```

```

    return nullptr;

    // Pop all our variables from scope.
    for (unsigned i = 0, e = VarNames.size(); i != e; ++i)
        NamedValues[VarNames[i].first] = OldBindings[i];

    // Return the body computation.
    return BodyVal;
}

Function *PrototypeAST::codegen() {
    // Make the function type: double(double,double) etc.
    std::vector<Type*> Doubles(Args.size(), Type::getDoubleTy(TheContext));
    FunctionType *FT =
        FunctionType::get(Type::getDoubleTy(TheContext), Doubles, false);

    Function *F =
        Function::Create(FT, Function::ExternalLinkage, Name, TheModule.get());

    // Set names for all arguments.
    unsigned Idx = 0;
    for (auto &Arg : F->args())
        Arg.setName(Args[Idx++]);

    return F;
}

Function *FunctionAST::codegen() {
    // Transfer ownership of the prototype to the FunctionProtos map, but keep a
    // reference to it for use below.
    auto &P = *Proto;
    FunctionProtos[Proto->getName()] = std::move(Proto);
    Function *TheFunction = getFunction(P.getName());
    if (!TheFunction)
        return nullptr;

    // If this is an operator, install it.
    if (P.isBinaryOp())
        BinopPrecedence[P.getOperatorName()] = P.getBinaryPrecedence();

    // Create a new basic block to start insertion into.
    BasicBlock *BB = BasicBlock::Create(TheContext, "entry", TheFunction);
    Builder.SetInsertPoint(BB);

    // Record the function arguments in the NamedValues map.
    NamedValues.clear();
    for (auto &Arg : TheFunction->args()) {
        // Create an alloca for this variable.
        AllocaInst *Alloca = CreateEntryBlockAlloca(TheFunction, Arg.getName());

        // Store the initial value into the alloca.
        Builder.CreateStore(&Arg, Alloca);

        // Add arguments to variable symbol table.
        NamedValues[Arg.getName()] = Alloca;
    }

    if (Value *RetVal = Body->codegen()) {

```



```

    // Finish off the function.
    Builder.CreateRet(RetVal);

    // Validate the generated code, checking for consistency.
    verifyFunction(*TheFunction);

    // Run the optimizer on the function.
    TheFPM->run(*TheFunction);

    return TheFunction;
}

// Error reading body, remove function.
TheFunction->eraseFromParent();

if (P.isBinaryOp())
    BinopPrecedence.erase(P.getOperatorName());
return nullptr;
}

//===-----//
// Top-Level parsing and JIT Driver
//===-----//

static void InitializeModuleAndPassManager() {
    // Open a new module.
    TheModule = llvm::make_unique<Module>("my cool jit", TheContext);
    TheModule->setDataLayout(TheJIT->getTargetMachine().createDataLayout());

    // Create a new pass manager attached to it.
    TheFPM = llvm::make_unique<legacy::FunctionPassManager>(TheModule.get());

    // Promote allocas to registers.
    TheFPM->add(createPromoteMemoryToRegisterPass());
    // Do simple "peephole" optimizations and bit-twiddling optzns.
    TheFPM->add(createInstructionCombiningPass());
    // Reassociate expressions.
    TheFPM->add(createReassociatePass());
    // Eliminate Common SubExpressions.
    TheFPM->add(createGVNPass());
    // Simplify the control flow graph (deleting unreachable blocks, etc).
    TheFPM->add(createCFGSimplificationPass());

    TheFPM->doInitialization();
}

static void HandleDefinition() {
    if (auto FnAST = ParseDefinition()) {
        if (auto *FnIR = FnAST->codegen()) {
            fprintf(stderr, "Read function definition:");
            FnIR->print(errs());
            fprintf(stderr, "\n");
            TheJIT->addModule(std::move(TheModule));
            InitializeModuleAndPassManager();
        }
    } else {
        // Skip token for error recovery.
        getNextToken();
    }
}

```

```

    }
}

static void HandleExtern() {
    if (auto ProtoAST = ParseExtern()) {
        if (auto *FnIR = ProtoAST->codegen()) {
            fprintf(stderr, "Read extern: ");
            FnIR->print(errs());
            fprintf(stderr, "\n");
            FunctionProtos[ProtoAST->getName()] = std::move(ProtoAST);
        }
    } else {
        // Skip token for error recovery.
        getNextToken();
    }
}

static void HandleTopLevelExpression() {
    // Evaluate a top-level expression into an anonymous function.
    if (auto FnAST = ParseTopLevelExpr()) {
        if (FnAST->codegen()) {
            // JIT the module containing the anonymous expression, keeping a handle so
            // we can free it later.
            auto H = TheJIT->addModule(std::move(TheModule));
            InitializeModuleAndPassManager();

            // Search the JIT for the __anon_expr symbol.
            auto ExprSymbol = TheJIT->findSymbol("__anon_expr");
            assert(ExprSymbol && "Function not found");

            // Get the symbol's address and cast it to the right type (takes no
            // arguments, returns a double) so we can call it as a native function.
            double (*FP)() = (double (*)(intptr_t))cantFail(ExprSymbol.getAddress());
            fprintf(stderr, "Evaluated to %f\n", FP());

            // Delete the anonymous expression module from the JIT.
            TheJIT->removeModule(H);
        }
    } else {
        // Skip token for error recovery.
        getNextToken();
    }
}

/// top ::= definition | external | expression | ';'
static void MainLoop() {
    while (true) {
        fprintf(stderr, "ready> ");
        switch (CurTok) {
            case tok_eof:
                return;
            case ';': // ignore top-level semicolons.
                getNextToken();
                break;
            case tok_def:
                HandleDefinition();
                break;
            case tok_extern:

```

```

    HandleExtern();
    break;
default:
    HandleTopLevelExpression();
    break;
}
}
}

//===-----//
// "Library" functions that can be "extern'd" from user code.
//===-----//

#ifdef LLVM_ON_WIN32
#define DLLEXPORT __declspec(dllexport)
#else
#define DLLEXPORT
#endif

// putchar - putchar that takes a double and returns 0.
extern "C" DLLEXPORT double putchar(double X) {
    fputc((char)X, stderr);
    return 0;
}

// printf - printf that takes a double prints it as "%f\n", returning 0.
extern "C" DLLEXPORT double printf(double X) {
    fprintf(stderr, "%f\n", X);
    return 0;
}

//===-----//
// Main driver code.
//===-----//

int main() {
    InitializeNativeTarget();
    InitializeNativeTargetAsmPrinter();
    InitializeNativeTargetAsmParser();

    // Install standard binary operators.
    // 1 is lowest precedence.
    BinopPrecedence['='] = 2;
    BinopPrecedence['<'] = 10;
    BinopPrecedence['+'] = 20;
    BinopPrecedence['-'] = 20;
    BinopPrecedence['*'] = 40; // highest.

    // Prime the first token.
    fprintf(stderr, "ready> ");
    getNextToken();

    TheJIT = llvm::make_unique<KaleidoscopeJIT>();

    InitializeModuleAndPassManager();

    // Run the main "interpreter loop" now.
    MainLoop();
}

```

```
    return 0;
}
```

Next: [Compiling to Object Code](#)

Kaleidoscope: Compiling to Object Code

- *Chapter 8 Introduction*
- *Choosing a target*
- *Target Machine*
- *Configuring the Module*
- *Emit Object Code*
- *Putting It All Together*
- *Full Code Listing*

Chapter 8 Introduction

Welcome to Chapter 8 of the “[Implementing a language with LLVM](#)” tutorial. This chapter describes how to compile our language down to object files.

Choosing a target

LLVM has native support for cross-compilation. You can compile to the architecture of your current machine, or just as easily compile for other architectures. In this tutorial, we’ll target the current machine.

To specify the architecture that you want to target, we use a string called a “target triple”. This takes the form `<arch><sub><vendor><sys><abi>` (see the [cross compilation docs](#)).

As an example, we can see what clang thinks is our current target triple:

```
$ clang --version | grep Target
Target: x86_64-unknown-linux-gnu
```

Running this command may show something different on your machine as you might be using a different architecture or operating system to me.

Fortunately, we don’t need to hard-code a target triple to target the current machine. LLVM provides `sys::getDefaultTargetTriple`, which returns the target triple of the current machine.

```
auto TargetTriple = sys::getDefaultTargetTriple();
```

LLVM doesn’t require us to link in all the target functionality. For example, if we’re just using the JIT, we don’t need the assembly printers. Similarly, if we’re only targeting certain architectures, we can only link in the functionality for those architectures.

For this example, we’ll initialize all the targets for emitting object code.

```
InitializeAllTargetInfos();
InitializeAllTargets();
InitializeAllTargetMCs();
InitializeAllAsmParsers();
InitializeAllAsmPrinters();
```

We can now use our target triple to get a Target:

```
std::string Error;
auto Target = TargetRegistry::lookupTarget(TargetTriple, Error);

// Print an error and exit if we couldn't find the requested target.
// This generally occurs if we've forgotten to initialise the
// TargetRegistry or we have a bogus target triple.
if (!Target) {
    errs() << Error;
    return 1;
}
```

Target Machine

We will also need a TargetMachine. This class provides a complete machine description of the machine we're targeting. If we want to target a specific feature (such as SSE) or a specific CPU (such as Intel's SandyLake), we do so now.

To see which features and CPUs that LLVM knows about, we can use `llc`. For example, let's look at x86:

```
$ llvm-as < /dev/null | llc -march=x86 -mattr=help
Available CPUs for this target:

    amdfam10      - Select the amdfam10 processor.
    athlon        - Select the athlon processor.
    athlon-4      - Select the athlon-4 processor.
    ...

Available features for this target:

    16bit-mode     - 16-bit mode (i8086).
    32bit-mode     - 32-bit mode (80386).
    3dnow          - Enable 3DNow! instructions.
    3dnowa         - Enable 3DNow! Athlon instructions.
    ...
```

For our example, we'll use the generic CPU without any additional features, options or relocation model.

```
auto CPU = "generic";
auto Features = "";

TargetOptions opt;
auto RM = Optional<Reloc::Model>();
auto TargetMachine = Target->createTargetMachine(TargetTriple, CPU, Features, opt,
    ↪RM);
```

Configuring the Module

We're now ready to configure our module, to specify the target and data layout. This isn't strictly necessary, but the [frontend performance guide](#) recommends this. Optimizations benefit from knowing about the target and data layout.

```
TheModule->setDataLayout(TargetMachine->createDataLayout());
TheModule->setTargetTriple(TargetTriple);
```

Emit Object Code

We're ready to emit object code! Let's define where we want to write our file to:

```
auto Filename = "output.o";
std::error_code EC;
raw_fd_ostream dest(Filename, EC, sys::fs::F_None);

if (EC) {
    errs() << "Could not open file: " << EC.message();
    return 1;
}
```

Finally, we define a pass that emits object code, then we run that pass:

```
legacy::PassManager pass;
auto FileType = TargetMachine::CGFT_ObjectFile;

if (TargetMachine->addPassesToEmitFile(pass, dest, FileType)) {
    errs() << "TargetMachine can't emit a file of this type";
    return 1;
}

pass.run(*TheModule);
dest.flush();
```

Putting It All Together

Does it work? Let's give it a try. We need to compile our code, but note that the arguments to `llvm-config` are different to the previous chapters.

```
$ clang++ -g -O3 toy.cpp `llvm-config --cxxflags --ldflags --system-libs --libs all` -
-o toy
```

Let's run it, and define a simple average function. Press Ctrl-D when you're done.

```
$ ./toy
ready> def average(x y) (x + y) * 0.5;
^D
Wrote output.o
```

We have an object file! To test it, let's write a simple program and link it with our output. Here's the source code:

```
#include <iostream>

extern "C" {
```

```

    double average(double, double);
}

int main() {
    std::cout << "average of 3.0 and 4.0: " << average(3.0, 4.0) << std::endl;
}

```

We link our program to output.o and check the result is what we expected:

```

$ clang++ main.cpp output.o -o main
$ ./main
average of 3.0 and 4.0: 3.5

```

Full Code Listing

```

#include "llvm/ADT/APFloat.h"
#include "llvm/ADT/Optional.h"
#include "llvm/ADT/STLExtras.h"
#include "llvm/IR/BasicBlock.h"
#include "llvm/IR/Constants.h"
#include "llvm/IR/DerivedTypes.h"
#include "llvm/IR/Function.h"
#include "llvm/IR/Instructions.h"
#include "llvm/IR/IRBuilder.h"
#include "llvm/IR/LLVMContext.h"
#include "llvm/IR/LegacyPassManager.h"
#include "llvm/IR/Module.h"
#include "llvm/IR/Type.h"
#include "llvm/IR/Verifier.h"
#include "llvm/Support/FileSystem.h"
#include "llvm/Support/Host.h"
#include "llvm/Support/raw_ostream.h"
#include "llvm/Support/TargetRegistry.h"
#include "llvm/Support/TargetSelect.h"
#include "llvm/Target/TargetMachine.h"
#include "llvm/Target/TargetOptions.h"
#include <algorithm>
#include <cassert>
#include <cctype>
#include <cstdio>
#include <cstdlib>
#include <map>
#include <memory>
#include <string>
#include <system_error>
#include <utility>
#include <vector>

using namespace llvm;
using namespace llvm::sys;

//===----- Lexer
// The lexer returns tokens [0-255] if it is an unknown character, otherwise one

```

```
// of these for known things.
enum Token {
    tok_eof = -1,

    // commands
    tok_def = -2,
    tok_extern = -3,

    // primary
    tok_identifier = -4,
    tok_number = -5,

    // control
    tok_if = -6,
    tok_then = -7,
    tok_else = -8,
    tok_for = -9,
    tok_in = -10,

    // operators
    tok_binary = -11,
    tok_unary = -12,

    // var definition
    tok_var = -13
};

static std::string IdentifierStr; // Filled in if tok_identifier
static double NumVal;           // Filled in if tok_number

/// gettok - Return the next token from standard input.
static int gettok() {
    static int LastChar = ' ';

    // Skip any whitespace.
    while (isspace(LastChar))
        LastChar = getchar();

    if (isalpha(LastChar)) { // identifier: [a-zA-Z][a-zA-Z0-9]*
        IdentifierStr = LastChar;
        while (isalnum((LastChar = getchar())))
            IdentifierStr += LastChar;

        if (IdentifierStr == "def")
            return tok_def;
        if (IdentifierStr == "extern")
            return tok_extern;
        if (IdentifierStr == "if")
            return tok_if;
        if (IdentifierStr == "then")
            return tok_then;
        if (IdentifierStr == "else")
            return tok_else;
        if (IdentifierStr == "for")
            return tok_for;
        if (IdentifierStr == "in")
            return tok_in;
        if (IdentifierStr == "binary")
```



```

    return tok_binary;
    if (IdentifierStr == "unary")
        return tok_unary;
    if (IdentifierStr == "var")
        return tok_var;
    return tok_identifier;
}

if (isdigit>LastChar) || LastChar == '.') { // Number: [0-9.]+
    std::string NumStr;
    do {
        NumStr += LastChar;
        LastChar = getchar();
    } while (isdigit>LastChar) || LastChar == '.');

    NumVal = strtod(NumStr.c_str(), nullptr);
    return tok_number;
}

if (LastChar == '#') {
    // Comment until end of line.
    do
        LastChar = getchar();
    while (LastChar != EOF && LastChar != '\n' && LastChar != '\r');

    if (LastChar != EOF)
        return gettok();
}

// Check for end of file. Don't eat the EOF.
if (LastChar == EOF)
    return tok_eof;

// Otherwise, just return the character as its ascii value.
int ThisChar = LastChar;
LastChar = getchar();
return ThisChar;
}

//===-----//
// Abstract Syntax Tree (aka Parse Tree)
//===-----//

namespace {

/// ExprAST - Base class for all expression nodes.
class ExprAST {
public:
    virtual ~ExprAST() = default;

    virtual Value *codegen() = 0;
};

/// NumberExprAST - Expression class for numeric literals like "1.0".
class NumberExprAST : public ExprAST {
    double Val;

public:

```

```
NumberExprAST(double Val) : Val(Val) {}

Value *codegen() override;
};

/// VariableExprAST - Expression class for referencing a variable, like "a".
class VariableExprAST : public ExprAST {
    std::string Name;

public:
    VariableExprAST(const std::string &Name) : Name(Name) {}

    Value *codegen() override;
    const std::string &getName() const { return Name; }
};

/// UnaryExprAST - Expression class for a unary operator.
class UnaryExprAST : public ExprAST {
    char Opcode;
    std::unique_ptr<ExprAST> Operand;

public:
    UnaryExprAST(char Opcode, std::unique_ptr<ExprAST> Operand)
        : Opcode(Opcode), Operand(std::move(Operand)) {}

    Value *codegen() override;
};

/// BinaryExprAST - Expression class for a binary operator.
class BinaryExprAST : public ExprAST {
    char Op;
    std::unique_ptr<ExprAST> LHS, RHS;

public:
    BinaryExprAST(char Op, std::unique_ptr<ExprAST> LHS,
                  std::unique_ptr<ExprAST> RHS)
        : Op(Op), LHS(std::move(LHS)), RHS(std::move(RHS)) {}

    Value *codegen() override;
};

/// CallExprAST - Expression class for function calls.
class CallExprAST : public ExprAST {
    std::string Callee;
    std::vector<std::unique_ptr<ExprAST>> Args;

public:
    CallExprAST(const std::string &Callee,
                std::vector<std::unique_ptr<ExprAST>> Args)
        : Callee(Callee), Args(std::move(Args)) {}

    Value *codegen() override;
};

/// IfExprAST - Expression class for if/then/else.
class IfExprAST : public ExprAST {
    std::unique_ptr<ExprAST> Cond, Then, Else;
```

```

public:
    IfExprAST(std::unique_ptr<ExprAST> Cond, std::unique_ptr<ExprAST> Then,
              std::unique_ptr<ExprAST> Else)
        : Cond(std::move(Cond)), Then(std::move(Then)), Else(std::move(Else)) {}

    Value *codegen() override;
};

/// ForExprAST - Expression class for for/in.
class ForExprAST : public ExprAST {
    std::string VarName;
    std::unique_ptr<ExprAST> Start, End, Step, Body;

public:
    ForExprAST(const std::string &VarName, std::unique_ptr<ExprAST> Start,
              std::unique_ptr<ExprAST> End, std::unique_ptr<ExprAST> Step,
              std::unique_ptr<ExprAST> Body)
        : VarName(VarName), Start(std::move(Start)), End(std::move(End)),
          Step(std::move(Step)), Body(std::move(Body)) {}

    Value *codegen() override;
};

/// VarExprAST - Expression class for var/in
class VarExprAST : public ExprAST {
    std::vector<std::pair<std::string, std::unique_ptr<ExprAST>>> VarNames;
    std::unique_ptr<ExprAST> Body;

public:
    VarExprAST(
        std::vector<std::pair<std::string, std::unique_ptr<ExprAST>>> VarNames,
        std::unique_ptr<ExprAST> Body)
        : VarNames(std::move(VarNames)), Body(std::move(Body)) {}

    Value *codegen() override;
};

/// PrototypeAST - This class represents the "prototype" for a function,
/// which captures its name, and its argument names (thus implicitly the number
/// of arguments the function takes), as well as if it is an operator.
class PrototypeAST {
    std::string Name;
    std::vector<std::string> Args;
    bool IsOperator;
    unsigned Precedence; // Precedence if a binary op.

public:
    PrototypeAST(const std::string &Name, std::vector<std::string> Args,
                bool IsOperator = false, unsigned Prec = 0)
        : Name(Name), Args(std::move(Args)), IsOperator(IsOperator),
          Precedence(Prec) {}

    Function *codegen();
    const std::string &getName() const { return Name; }

    bool isUnaryOp() const { return IsOperator && Args.size() == 1; }
    bool isBinaryOp() const { return IsOperator && Args.size() == 2; }

```

```

char getOperatorName() const {
    assert(isUnaryOp() || isBinaryOp());
    return Name[Name.size() - 1];
}

unsigned getBinaryPrecedence() const { return Precedence; }
};

/// FunctionAST - This class represents a function definition itself.
class FunctionAST {
    std::unique_ptr<PrototypeAST> Proto;
    std::unique_ptr<ExprAST> Body;

public:
    FunctionAST(std::unique_ptr<PrototypeAST> Proto,
                std::unique_ptr<ExprAST> Body)
        : Proto(std::move(Proto)), Body(std::move(Body)) {}

    Function *codegen();
};

} // end anonymous namespace

//=====//
// Parser
//=====//

/// CurTok/getNextToken - Provide a simple token buffer. CurTok is the current
/// token the parser is looking at. getNextToken reads another token from the
/// lexer and updates CurTok with its results.
static int CurTok;
static int getNextToken() { return CurTok = gettok(); }

/// BinopPrecedence - This holds the precedence for each binary operator that is
/// defined.
static std::map<char, int> BinopPrecedence;

/// GetTokPrecedence - Get the precedence of the pending binary operator token.
static int GetTokPrecedence() {
    if (!isascii(CurTok))
        return -1;

    // Make sure it's a declared binop.
    int TokPrec = BinopPrecedence[CurTok];
    if (TokPrec <= 0)
        return -1;
    return TokPrec;
}

/// LogError* - These are little helper functions for error handling.
std::unique_ptr<ExprAST> LogError(const char *Str) {
    fprintf(stderr, "Error: %s\n", Str);
    return nullptr;
}

std::unique_ptr<PrototypeAST> LogErrorP(const char *Str) {
    LogError(Str);
    return nullptr;
}

```

```

}

static std::unique_ptr<ExprAST> ParseExpression();

/// numberexpr ::= number
static std::unique_ptr<ExprAST> ParseNumberExpr() {
    auto Result = llvm::make_unique<NumberExprAST>(NumVal);
    getNextToken(); // consume the number
    return std::move(Result);
}

/// parenexpr ::= '(' expression ')'
static std::unique_ptr<ExprAST> ParseParenExpr() {
    getNextToken(); // eat (.
    auto V = ParseExpression();
    if (!V)
        return nullptr;

    if (CurTok != ')')
        return LogError("expected ') '");
    getNextToken(); // eat ).
    return V;
}

/// identifierexpr
/// ::= identifier
/// ::= identifier '(' expression* ')'
static std::unique_ptr<ExprAST> ParseIdentifierExpr() {
    std::string IdName = IdentifierStr;

    getNextToken(); // eat identifier.

    if (CurTok != '(') // Simple variable ref.
        return llvm::make_unique<VariableExprAST>(IdName);

    // Call.
    getNextToken(); // eat (
    std::vector<std::unique_ptr<ExprAST>> Args;
    if (CurTok != ')') {
        while (true) {
            if (auto Arg = ParseExpression())
                Args.push_back(std::move(Arg));
            else
                return nullptr;

            if (CurTok == ')')
                break;

            if (CurTok != ',')
                return LogError("Expected ')' or ',' in argument list");
            getNextToken();
        }
    }

    // Eat the ') '.
    getNextToken();

    return llvm::make_unique<CallExprAST>(IdName, std::move(Args));
}

```

```

}

/// ifexpr ::= 'if' expression 'then' expression 'else' expression
static std::unique_ptr<ExprAST> ParseIfExpr() {
    getNextToken(); // eat the if.

    // condition.
    auto Cond = ParseExpression();
    if (!Cond)
        return nullptr;

    if (CurTok != tok_then)
        return LogError("expected then");
    getNextToken(); // eat the then

    auto Then = ParseExpression();
    if (!Then)
        return nullptr;

    if (CurTok != tok_else)
        return LogError("expected else");

    getNextToken();

    auto Else = ParseExpression();
    if (!Else)
        return nullptr;

    return llvm::make_unique<IfExprAST>(std::move(Cond), std::move(Then),
                                        std::move(Else));
}

/// forexpr ::= 'for' identifier '=' expr ',' expr (',' expr)? 'in' expression
static std::unique_ptr<ExprAST> ParseForExpr() {
    getNextToken(); // eat the for.

    if (CurTok != tok_identifier)
        return LogError("expected identifier after for");

    std::string IdName = IdentifierStr;
    getNextToken(); // eat identifier.

    if (CurTok != '=')
        return LogError("expected '=' after for");
    getNextToken(); // eat '='.

    auto Start = ParseExpression();
    if (!Start)
        return nullptr;
    if (CurTok != ',')
        return LogError("expected ',' after for start value");
    getNextToken();

    auto End = ParseExpression();
    if (!End)
        return nullptr;

    // The step value is optional.

```

```

std::unique_ptr<ExprAST> Step;
if (CurTok == ',') {
    getNextToken();
    Step = ParseExpression();
    if (!Step)
        return nullptr;
}

if (CurTok != tok_in)
    return LogError("expected 'in' after for");
getNextToken(); // eat 'in'.

auto Body = ParseExpression();
if (!Body)
    return nullptr;

return llvm::make_unique<ForExprAST>(IdName, std::move(Start), std::move(End),
                                     std::move(Step), std::move(Body));
}

/// varexpr ::= 'var' identifier ('=' expression)?
///           (' identifier ('=' expression)?)* 'in' expression
static std::unique_ptr<ExprAST> ParseVarExpr() {
    getNextToken(); // eat the var.

    std::vector<std::pair<std::string, std::unique_ptr<ExprAST>>> VarNames;

    /// At least one variable name is required.
    if (CurTok != tok_identifier)
        return LogError("expected identifier after var");

    while (true) {
        std::string Name = IdentifierStr;
        getNextToken(); // eat identifier.

        /// Read the optional initializer.
        std::unique_ptr<ExprAST> Init = nullptr;
        if (CurTok == '=') {
            getNextToken(); // eat the '='.

            Init = ParseExpression();
            if (!Init)
                return nullptr;
        }

        VarNames.push_back(std::make_pair(Name, std::move(Init)));

        /// End of var list, exit loop.
        if (CurTok != ',')
            break;
        getNextToken(); // eat the ','.

        if (CurTok != tok_identifier)
            return LogError("expected identifier list after var");
    }

    /// At this point, we have to have 'in'.
    if (CurTok != tok_in)

```

```

    return LogError("expected 'in' keyword after 'var'");
getNextToken(); // eat 'in'.

auto Body = ParseExpression();
if (!Body)
    return nullptr;

return llvm::make_unique<VarExprAST>(std::move(VarNames), std::move(Body));
}

/// primary
/// ::= identifierexpr
/// ::= numberexpr
/// ::= parenexpr
/// ::= ifexpr
/// ::= forexpr
/// ::= varexpr
static std::unique_ptr<ExprAST> ParsePrimary() {
    switch (CurTok) {
    default:
        return LogError("unknown token when expecting an expression");
    case tok_identifier:
        return ParseIdentifierExpr();
    case tok_number:
        return ParseNumberExpr();
    case '(':
        return ParseParenExpr();
    case tok_if:
        return ParseIfExpr();
    case tok_for:
        return ParseForExpr();
    case tok_var:
        return ParseVarExpr();
    }
}

/// unary
/// ::= primary
/// ::= '!' unary
static std::unique_ptr<ExprAST> ParseUnary() {
    // If the current token is not an operator, it must be a primary expr.
    if (!isascii(CurTok) || CurTok == '(' || CurTok == ',')
        return ParsePrimary();

    // If this is a unary operator, read it.
    int Opc = CurTok;
    getNextToken();
    if (auto Operand = ParseUnary())
        return llvm::make_unique<UnaryExprAST>(Opc, std::move(Operand));
    return nullptr;
}

/// binoprhs
/// ::= ('+' unary)*
static std::unique_ptr<ExprAST> ParseBinOpRHS(int ExprPrec,
                                              std::unique_ptr<ExprAST> LHS) {
    // If this is a binop, find its precedence.
    while (true) {

```



```

    int TokPrec = GetTokPrecedence();

    // If this is a binop that binds at least as tightly as the current binop,
    // consume it, otherwise we are done.
    if (TokPrec < ExprPrec)
        return LHS;

    // Okay, we know this is a binop.
    int BinOp = CurTok;
    getNextToken(); // eat binop

    // Parse the unary expression after the binary operator.
    auto RHS = ParseUnary();
    if (!RHS)
        return nullptr;

    // If BinOp binds less tightly with RHS than the operator after RHS, let
    // the pending operator take RHS as its LHS.
    int NextPrec = GetTokPrecedence();
    if (TokPrec < NextPrec) {
        RHS = ParseBinOpRHS(TokPrec + 1, std::move(RHS));
        if (!RHS)
            return nullptr;
    }

    // Merge LHS/RHS.
    LHS =
        llvm::make_unique<BinaryExprAST>(BinOp, std::move(LHS), std::move(RHS));
}

/// expression
/// ::= unary binoprhs
///
static std::unique_ptr<ExprAST> ParseExpression() {
    auto LHS = ParseUnary();
    if (!LHS)
        return nullptr;

    return ParseBinOpRHS(0, std::move(LHS));
}

/// prototype
/// ::= id '(' id* ')'
/// ::= binary LETTER number? (id, id)
/// ::= unary LETTER (id)
static std::unique_ptr<PrototypeAST> ParsePrototype() {
    std::string FnName;

    unsigned Kind = 0; // 0 = identifier, 1 = unary, 2 = binary.
    unsigned BinaryPrecedence = 30;

    switch (CurTok) {
    default:
        return LogErrorP("Expected function name in prototype");
    case tok_identifier:
        FnName = IdentifierStr;
        Kind = 0;

```

```

    getNextToken();
    break;
case tok_unary:
    getNextToken();
    if (!isascii(CurTok))
        return LogErrorP("Expected unary operator");
    FnName = "unary";
    FnName += (char)CurTok;
    Kind = 1;
    getNextToken();
    break;
case tok_binary:
    getNextToken();
    if (!isascii(CurTok))
        return LogErrorP("Expected binary operator");
    FnName = "binary";
    FnName += (char)CurTok;
    Kind = 2;
    getNextToken();

    // Read the precedence if present.
    if (CurTok == tok_number) {
        if (NumVal < 1 || NumVal > 100)
            return LogErrorP("Invalid precedence: must be 1..100");
        BinaryPrecedence = (unsigned)NumVal;
        getNextToken();
    }
    break;
}

if (CurTok != '(')
    return LogErrorP("Expected '(' in prototype");

std::vector<std::string> ArgNames;
while (getNextToken() == tok_identifier)
    ArgNames.push_back(IdentifierStr);
if (CurTok != ')')
    return LogErrorP("Expected ')' in prototype");

// success.
getNextToken(); // eat ')'.

// Verify right number of names for operator.
if (Kind && ArgNames.size() != Kind)
    return LogErrorP("Invalid number of operands for operator");

return llvm::make_unique<PrototypeAST>(FnName, ArgNames, Kind != 0,
                                       BinaryPrecedence);
}

/// definition ::= 'def' prototype expression
static std::unique_ptr<FunctionAST> ParseDefinition() {
    getNextToken(); // eat def.
    auto Proto = ParsePrototype();
    if (!Proto)
        return nullptr;

    if (auto E = ParseExpression())

```

```

    return llvm::make_unique<FunctionAST>(std::move(Proto), std::move(E));
    return nullptr;
}

/// toplevelexpr ::= expression
static std::unique_ptr<FunctionAST> ParseTopLevelExpr() {
    if (auto E = ParseExpression()) {
        // Make an anonymous proto.
        auto Proto = llvm::make_unique<PrototypeAST>("__anon_expr",
                                                    std::vector<std::string>());
        return llvm::make_unique<FunctionAST>(std::move(Proto), std::move(E));
    }
    return nullptr;
}

/// external ::= 'extern' prototype
static std::unique_ptr<PrototypeAST> ParseExtern() {
    getNextToken(); // eat extern.
    return ParsePrototype();
}

//=====//
// Code Generation
//=====//

static LLVMContext TheContext;
static IRBuilder<> Builder(TheContext);
static std::unique_ptr<Module> TheModule;
static std::map<std::string, AllocaInst *> NamedValues;
static std::map<std::string, std::unique_ptr<PrototypeAST>> FunctionProtos;

Value *LogErrorV(const char *Str) {
    LogError(Str);
    return nullptr;
}

Function *getFunction(std::string Name) {
    // First, see if the function has already been added to the current module.
    if (auto *F = TheModule->getFunction(Name))
        return F;

    // If not, check whether we can codegen the declaration from some existing
// prototype.
    auto FI = FunctionProtos.find(Name);
    if (FI != FunctionProtos.end())
        return FI->second->codegen();

    // If no existing prototype exists, return null.
    return nullptr;
}

/// CreateEntryBlockAlloca - Create an alloca instruction in the entry block of
/// the function. This is used for mutable variables etc.
static AllocaInst *CreateEntryBlockAlloca(Function *TheFunction,
                                           const std::string &VarName) {
    IRBuilder<> TmpB(&TheFunction->getEntryBlock(),
                  TheFunction->getEntryBlock().begin());
    return TmpB.CreateAlloca(Type::getDoubleTy(TheContext), nullptr, VarName);
}

```

```
}

Value *NumberExprAST::codegen() {
    return ConstantFP::get(TheContext, APFloat(Val));
}

Value *VariableExprAST::codegen() {
    // Look this variable up in the function.
    Value *V = NamedValues[Name];
    if (!V)
        return LogErrorV("Unknown variable name");

    // Load the value.
    return Builder.CreateLoad(V, Name.c_str());
}

Value *UnaryExprAST::codegen() {
    Value *OperandV = Operand->codegen();
    if (!OperandV)
        return nullptr;

    Function *F = getFunction(std::string("unary") + Opcode);
    if (!F)
        return LogErrorV("Unknown unary operator");

    return Builder.CreateCall(F, OperandV, "unop");
}

Value *BinaryExprAST::codegen() {
    // Special case '=' because we don't want to emit the LHS as an expression.
    if (Op == '=') {
        // Assignment requires the LHS to be an identifier.
        // This assume we're building without RTTI because LLVM builds that way by
        // default. If you build LLVM with RTTI this can be changed to a
        // dynamic_cast for automatic error checking.
        VariableExprAST *LHSE = static_cast<VariableExprAST *>(LHS.get());
        if (!LHSE)
            return LogErrorV("destination of '=' must be a variable");
        // Codegen the RHS.
        Value *Val = RHS->codegen();
        if (!Val)
            return nullptr;

        // Look up the name.
        Value *Variable = NamedValues[LHSE->getName()];
        if (!Variable)
            return LogErrorV("Unknown variable name");

        Builder.CreateStore(Val, Variable);
        return Val;
    }

    Value *L = LHS->codegen();
    Value *R = RHS->codegen();
    if (!L || !R)
        return nullptr;

    switch (Op) {
```

```

    case '+':
        return Builder.CreateFAdd(L, R, "addtmp");
    case '-':
        return Builder.CreateFSub(L, R, "subtmp");
    case '*':
        return Builder.CreateFMul(L, R, "multmp");
    case '<':
        L = Builder.CreateFCmpULT(L, R, "cmptmp");
        // Convert bool 0/1 to double 0.0 or 1.0
        return Builder.CreateUIToFP(L, Type::getDoubleTy(TheContext), "booltmp");
    default:
        break;
}

// If it wasn't a builtin binary operator, it must be a user defined one. Emit
// a call to it.
Function *F = getFunction(std::string("binary") + Op);
assert(F && "binary operator not found!");

Value *Ops[] = {L, R};
return Builder.CreateCall(F, Ops, "binop");
}

Value *CallExprAST::codegen() {
    // Look up the name in the global module table.
    Function *CalleeF = getFunction(Callee);
    if (!CalleeF)
        return LogErrorV("Unknown function referenced");

    // If argument mismatch error.
    if (CalleeF->arg_size() != Args.size())
        return LogErrorV("Incorrect # arguments passed");

    std::vector<Value *> ArgsV;
    for (unsigned i = 0, e = Args.size(); i != e; ++i) {
        ArgsV.push_back(Args[i]->codegen());
        if (!ArgsV.back())
            return nullptr;
    }

    return Builder.CreateCall(CalleeF, ArgsV, "calltmp");
}

Value *IfExprAST::codegen() {
    Value *CondV = Cond->codegen();
    if (!CondV)
        return nullptr;

    // Convert condition to a bool by comparing non-equal to 0.0.
    CondV = Builder.CreateFCmpONE(
        CondV, ConstantFP::get(TheContext, APFloat(0.0)), "ifcond");

    Function *TheFunction = Builder.GetInsertBlock()->getParent();

    // Create blocks for the then and else cases. Insert the 'then' block at the
    // end of the function.
    BasicBlock *ThenBB = BasicBlock::Create(TheContext, "then", TheFunction);
    BasicBlock *ElseBB = BasicBlock::Create(TheContext, "else");

```

```

BasicBlock *MergeBB = BasicBlock::Create(TheContext, "ifcont");

Builder.CreateCondBr(CondV, ThenBB, ElseBB);

// Emit then value.
Builder.SetInsertPoint(ThenBB);

Value *ThenV = Then->codegen();
if (!ThenV)
    return nullptr;

Builder.CreateBr(MergeBB);
// Codegen of 'Then' can change the current block, update ThenBB for the PHI.
ThenBB = Builder.GetInsertBlock();

// Emit else block.
TheFunction->getBasicBlockList().push_back(ElseBB);
Builder.SetInsertPoint(ElseBB);

Value *ElseV = Else->codegen();
if (!ElseV)
    return nullptr;

Builder.CreateBr(MergeBB);
// Codegen of 'Else' can change the current block, update ElseBB for the PHI.
ElseBB = Builder.GetInsertBlock();

// Emit merge block.
TheFunction->getBasicBlockList().push_back(MergeBB);
Builder.SetInsertPoint(MergeBB);
PHINode *PN = Builder.CreatePHI(Type::getDoubleTy(TheContext), 2, "iftmp");

PN->addIncoming(ThenV, ThenBB);
PN->addIncoming(ElseV, ElseBB);
return PN;
}

// Output for-loop as:
//  var = alloca double
//  ...
//  start = startexpr
//  store start -> var
//  goto loop
// loop:
//  ...
//  bodyexpr
//  ...
// loopend:
//  step = stepexpr
//  endcond = endexpr
//
//  curvar = load var
//  nextvar = curvar + step
//  store nextvar -> var
//  br endcond, loop, endloop
// outloop:
Value *ForExprAST::codegen() {
    Function *TheFunction = Builder.GetInsertBlock()->getParent();

```

```

// Create an alloca for the variable in the entry block.
AllocaInst *Alloca = CreateEntryBlockAlloca(TheFunction, VarName);

// Emit the start code first, without 'variable' in scope.
Value *StartVal = Start->codegen();
if (!StartVal)
    return nullptr;

// Store the value into the alloca.
Builder.CreateStore(StartVal, Alloca);

// Make the new basic block for the loop header, inserting after current
// block.
BasicBlock *LoopBB = BasicBlock::Create(TheContext, "loop", TheFunction);

// Insert an explicit fall through from the current block to the LoopBB.
Builder.CreateBr(LoopBB);

// Start insertion in LoopBB.
Builder.SetInsertPoint(LoopBB);

// Within the loop, the variable is defined equal to the PHI node. If it
// shadows an existing variable, we have to restore it, so save it now.
AllocaInst *OldVal = NamedValues[VarName];
NamedValues[VarName] = Alloca;

// Emit the body of the loop. This, like any other expr, can change the
// current BB. Note that we ignore the value computed by the body, but don't
// allow an error.
if (!Body->codegen())
    return nullptr;

// Emit the step value.
Value *StepVal = nullptr;
if (Step) {
    StepVal = Step->codegen();
    if (!StepVal)
        return nullptr;
} else {
    // If not specified, use 1.0.
    StepVal = ConstantFP::get(TheContext, APFloat(1.0));
}

// Compute the end condition.
Value *EndCond = End->codegen();
if (!EndCond)
    return nullptr;

// Reload, increment, and restore the alloca. This handles the case where
// the body of the loop mutates the variable.
Value *CurVar = Builder.CreateLoad(Alloca, VarName.c_str());
Value *NextVar = Builder.CreateFAdd(CurVar, StepVal, "nextvar");
Builder.CreateStore(NextVar, Alloca);

// Convert condition to a bool by comparing non-equal to 0.0.
EndCond = Builder.CreateFCmpONE(
    EndCond, ConstantFP::get(TheContext, APFloat(0.0)), "loopcond");

```

```

// Create the "after loop" block and insert it.
BasicBlock *AfterBB =
    BasicBlock::Create(TheContext, "afterloop", TheFunction);

// Insert the conditional branch into the end of LoopEndBB.
Builder.CreateCondBr(EndCond, LoopBB, AfterBB);

// Any new code will be inserted in AfterBB.
Builder.SetInsertPoint(AfterBB);

// Restore the unshadowed variable.
if (OldVal)
    NamedValues[VarName] = OldVal;
else
    NamedValues.erase(VarName);

// for expr always returns 0.0.
return Constant::getNullValue(Type::getDoubleTy(TheContext));
}

Value *VarExprAST::codegen() {
    std::vector<AllocaInst *> OldBindings;

    Function *TheFunction = Builder.GetInsertBlock()->getParent();

    // Register all variables and emit their initializer.
    for (unsigned i = 0, e = VarNames.size(); i != e; ++i) {
        const std::string &VarName = VarNames[i].first;
        ExprAST *Init = VarNames[i].second.get();

        // Emit the initializer before adding the variable to scope, this prevents
        // the initializer from referencing the variable itself, and permits stuff
        // like this:
        //   var a = 1 in
        //   var a = a in ... # refers to outer 'a'.
        Value *InitVal;
        if (Init) {
            InitVal = Init->codegen();
            if (!InitVal)
                return nullptr;
        } else { // If not specified, use 0.0.
            InitVal = ConstantFP::get(TheContext, APFloat(0.0));
        }

        AllocaInst *Alloca = CreateEntryBlockAlloca(TheFunction, VarName);
        Builder.CreateStore(InitVal, Alloca);

        // Remember the old variable binding so that we can restore the binding when
        // we unrecurse.
        OldBindings.push_back(NamedValues[VarName]);

        // Remember this binding.
        NamedValues[VarName] = Alloca;
    }

    // Codegen the body, now that all vars are in scope.
    Value *BodyVal = Body->codegen();

```



```

if (!BodyVal)
    return nullptr;

// Pop all our variables from scope.
for (unsigned i = 0, e = VarNames.size(); i != e; ++i)
    NamedValues[VarNames[i].first] = OldBindings[i];

// Return the body computation.
return BodyVal;
}

Function *PrototypeAST::codegen() {
    // Make the function type: double(double,double) etc.
    std::vector<Type *> Doubles(Args.size(), Type::getDoubleTy(TheContext));
    FunctionType *FT =
        FunctionType::get(Type::getDoubleTy(TheContext), Doubles, false);

    Function *F =
        Function::Create(FT, Function::ExternalLinkage, Name, TheModule.get());

    // Set names for all arguments.
    unsigned Idx = 0;
    for (auto &Arg : F->args())
        Arg.setName(Args[Idx++]);

    return F;
}

Function *FunctionAST::codegen() {
    // Transfer ownership of the prototype to the FunctionProtos map, but keep a
    // reference to it for use below.
    auto &P = *Proto;
    FunctionProtos[Proto->getName()] = std::move(Proto);
    Function *TheFunction = getFunction(P.getName());
    if (!TheFunction)
        return nullptr;

    // If this is an operator, install it.
    if (P.isBinaryOp())
        BinopPrecedence[P.getOperatorName()] = P.getBinaryPrecedence();

    // Create a new basic block to start insertion into.
    BasicBlock *BB = BasicBlock::Create(TheContext, "entry", TheFunction);
    Builder.SetInsertPoint(BB);

    // Record the function arguments in the NamedValues map.
    NamedValues.clear();
    for (auto &Arg : TheFunction->args()) {
        // Create an alloca for this variable.
        AllocaInst *Alloca = CreateEntryBlockAlloca(TheFunction, Arg.getName());

        // Store the initial value into the alloca.
        Builder.CreateStore(&Arg, Alloca);

        // Add arguments to variable symbol table.
        NamedValues[Arg.getName()] = Alloca;
    }
}

```

```
if (Value *RetVal = Body->codegen()) {
    // Finish off the function.
    Builder.CreateRet(RetVal);

    // Validate the generated code, checking for consistency.
    verifyFunction(*TheFunction);

    return TheFunction;
}

// Error reading body, remove function.
TheFunction->eraseFromParent();

if (P.isBinaryOp())
    BinopPrecedence.erase(P.getOperatorName());
return nullptr;
}

//===-----//
// Top-Level parsing and JIT Driver
//===-----//

static void InitializeModuleAndPassManager() {
    // Open a new module.
    TheModule = llvm::make_unique<Module>("my cool jit", TheContext);
}

static void HandleDefinition() {
    if (auto FnAST = ParseDefinition()) {
        if (auto *FnIR = FnAST->codegen()) {
            fprintf(stderr, "Read function definition:");
            FnIR->print(errs());
            fprintf(stderr, "\n");
        }
    } else {
        // Skip token for error recovery.
        getNextToken();
    }
}

static void HandleExtern() {
    if (auto ProtoAST = ParseExtern()) {
        if (auto *FnIR = ProtoAST->codegen()) {
            fprintf(stderr, "Read extern: ");
            FnIR->print(errs());
            fprintf(stderr, "\n");
            FunctionProtos[ProtoAST->getName()] = std::move(ProtoAST);
        }
    } else {
        // Skip token for error recovery.
        getNextToken();
    }
}

static void HandleTopLevelExpression() {
    // Evaluate a top-level expression into an anonymous function.
    if (auto FnAST = ParseTopLevelExpr()) {
        FnAST->codegen();
    }
}
```

```

    } else {
        // Skip token for error recovery.
        getNextToken();
    }
}

// top ::= definition | external | expression | ';'
static void MainLoop() {
    while (true) {
        switch (CurTok) {
            case tok_eof:
                return;
            case ';': // ignore top-level semicolons.
                getNextToken();
                break;
            case tok_def:
                HandleDefinition();
                break;
            case tok_extern:
                HandleExtern();
                break;
            default:
                HandleTopLevelExpression();
                break;
        }
    }
}

//=====//
// "Library" functions that can be "extern'd" from user code.
//=====//

#ifdef LLVM_ON_WIN32
#define DLLEXPORT __declspec(dllexport)
#else
#define DLLEXPORT
#endif

// putchar - putchar that takes a double and returns 0.
extern "C" DLLEXPORT double putchar(double X) {
    fputc((char)X, stderr);
    return 0;
}

// printf - printf that takes a double prints it as "%f\n", returning 0.
extern "C" DLLEXPORT double printf(double X) {
    fprintf(stderr, "%f\n", X);
    return 0;
}

//=====//
// Main driver code.
//=====//

int main() {
    // Install standard binary operators.
    // 1 is lowest precedence.
    BinopPrecedence['<'] = 10;

```

```
BinopPrecedence['+'] = 20;
BinopPrecedence['-'] = 20;
BinopPrecedence['*'] = 40; // highest.

// Prime the first token.
fprintf(stderr, "ready> ");
getNextToken();

InitializeModuleAndPassManager();

// Run the main "interpreter loop" now.
MainLoop();

// Initialize the target registry etc.
InitializeAllTargetInfos();
InitializeAllTargets();
InitializeAllTargetMCs();
InitializeAllAsmParsers();
InitializeAllAsmPrinters();

auto TargetTriple = sys::getDefaultTargetTriple();
TheModule->setTargetTriple(TargetTriple);

std::string Error;
auto Target = TargetRegistry::lookupTarget(TargetTriple, Error);

// Print an error and exit if we couldn't find the requested target.
// This generally occurs if we've forgotten to initialise the
// TargetRegistry or we have a bogus target triple.
if (!Target) {
    errs() << Error;
    return 1;
}

auto CPU = "generic";
auto Features = "";

TargetOptions opt;
auto RM = Optional<Reloc::Model>();
auto TheTargetMachine =
    Target->createTargetMachine(TargetTriple, CPU, Features, opt, RM);

TheModule->setDataLayout(TheTargetMachine->createDataLayout());

auto Filename = "output.o";
std::error_code EC;
raw_fd_ostream dest(Filename, EC, sys::fs::F_None);

if (EC) {
    errs() << "Could not open file: " << EC.message();
    return 1;
}

legacy::PassManager pass;
auto FileType = TargetMachine::CGFT_ObjectFile;

if (TheTargetMachine->addPassesToEmitFile(pass, dest, FileType)) {
    errs() << "TheTargetMachine can't emit a file of this type";
}
```

```

    return 1;
}

pass.run(*TheModule);
dest.flush();

outs() << "Wrote " << Filename << "\n";

return 0;
}

```

Next: Adding Debug Information

Kaleidoscope: Adding Debug Information

- *Chapter 9 Introduction*
- *Why is this a hard problem?*
- *Ahead-of-Time Compilation Mode*
- *Compile Unit*
- *DWARF Emission Setup*
- *Functions*
- *Source Locations*
- *Variables*
- *Full Code Listing*

Chapter 9 Introduction

Welcome to Chapter 9 of the “[Implementing a language with LLVM](#)” tutorial. In chapters 1 through 8, we’ve built a decent little programming language with functions and variables. What happens if something goes wrong though, how do you debug your program?

Source level debugging uses formatted data that helps a debugger translate from binary and the state of the machine back to the source that the programmer wrote. In LLVM we generally use a format called [DWARF](#). DWARF is a compact encoding that represents types, source locations, and variable locations.

The short summary of this chapter is that we’ll go through the various things you have to add to a programming language to support debug info, and how you translate that into DWARF.

Caveat: For now we can’t debug via the JIT, so we’ll need to compile our program down to something small and standalone. As part of this we’ll make a few modifications to the running of the language and how programs are compiled. This means that we’ll have a source file with a simple program written in Kaleidoscope rather than the interactive JIT. It does involve a limitation that we can only have one “top level” command at a time to reduce the number of changes necessary.

Here’s the sample program we’ll be compiling:

```

def fib(x)
  if x < 3 then

```

```
1
else
    fib(x-1)+fib(x-2);

fib(10)
```

Why is this a hard problem?

Debug information is a hard problem for a few different reasons - mostly centered around optimized code. First, optimization makes keeping source locations more difficult. In LLVM IR we keep the original source location for each IR level instruction on the instruction. Optimization passes should keep the source locations for newly created instructions, but merged instructions only get to keep a single location - this can cause jumping around when stepping through optimized programs. Secondly, optimization can move variables in ways that are either optimized out, shared in memory with other variables, or difficult to track. For the purposes of this tutorial we're going to avoid optimization (as you'll see with one of the next sets of patches).

Ahead-of-Time Compilation Mode

To highlight only the aspects of adding debug information to a source language without needing to worry about the complexities of JIT debugging we're going to make a few changes to Kaleidoscope to support compiling the IR emitted by the front end into a simple standalone program that you can execute, debug, and see results.

First we make our anonymous function that contains our top level statement be our "main":

```
- auto Proto = llvm::make_unique<PrototypeAST>("", std::vector<std::string>());
+ auto Proto = llvm::make_unique<PrototypeAST>("main", std::vector<std::string>());
```

just with the simple change of giving it a name.

Then we're going to remove the command line code wherever it exists:

```
@@ -1129,7 +1129,6 @@ static void HandleTopLevelExpression() {
    /// top ::= definition | external | expression | ';'
    static void MainLoop() {
        while (1) {
-         fprintf(stderr, "ready> ");
            switch (CurTok) {
                case tok_eof:
                    return;
@@ -1184,7 +1183,6 @@ int main() {
    BinopPrecedence['*'] = 40; // highest.

    // Prime the first token.
-    fprintf(stderr, "ready> ");
    getNextToken();
```

Lastly we're going to disable all of the optimization passes and the JIT so that the only thing that happens after we're done parsing and generating code is that the LLVM IR goes to standard error:

```
@@ -1108,17 +1108,8 @@ static void HandleExtern() {
    static void HandleTopLevelExpression() {
        // Evaluate a top-level expression into an anonymous function.
        if (auto FnAST = ParseTopLevelExpr()) {
-         if (auto *FnIR = FnAST->codegen()) {
```

```

-      // We're just doing this to make sure it executes.
-      TheExecutionEngine->finalizeObject();
-      // JIT the function, returning a function pointer.
-      void *FPtr = TheExecutionEngine->getPointerToFunction(FnIR);
-
-      // Cast it to the right type (takes no arguments, returns a double) so we
-      // can call it as a native function.
-      double (*FP)() = (double (*)(()))(intptr_t)FPtr;
-      // Ignore the return value for this.
-      (void)FP;
+      if (!F->codegen()) {
+          fprintf(stderr, "Error generating code for top level expr");
+      }
+      } else {
+          // Skip token for error recovery.
@@ -1439,11 +1459,11 @@ int main() {
+      // target lays out data structures.
+      TheModule->setDataLayout(TheExecutionEngine->getDataLayout());
+      OurFPM.add(new DataLayoutPass());
+  #if 0
+      OurFPM.add(createBasicAliasAnalysisPass());
+      // Promote allocas to registers.
+      OurFPM.add(createPromoteMemoryToRegisterPass());
@@ -1218,7 +1210,7 @@ int main() {
+      OurFPM.add(createGVNPass());
+      // Simplify the control flow graph (deleting unreachable blocks, etc).
+      OurFPM.add(createCFGSimplificationPass());
-
+  #endif
+      OurFPM.doInitialization();

+      // Set the global so the code gen can use this.

```

This relatively small set of changes get us to the point that we can compile our piece of Kaleidoscope language down to an executable program via this command line:

```
Kaleidoscope-Ch9 < fib.ks | & clang -x ir -
```

which gives an a.out/a.exe in the current working directory.

Compile Unit

The top level container for a section of code in DWARF is a compile unit. This contains the type and function data for an individual translation unit (read: one file of source code). So the first thing we need to do is construct one for our fib.ks file.

DWARF Emission Setup

Similar to the `IRBuilder` class we have a `DIBuilder` class that helps in constructing debug metadata for an LLVM IR file. It corresponds 1:1 similarly to `IRBuilder` and LLVM IR, but with nicer names. Using it does require that you be more familiar with DWARF terminology than you needed to be with `IRBuilder` and `Instruction` names, but if you read through the general documentation on the [Metadata Format](#) it should be a little more clear. We'll be using this class to construct all of our IR level descriptions. Construction for it takes a module so we need to construct it shortly after we construct our module. We've left it as a global static variable to make it a bit easier to use.

Next we’re going to create a small container to cache some of our frequent data. The first will be our compile unit, but we’ll also write a bit of code for our one type since we won’t have to worry about multiple typed expressions:

```
static DIBuilder *DBuilder;

struct DebugInfo {
    DIBuilder *TheCU;
    DIBuilder *DbgTy;

    DIBuilder *getDoubleTy();
} KSDbgInfo;

DIBuilder *DebugInfo::getDoubleTy() {
    if (DbgTy)
        return DbgTy;

    DbgTy = DBuilder->createBasicType("double", 64, 64, dwarf::DW_ATE_float);
    return DbgTy;
}
```

And then later on in main when we’re constructing our module:

```
DBuilder = new DIBuilder(*TheModule);

KSDbgInfo.TheCU = DBuilder->createCompileUnit(
    dwarf::DW_LANG_C, "fib.ks", ".", "Kaleidoscope Compiler", 0, "", 0);
```

There are a couple of things to note here. First, while we’re producing a compile unit for a language called Kaleidoscope we used the language constant for C. This is because a debugger wouldn’t necessarily understand the calling conventions or default ABI for a language it doesn’t recognize and we follow the C ABI in our LLVM code generation so it’s the closest thing to accurate. This ensures we can actually call functions from the debugger and have them execute. Secondly, you’ll see the “fib.ks” in the call to `createCompileUnit`. This is a default hard coded value since we’re using shell redirection to put our source into the Kaleidoscope compiler. In a usual front end you’d have an input file name and it would go there.

One last thing as part of emitting debug information via DIBuilder is that we need to “finalize” the debug information. The reasons are part of the underlying API for DIBuilder, but make sure you do this near the end of main:

```
DBuilder->finalize();
```

before you dump out the module.

Functions

Now that we have our `Compile Unit` and our source locations, we can add function definitions to the debug info. So in `PrototypeAST::codegen()` we add a few lines of code to describe a context for our subprogram, in this case the “File”, and the actual definition of the function itself.

So the context:

```
DIFile *Unit = DBuilder->createFile(KSDbgInfo.TheCU.getFilename(),
                                   KSDbgInfo.TheCU.getDirectory());
```

giving us an `DIFile` and asking the `Compile Unit` we created above for the directory and filename where we are currently. Then, for now, we use some source locations of 0 (since our AST doesn’t currently have source location information) and construct our function definition:


```
DIScope *FContext = Unit;
unsigned LineNo = 0;
unsigned ScopeLine = 0;
DISubprogram *SP = DBuilder->createFunction(
    FContext, P.getName(), StringRef(), Unit, LineNo,
    CreateFunctionType(TheFunction->arg_size(), Unit),
    false /* internal linkage */, true /* definition */, ScopeLine,
    DINode::FlagPrototyped, false);
TheFunction->setSubprogram(SP);
```

and we now have an DISubprogram that contains a reference to all of our metadata for the function.

Source Locations

The most important thing for debug information is accurate source location - this makes it possible to map your source code back. We have a problem though, Kaleidoscope really doesn't have any source location information in the lexer or parser so we'll need to add it.

```
struct SourceLocation {
    int Line;
    int Col;
};
static SourceLocation CurLoc;
static SourceLocation LexLoc = {1, 0};

static int advance() {
    int LastChar = getchar();

    if (LastChar == '\n' || LastChar == '\r') {
        LexLoc.Line++;
        LexLoc.Col = 0;
    } else
        LexLoc.Col++;
    return LastChar;
}
```

In this set of code we've added some functionality on how to keep track of the line and column of the "source file". As we lex every token we set our current "lexical location" to the assorted line and column for the beginning of the token. We do this by overriding all of the previous calls to `getchar()` with our new `advance()` that keeps track of the information and then we have added to all of our AST classes a source location:

```
class ExprAST {
    SourceLocation Loc;

public:
    ExprAST(SourceLocation Loc = CurLoc) : Loc(Loc) {}
    virtual ~ExprAST() {}
    virtual Value* codegen() = 0;
    int getLine() const { return Loc.Line; }
    int getCol() const { return Loc.Col; }
    virtual raw_ostream &dump(raw_ostream &out, int ind) {
        return out << ':' << getLine() << ':' << getCol() << '\n';
    }
}
```

that we pass down through when we create a new expression:

```
LHS = llvm::make_unique<BinaryExprAST>(BinLoc, BinOp, std::move(LHS),
                                       std::move(RHS));
```

giving us locations for each of our expressions and variables.

To make sure that every instruction gets proper source location information, we have to tell Builder whenever we're at a new source location. We use a small helper function for this:

```
void DebugInfo::emitLocation(ExprAST *AST) {
    DIScope *Scope;
    if (LexicalBlocks.empty())
        Scope = TheCU;
    else
        Scope = LexicalBlocks.back();
    Builder.SetCurrentDebugLocation(
        DebugLoc::get(AST->getLine(), AST->getCol(), Scope));
}
```

This both tells the main IRBuilder where we are, but also what scope we're in. The scope can either be on compile-unit level or be the nearest enclosing lexical block like the current function. To represent this we create a stack of scopes:

```
std::vector<DIScope *> LexicalBlocks;
```

and push the scope (function) to the top of the stack when we start generating the code for each function:

```
KSDBGInfo.LexicalBlocks.push_back(SP);
```

Also, we may not forget to pop the scope back off of the scope stack at the end of the code generation for the function:

```
// Pop off the lexical block for the function since we added it
// unconditionally.
KSDBGInfo.LexicalBlocks.pop_back();
```

Then we make sure to emit the location every time we start to generate code for a new AST object:

```
KSDBGInfo.emitLocation(this);
```

Variables

Now that we have functions, we need to be able to print out the variables we have in scope. Let's get our function arguments set up so we can get decent backtraces and see how our functions are being called. It isn't a lot of code, and we generally handle it when we're creating the argument allocas in `FunctionAST::codegen`.

```
// Record the function arguments in the NamedValues map.
NamedValues.clear();
unsigned ArgIdx = 0;
for (auto &Arg : TheFunction->args()) {
    // Create an alloca for this variable.
    AllocaInst *Alloca = CreateEntryBlockAlloca(TheFunction, Arg.getName());

    // Create a debug descriptor for the variable.
    DILocalVariable *D = DBuilder->createParameterVariable(
        SP, Arg.getName(), ++ArgIdx, Unit, LineNo, KSDBGInfo.getDoubleTy(),
        true);
```

```

DBuilder->insertDeclare(Alloca, D, DBuilder->createExpression(),
                        DebugLoc::get(LineNo, 0, SP),
                        Builder.GetInsertBlock());

// Store the initial value into the alloca.
Builder.CreateStore(&Arg, Alloca);

// Add arguments to variable symbol table.
NamedValues[Arg.getName()] = Alloca;
}

```

Here we're first creating the variable, giving it the scope (SP), the name, source location, type, and since it's an argument, the argument index. Next, we create an `lvm.dbg.declare` call to indicate at the IR level that we've got a variable in an alloca (and it gives a starting location for the variable), and setting a source location for the beginning of the scope on the declare.

One interesting thing to note at this point is that various debuggers have assumptions based on how code and debug information was generated for them in the past. In this case we need to do a little bit of a hack to avoid generating line information for the function prologue so that the debugger knows to skip over those instructions when setting a breakpoint. So in `FunctionAST::CodeGen` we add some more lines:

```

// Unset the location for the prologue emission (leading instructions with no
// location in a function are considered part of the prologue and the debugger
// will run past them when breaking on a function)
KSDbgInfo.emitLocation(nullptr);

```

and then emit a new location when we actually start generating code for the body of the function:

```

KSDbgInfo.emitLocation(Body.get());

```

With this we have enough debug information to set breakpoints in functions, print out argument variables, and call functions. Not too bad for just a few simple lines of code!

Full Code Listing

Here is the complete code listing for our running example, enhanced with debug information. To build this example, use:

```

# Compile
clang++ -g toy.cpp `llvm-config --cxxflags --ldflags --system-libs --libs core mcjit_
↪native` -O3 -o toy
# Run
./toy

```

Here is the code:

```

#include "llvm/ADT/STLExtras.h"
#include "llvm/Analysis/BasicAliasAnalysis.h"
#include "llvm/Analysis/Passes.h"
#include "llvm/IR/DIBuilder.h"
#include "llvm/IR/IRBuilder.h"
#include "llvm/IR/LLVMContext.h"
#include "llvm/IR/LegacyPassManager.h"
#include "llvm/IR/Module.h"
#include "llvm/IR/Verifier.h"
#include "llvm/Support/TargetSelect.h"

```

```
#include "llvm/Transforms/Scalar.h"
#include <cctype>
#include <cstdio>
#include <map>
#include <string>
#include <vector>
#include "../include/KaleidoscopeJIT.h"

using namespace llvm;
using namespace llvm::orc;

//=====//
// Lexer
//=====//

// The lexer returns tokens [0-255] if it is an unknown character, otherwise one
// of these for known things.
enum Token {
    tok_eof = -1,

    // commands
    tok_def = -2,
    tokExtern = -3,

    // primary
    tok_identifier = -4,
    tok_number = -5,

    // control
    tok_if = -6,
    tok_then = -7,
    tok_else = -8,
    tok_for = -9,
    tok_in = -10,

    // operators
    tok_binary = -11,
    tok_unary = -12,

    // var definition
    tok_var = -13
};

std::string getTokName(int Tok) {
    switch (Tok) {
    case tok_eof:
        return "eof";
    case tok_def:
        return "def";
    case tokExtern:
        return "extern";
    case tok_identifier:
        return "identifier";
    case tok_number:
        return "number";
    case tok_if:
        return "if";
    case tok_then:
        return "then";
    case tok_else:
        return "else";
    case tok_for:
        return "for";
    case tok_in:
        return "in";
    case tok_binary:
        return "binary";
    case tok_unary:
        return "unary";
    case tok_var:
        return "var";
    }
}
```

```

    return "then";
case tok_else:
    return "else";
case tok_for:
    return "for";
case tok_in:
    return "in";
case tok_binary:
    return "binary";
case tok_unary:
    return "unary";
case tok_var:
    return "var";
}
return std::string(1, (char)Tok);
}

namespace {
class PrototypeAST;
class ExprAST;
}

static LLVMContext TheContext;
static IRBuilder<> Builder(TheContext);
struct DebugInfo {
    DICompileUnit *TheCU;
    DIType *DblTy;
    std::vector<DIScope *> LexicalBlocks;

    void emitLocation(ExprAST *AST);
    DIType *getDoubleTy();
} KSDbgInfo;

struct SourceLocation {
    int Line;
    int Col;
};

static SourceLocation CurLoc;
static SourceLocation LexLoc = {1, 0};

static int advance() {
    int LastChar = getchar();

    if (LastChar == '\n' || LastChar == '\r') {
        LexLoc.Line++;
        LexLoc.Col = 0;
    } else
        LexLoc.Col++;
    return LastChar;
}

static std::string IdentifierStr; // Filled in if tok_identifier
static double NumVal;           // Filled in if tok_number

/// gettok - Return the next token from standard input.
static int gettok() {
    static int LastChar = ' ';

    // Skip any whitespace.

```

```
while (isspace>LastChar))
    LastChar = advance();

CurLoc = LexLoc;

if (isalpha>LastChar)) { // identifier: [a-zA-Z][a-zA-Z0-9]*
    IdentifierStr = LastChar;
    while (isalnum((LastChar = advance()))))
        IdentifierStr += LastChar;

    if (IdentifierStr == "def")
        return tok_def;
    if (IdentifierStr == "extern")
        return tok_extern;
    if (IdentifierStr == "if")
        return tok_if;
    if (IdentifierStr == "then")
        return tok_then;
    if (IdentifierStr == "else")
        return tok_else;
    if (IdentifierStr == "for")
        return tok_for;
    if (IdentifierStr == "in")
        return tok_in;
    if (IdentifierStr == "binary")
        return tok_binary;
    if (IdentifierStr == "unary")
        return tok_unary;
    if (IdentifierStr == "var")
        return tok_var;
    return tok_identifier;
}

if (isdigit>LastChar) || LastChar == '.') { // Number: [0-9.]+
    std::string NumStr;
    do {
        NumStr += LastChar;
        LastChar = advance();
    } while (isdigit>LastChar) || LastChar == '.');

    NumVal = strtod(NumStr.c_str(), nullptr);
    return tok_number;
}

if (LastChar == '#') {
    // Comment until end of line.
    do
        LastChar = advance();
    while (LastChar != EOF && LastChar != '\n' && LastChar != '\r');

    if (LastChar != EOF)
        return gettok();
}

// Check for end of file. Don't eat the EOF.
if (LastChar == EOF)
    return tok_eof;
```

```

    // Otherwise, just return the character as its ascii value.
    int ThisChar = LastChar;
    LastChar = advance();
    return ThisChar;
}

//===-----//
// Abstract Syntax Tree (aka Parse Tree)
//===-----//
namespace {

raw_ostream &indent(raw_ostream &O, int size) {
    return O << std::string(size, ' ');
}

/// ExprAST - Base class for all expression nodes.
class ExprAST {
    SourceLocation Loc;

public:
    ExprAST(SourceLocation Loc = CurLoc) : Loc(Loc) {}
    virtual ~ExprAST() {}
    virtual Value *codegen() = 0;
    int getLine() const { return Loc.Line; }
    int getCol() const { return Loc.Col; }
    virtual raw_ostream &dump(raw_ostream &out, int ind) {
        return out << ':' << getLine() << ':' << getCol() << '\n';
    }
};

/// NumberExprAST - Expression class for numeric literals like "1.0".
class NumberExprAST : public ExprAST {
    double Val;

public:
    NumberExprAST(double Val) : Val(Val) {}
    raw_ostream &dump(raw_ostream &out, int ind) override {
        return ExprAST::dump(out << Val, ind);
    }
    Value *codegen() override;
};

/// VariableExprAST - Expression class for referencing a variable, like "a".
class VariableExprAST : public ExprAST {
    std::string Name;

public:
    VariableExprAST(SourceLocation Loc, const std::string &Name)
        : ExprAST(Loc), Name(Name) {}
    const std::string &getName() const { return Name; }
    Value *codegen() override;
    raw_ostream &dump(raw_ostream &out, int ind) override {
        return ExprAST::dump(out << Name, ind);
    }
};

/// UnaryExprAST - Expression class for a unary operator.
class UnaryExprAST : public ExprAST {

```

```

    char Opcode;
    std::unique_ptr<ExprAST> Operand;

public:
    UnaryExprAST(char Opcode, std::unique_ptr<ExprAST> Operand)
        : Opcode(Opcode), Operand(std::move(Operand)) {}
    Value *codegen() override;
    raw_ostream &dump(raw_ostream &out, int ind) override {
        ExprAST::dump(out << "unary" << Opcode, ind);
        Operand->dump(out, ind + 1);
        return out;
    }
};

/// BinaryExprAST - Expression class for a binary operator.
class BinaryExprAST : public ExprAST {
    char Op;
    std::unique_ptr<ExprAST> LHS, RHS;

public:
    BinaryExprAST(SourceLocation Loc, char Op, std::unique_ptr<ExprAST> LHS,
                  std::unique_ptr<ExprAST> RHS)
        : ExprAST(Loc), Op(Op), LHS(std::move(LHS)), RHS(std::move(RHS)) {}
    Value *codegen() override;
    raw_ostream &dump(raw_ostream &out, int ind) override {
        ExprAST::dump(out << "binary" << Op, ind);
        LHS->dump(indent(out, ind) << "LHS:", ind + 1);
        RHS->dump(indent(out, ind) << "RHS:", ind + 1);
        return out;
    }
};

/// CallExprAST - Expression class for function calls.
class CallExprAST : public ExprAST {
    std::string Callee;
    std::vector<std::unique_ptr<ExprAST>> Args;

public:
    CallExprAST(SourceLocation Loc, const std::string &Callee,
                std::vector<std::unique_ptr<ExprAST>> Args)
        : ExprAST(Loc), Callee(Callee), Args(std::move(Args)) {}
    Value *codegen() override;
    raw_ostream &dump(raw_ostream &out, int ind) override {
        ExprAST::dump(out << "call " << Callee, ind);
        for (const auto &Arg : Args)
            Arg->dump(indent(out, ind + 1), ind + 1);
        return out;
    }
};

/// IfExprAST - Expression class for if/then/else.
class IfExprAST : public ExprAST {
    std::unique_ptr<ExprAST> Cond, Then, Else;

public:
    IfExprAST(SourceLocation Loc, std::unique_ptr<ExprAST> Cond,
              std::unique_ptr<ExprAST> Then, std::unique_ptr<ExprAST> Else)
        : ExprAST(Loc), Cond(std::move(Cond)), Then(std::move(Then)),

```



```

        Else(std::move(Else)) {}
Value *codegen() override;
raw_ostream &dump(raw_ostream &out, int ind) override {
    ExprAST::dump(out << "if", ind);
    Cond->dump(indent(out, ind) << "Cond:", ind + 1);
    Then->dump(indent(out, ind) << "Then:", ind + 1);
    Else->dump(indent(out, ind) << "Else:", ind + 1);
    return out;
}
};

/// ForExprAST - Expression class for for/in.
class ForExprAST : public ExprAST {
    std::string VarName;
    std::unique_ptr<ExprAST> Start, End, Step, Body;

public:
    ForExprAST(const std::string &VarName, std::unique_ptr<ExprAST> Start,
               std::unique_ptr<ExprAST> End, std::unique_ptr<ExprAST> Step,
               std::unique_ptr<ExprAST> Body)
        : VarName(VarName), Start(std::move(Start)), End(std::move(End)),
          Step(std::move(Step)), Body(std::move(Body)) {}
Value *codegen() override;
raw_ostream &dump(raw_ostream &out, int ind) override {
    ExprAST::dump(out << "for", ind);
    Start->dump(indent(out, ind) << "Cond:", ind + 1);
    End->dump(indent(out, ind) << "End:", ind + 1);
    Step->dump(indent(out, ind) << "Step:", ind + 1);
    Body->dump(indent(out, ind) << "Body:", ind + 1);
    return out;
}
};

/// VarExprAST - Expression class for var/in
class VarExprAST : public ExprAST {
    std::vector<std::pair<std::string, std::unique_ptr<ExprAST>>> VarNames;
    std::unique_ptr<ExprAST> Body;

public:
    VarExprAST(
        std::vector<std::pair<std::string, std::unique_ptr<ExprAST>>> VarNames,
        std::unique_ptr<ExprAST> Body)
        : VarNames(std::move(VarNames)), Body(std::move(Body)) {}
Value *codegen() override;
raw_ostream &dump(raw_ostream &out, int ind) override {
    ExprAST::dump(out << "var", ind);
    for (const auto &NamedVar : VarNames)
        NamedVar.second->dump(indent(out, ind) << NamedVar.first << ":", ind + 1);
    Body->dump(indent(out, ind) << "Body:", ind + 1);
    return out;
}
};

/// PrototypeAST - This class represents the "prototype" for a function,
/// which captures its name, and its argument names (thus implicitly the number
/// of arguments the function takes), as well as if it is an operator.
class PrototypeAST {
    std::string Name;

```

```

std::vector<std::string> Args;
bool IsOperator;
unsigned Precedence; // Precedence if a binary op.
int Line;

public:
    PrototypeAST(SourceLocation Loc, const std::string &Name,
                std::vector<std::string> Args, bool IsOperator = false,
                unsigned Prec = 0)
        : Name(Name), Args(std::move(Args)), IsOperator(IsOperator),
          Precedence(Prec), Line(Loc.Line) {}
    Function *codegen();
    const std::string &getName() const { return Name; }

    bool isUnaryOp() const { return IsOperator && Args.size() == 1; }
    bool isBinaryOp() const { return IsOperator && Args.size() == 2; }

    char getOperatorName() const {
        assert(isUnaryOp() || isBinaryOp());
        return Name[Name.size() - 1];
    }

    unsigned getBinaryPrecedence() const { return Precedence; }
    int getLine() const { return Line; }
};

/// FunctionAST - This class represents a function definition itself.
class FunctionAST {
    std::unique_ptr<PrototypeAST> Proto;
    std::unique_ptr<ExprAST> Body;

public:
    FunctionAST(std::unique_ptr<PrototypeAST> Proto,
                std::unique_ptr<ExprAST> Body)
        : Proto(std::move(Proto)), Body(std::move(Body)) {}
    Function *codegen();
    raw_ostream &dump(raw_ostream &out, int ind) {
        indent(out, ind) << "FunctionAST\n";
        ++ind;
        indent(out, ind) << "Body:";
        return Body ? Body->dump(out, ind) : out << "null\n";
    }
};
} // end anonymous namespace

//===-----
// Parser
//===-----

/// CurTok/getNextToken - Provide a simple token buffer. CurTok is the current
/// token the parser is looking at. getNextToken reads another token from the
/// lexer and updates CurTok with its results.
static int CurTok;
static int getNextToken() { return CurTok = gettok(); }

/// BinopPrecedence - This holds the precedence for each binary operator that is
/// defined.
static std::map<char, int> BinopPrecedence;

```

```

/// GetTokPrecedence - Get the precedence of the pending binary operator token.
static int GetTokPrecedence() {
    if (!isascii(CurTok))
        return -1;

    // Make sure it's a declared binop.
    int TokPrec = BinopPrecedence[CurTok];
    if (TokPrec <= 0)
        return -1;
    return TokPrec;
}

/// LogError* - These are little helper functions for error handling.
std::unique_ptr<ExprAST> LogError(const char *Str) {
    fprintf(stderr, "Error: %s\n", Str);
    return nullptr;
}

std::unique_ptr<PrototypeAST> LogErrorP(const char *Str) {
    LogError(Str);
    return nullptr;
}

static std::unique_ptr<ExprAST> ParseExpression();

/// numberexpr ::= number
static std::unique_ptr<ExprAST> ParseNumberExpr() {
    auto Result = llvm::make_unique<NumberExprAST>(NumVal);
    getNextToken(); // consume the number
    return std::move(Result);
}

/// parenexpr ::= '(' expression ')'
static std::unique_ptr<ExprAST> ParseParenExpr() {
    getNextToken(); // eat '('
    auto V = ParseExpression();
    if (!V)
        return nullptr;

    if (CurTok != ')')
        return LogError("expected ')'");
    getNextToken(); // eat ')'
    return V;
}

/// identifierexpr
/// ::= identifier
/// ::= identifier '(' expression* ')'
static std::unique_ptr<ExprAST> ParseIdentifierExpr() {
    std::string IdName = IdentifierStr;

    SourceLocation LitLoc = CurLoc;

    getNextToken(); // eat identifier.

    if (CurTok != '(') // Simple variable ref.
        return llvm::make_unique<VariableExprAST>(LitLoc, IdName);

```

```
// Call.
getNextToken(); // eat (
std::vector<std::unique_ptr<ExprAST>> Args;
if (CurTok != ')') {
    while (1) {
        if (auto Arg = ParseExpression())
            Args.push_back(std::move(Arg));
        else
            return nullptr;

        if (CurTok == ')')
            break;

        if (CurTok != ',')
            return LogError("Expected ')' or ',' in argument list");
        getNextToken();
    }
}

// Eat the ')'.
getNextToken();

return llvm::make_unique<CallExprAST>(LitLoc, IdName, std::move(Args));
}

/// ifexpr ::= 'if' expression 'then' expression 'else' expression
static std::unique_ptr<ExprAST> ParseIfExpr() {
    SourceLocation IfLoc = CurLoc;

    getNextToken(); // eat the if.

    // condition.
    auto Cond = ParseExpression();
    if (!Cond)
        return nullptr;

    if (CurTok != tok_then)
        return LogError("expected then");
    getNextToken(); // eat the then

    auto Then = ParseExpression();
    if (!Then)
        return nullptr;

    if (CurTok != tok_else)
        return LogError("expected else");

    getNextToken();

    auto Else = ParseExpression();
    if (!Else)
        return nullptr;

    return llvm::make_unique<IfExprAST>(IfLoc, std::move(Cond), std::move(Then),
                                         std::move(Else));
}
```

```

/// forexpr ::= 'for' identifier '=' expr ',' expr (',' expr)? 'in' expression
static std::unique_ptr<ExprAST> ParseForExpr() {
    getNextToken(); // eat the for.

    if (CurTok != tok_identifier)
        return LogError("expected identifier after for");

    std::string IdName = IdentifierStr;
    getNextToken(); // eat identifier.

    if (CurTok != '=')
        return LogError("expected '=' after for");
    getNextToken(); // eat '='.

    auto Start = ParseExpression();
    if (!Start)
        return nullptr;
    if (CurTok != ',')
        return LogError("expected ',' after for start value");
    getNextToken();

    auto End = ParseExpression();
    if (!End)
        return nullptr;

    // The step value is optional.
    std::unique_ptr<ExprAST> Step;
    if (CurTok == ',') {
        getNextToken();
        Step = ParseExpression();
        if (!Step)
            return nullptr;
    }

    if (CurTok != tok_in)
        return LogError("expected 'in' after for");
    getNextToken(); // eat 'in'.

    auto Body = ParseExpression();
    if (!Body)
        return nullptr;

    return llvm::make_unique<ForExprAST>(IdName, std::move(Start), std::move(End),
                                         std::move(Step), std::move(Body));
}

/// varexpr ::= 'var' identifier ('=' expression)?
///           (',' identifier ('=' expression)?)* 'in' expression
static std::unique_ptr<ExprAST> ParseVarExpr() {
    getNextToken(); // eat the var.

    std::vector<std::pair<std::string, std::unique_ptr<ExprAST>>> VarNames;

    // At least one variable name is required.
    if (CurTok != tok_identifier)
        return LogError("expected identifier after var");

    while (1) {

```

```

std::string Name = IdentifierStr;
getNextToken(); // eat identifier.

// Read the optional initializer.
std::unique_ptr<ExprAST> Init = nullptr;
if (CurTok == '=') {
    getNextToken(); // eat the '='.

    Init = ParseExpression();
    if (!Init)
        return nullptr;
}

VarNames.push_back(std::make_pair(Name, std::move(Init)));

// End of var list, exit loop.
if (CurTok != ',')
    break;
getNextToken(); // eat the ','.

if (CurTok != tok_identifier)
    return LogError("expected identifier list after var");
}

// At this point, we have to have 'in'.
if (CurTok != tok_in)
    return LogError("expected 'in' keyword after 'var'");
getNextToken(); // eat 'in'.

auto Body = ParseExpression();
if (!Body)
    return nullptr;

return llvm::make_unique<VarExprAST>(std::move(VarNames), std::move(Body));
}

/// primary
/// ::= identifierexpr
/// ::= numberexpr
/// ::= parenexpr
/// ::= ifexpr
/// ::= forexpr
/// ::= varexpr
static std::unique_ptr<ExprAST> ParsePrimary() {
    switch (CurTok) {
    default:
        return LogError("unknown token when expecting an expression");
    case tok_identifier:
        return ParseIdentifierExpr();
    case tok_number:
        return ParseNumberExpr();
    case '(':
        return ParseParenExpr();
    case tok_if:
        return ParseIfExpr();
    case tok_for:
        return ParseForExpr();
    case tok_var:

```

```

    return ParseVarExpr();
}
}

/// unary
///   ::= primary
///   ::= '!' unary
static std::unique_ptr<ExprAST> ParseUnary() {
    // If the current token is not an operator, it must be a primary expr.
    if (!isascii(CurTok) || CurTok == '(' || CurTok == ',')
        return ParsePrimary();

    // If this is a unary operator, read it.
    int Opc = CurTok;
    getNextToken();
    if (auto Operand = ParseUnary())
        return llvm::make_unique<UnaryExprAST>(Opc, std::move(Operand));
    return nullptr;
}

/// binoprhs
///   ::= ('+' unary)*
static std::unique_ptr<ExprAST> ParseBinOpRHS(int ExprPrec,
                                              std::unique_ptr<ExprAST> LHS) {
    // If this is a binop, find its precedence.
    while (1) {
        int TokPrec = GetTokPrecedence();

        // If this is a binop that binds at least as tightly as the current binop,
        // consume it, otherwise we are done.
        if (TokPrec < ExprPrec)
            return LHS;

        // Okay, we know this is a binop.
        int BinOp = CurTok;
        SourceLocation BinLoc = CurLoc;
        getNextToken(); // eat binop

        // Parse the unary expression after the binary operator.
        auto RHS = ParseUnary();
        if (!RHS)
            return nullptr;

        // If BinOp binds less tightly with RHS than the operator after RHS, let
        // the pending operator take RHS as its LHS.
        int NextPrec = GetTokPrecedence();
        if (TokPrec < NextPrec) {
            RHS = ParseBinOpRHS(TokPrec + 1, std::move(RHS));
            if (!RHS)
                return nullptr;
        }

        // Merge LHS/RHS.
        LHS = llvm::make_unique<BinaryExprAST>(BinLoc, BinOp, std::move(LHS),
                                              std::move(RHS));
    }
}
}

```

```

/// expression
/// ::= unary binoprhs
///
static std::unique_ptr<ExprAST> ParseExpression() {
    auto LHS = ParseUnary();
    if (!LHS)
        return nullptr;

    return ParseBinOpRHS(0, std::move(LHS));
}

/// prototype
/// ::= id '(' id* ')'
/// ::= binary LETTER number? (id, id)
/// ::= unary LETTER (id)
static std::unique_ptr<PrototypeAST> ParsePrototype() {
    std::string FnName;

    SourceLocation FnLoc = CurLoc;

    unsigned Kind = 0; // 0 = identifier, 1 = unary, 2 = binary.
    unsigned BinaryPrecedence = 30;

    switch (CurTok) {
    default:
        return LogErrorP("Expected function name in prototype");
    case tok_identifier:
        FnName = IdentifierStr;
        Kind = 0;
        getNextToken();
        break;
    case tok_unary:
        getNextToken();
        if (!isascii(CurTok))
            return LogErrorP("Expected unary operator");
        FnName = "unary";
        FnName += (char)CurTok;
        Kind = 1;
        getNextToken();
        break;
    case tok_binary:
        getNextToken();
        if (!isascii(CurTok))
            return LogErrorP("Expected binary operator");
        FnName = "binary";
        FnName += (char)CurTok;
        Kind = 2;
        getNextToken();

        // Read the precedence if present.
        if (CurTok == tok_number) {
            if (NumVal < 1 || NumVal > 100)
                return LogErrorP("Invalid precedence: must be 1..100");
            BinaryPrecedence = (unsigned)NumVal;
            getNextToken();
        }
        break;
    }
}

```



```

if (CurTok != '(')
    return LogErrorP("Expected '(' in prototype");

std::vector<std::string> ArgNames;
while (getNextToken() == tok_identifier)
    ArgNames.push_back(IdentifierStr);
if (CurTok != ')')
    return LogErrorP("Expected ')' in prototype");

// success.
getNextToken(); // eat ')'.

// Verify right number of names for operator.
if (Kind && ArgNames.size() != Kind)
    return LogErrorP("Invalid number of operands for operator");

return llvm::make_unique<PrototypeAST>(FnLoc, FnName, ArgNames, Kind != 0,
                                     BinaryPrecedence);
}

/// definition ::= 'def' prototype expression
static std::unique_ptr<FunctionAST> ParseDefinition() {
    getNextToken(); // eat def.
    auto Proto = ParsePrototype();
    if (!Proto)
        return nullptr;

    if (auto E = ParseExpression())
        return llvm::make_unique<FunctionAST>(std::move(Proto), std::move(E));
    return nullptr;
}

/// toplevelexpr ::= expression
static std::unique_ptr<FunctionAST> ParseTopLevelExpr() {
    SourceLocation FnLoc = CurLoc;
    if (auto E = ParseExpression()) {
        // Make an anonymous proto.
        auto Proto = llvm::make_unique<PrototypeAST>(FnLoc, "__anon_expr",
                                                    std::vector<std::string>());
        return llvm::make_unique<FunctionAST>(std::move(Proto), std::move(E));
    }
    return nullptr;
}

/// external ::= 'extern' prototype
static std::unique_ptr<PrototypeAST> ParseExtern() {
    getNextToken(); // eat extern.
    return ParsePrototype();
}

//=====//
// Debug Info Support
//=====//

static std::unique_ptr<DIBuilder> DBuilder;

DIType *DebugInfo::getDoubleTy() {

```

```

    if (DblTy)
        return DblTy;

    DblTy = DBuilder->createBasicType("double", 64, 64, dwarf::DW_ATE_float);
    return DblTy;
}

void DebugInfo::emitLocation(ExprAST *AST) {
    if (!AST)
        return Builder.SetCurrentDebugLocation(DebugLoc());
    DIScope *Scope;
    if (LexicalBlocks.empty())
        Scope = TheCU;
    else
        Scope = LexicalBlocks.back();
    Builder.SetCurrentDebugLocation(
        DebugLoc::get(AST->getLine(), AST->getCol(), Scope));
}

static DISubroutineType *CreateFunctionType(unsigned NumArgs, DIFile *Unit) {
    SmallVector<Metadata *, 8> EltTys;
    DIType *DblTy = KSDBGInfo.getDoubleTy();

    // Add the result type.
    EltTys.push_back(DblTy);

    for (unsigned i = 0, e = NumArgs; i != e; ++i)
        EltTys.push_back(DblTy);

    return DBuilder->createSubroutineType(DBuilder->getOrCreateTypeArray(EltTys));
}

//===-----//
// Code Generation
//===-----//

static std::unique_ptr<Module> TheModule;
static std::map<std::string, AllocaInst *> NamedValues;
static std::unique_ptr<KaleidoscopeJIT> TheJIT;
static std::map<std::string, std::unique_ptr<PrototypeAST>> FunctionProtos;

Value *LogErrorV(const char *Str) {
    LogError(Str);
    return nullptr;
}

Function *getFunction(std::string Name) {
    // First, see if the function has already been added to the current module.
    if (auto *F = TheModule->getFunction(Name))
        return F;

    // If not, check whether we can codegen the declaration from some existing
    // prototype.
    auto FI = FunctionProtos.find(Name);
    if (FI != FunctionProtos.end())
        return FI->second->codegen();

    // If no existing prototype exists, return null.

```

```

    return nullptr;
}

/// CreateEntryBlockAlloca - Create an alloca instruction in the entry block of
/// the function. This is used for mutable variables etc.
static AllocaInst *CreateEntryBlockAlloca(Function *TheFunction,
                                           const std::string &VarName) {
    IRBuilder<> TmpB(&TheFunction->getEntryBlock(),
                  TheFunction->getEntryBlock().begin());
    return TmpB.CreateAlloca(Type::getDoubleTy(TheContext), nullptr,
                             VarName.c_str());
}

Value *NumberExprAST::codegen() {
    KSDbgInfo.emitLocation(this);
    return ConstantFP::get(TheContext, APFloat(Val));
}

Value *VariableExprAST::codegen() {
    // Look this variable up in the function.
    Value *V = NamedValues[Name];
    if (!V)
        return LogErrorV("Unknown variable name");

    KSDbgInfo.emitLocation(this);
    // Load the value.
    return Builder.CreateLoad(V, Name.c_str());
}

Value *UnaryExprAST::codegen() {
    Value *OperandV = Operand->codegen();
    if (!OperandV)
        return nullptr;

    Function *F = getFunction(std::string("unary") + Opcode);
    if (!F)
        return LogErrorV("Unknown unary operator");

    KSDbgInfo.emitLocation(this);
    return Builder.CreateCall(F, OperandV, "unop");
}

Value *BinaryExprAST::codegen() {
    KSDbgInfo.emitLocation(this);

    // Special case '=' because we don't want to emit the LHS as an expression.
    if (Op == '=') {
        // Assignment requires the LHS to be an identifier.
        // This assume we're building without RTTI because LLVM builds that way by
        // default. If you build LLVM with RTTI this can be changed to a
        // dynamic_cast for automatic error checking.
        VariableExprAST *LHSE = static_cast<VariableExprAST *>(LHS.get());
        if (!LHSE)
            return LogErrorV("destination of '=' must be a variable");
        // Codegen the RHS.
        Value *Val = RHS->codegen();
        if (!Val)
            return nullptr;
    }
}

```

```

    // Look up the name.
    Value *Variable = NamedValues[LHSE->getName()];
    if (!Variable)
        return LogErrorV("Unknown variable name");

    Builder.CreateStore(Val, Variable);
    return Val;
}

Value *L = LHS->codegen();
Value *R = RHS->codegen();
if (!L || !R)
    return nullptr;

switch (Op) {
case '+':
    return Builder.CreateFAdd(L, R, "addtmp");
case '-':
    return Builder.CreateFSub(L, R, "subtmp");
case '*':
    return Builder.CreateFMul(L, R, "multmp");
case '<':
    L = Builder.CreateFCmpULT(L, R, "cmptmp");
    // Convert bool 0/1 to double 0.0 or 1.0
    return Builder.CreateUIToFP(L, Type::getDoubleTy(TheContext), "booltmp");
default:
    break;
}

// If it wasn't a builtin binary operator, it must be a user defined one. Emit
// a call to it.
Function *F = getFunction(std::string("binary") + Op);
assert(F && "binary operator not found!");

Value *Ops[] = {L, R};
return Builder.CreateCall(F, Ops, "binop");
}

Value *CallExprAST::codegen() {
    KSDbgInfo.emitLocation(this);

    // Look up the name in the global module table.
    Function *CalleeF = getFunction(Callee);
    if (!CalleeF)
        return LogErrorV("Unknown function referenced");

    // If argument mismatch error.
    if (CalleeF->arg_size() != Args.size())
        return LogErrorV("Incorrect # arguments passed");

    std::vector<Value *> ArgsV;
    for (unsigned i = 0, e = Args.size(); i != e; ++i) {
        ArgsV.push_back(Args[i]->codegen());
        if (!ArgsV.back())
            return nullptr;
    }
}

```

```

    return Builder.CreateCall(CalleeF, ArgsV, "calltmp");
}

Value *IfExprAST::codegen() {
    KSDbgInfo.emitLocation(this);

    Value *CondV = Cond->codegen();
    if (!CondV)
        return nullptr;

    // Convert condition to a bool by comparing non-equal to 0.0.
    CondV = Builder.CreateFCmpONE(
        CondV, ConstantFP::get(TheContext, APFloat(0.0)), "ifcond");

    Function *TheFunction = Builder.GetInsertBlock()->getParent();

    // Create blocks for the then and else cases. Insert the 'then' block at the
    // end of the function.
    BasicBlock *ThenBB = BasicBlock::Create(TheContext, "then", TheFunction);
    BasicBlock *ElseBB = BasicBlock::Create(TheContext, "else");
    BasicBlock *MergeBB = BasicBlock::Create(TheContext, "ifcont");

    Builder.CreateCondBr(CondV, ThenBB, ElseBB);

    // Emit then value.
    Builder.SetInsertPoint(ThenBB);

    Value *ThenV = Then->codegen();
    if (!ThenV)
        return nullptr;

    Builder.CreateBr(MergeBB);
    // Codegen of 'Then' can change the current block, update ThenBB for the PHI.
    ThenBB = Builder.GetInsertBlock();

    // Emit else block.
    TheFunction->getBasicBlockList().push_back(ElseBB);
    Builder.SetInsertPoint(ElseBB);

    Value *ElseV = Else->codegen();
    if (!ElseV)
        return nullptr;

    Builder.CreateBr(MergeBB);
    // Codegen of 'Else' can change the current block, update ElseBB for the PHI.
    ElseBB = Builder.GetInsertBlock();

    // Emit merge block.
    TheFunction->getBasicBlockList().push_back(MergeBB);
    Builder.SetInsertPoint(MergeBB);
    PHINode *PN = Builder.CreatePHI(Type::getDoubleTy(TheContext), 2, "iftmp");

    PN->addIncoming(ThenV, ThenBB);
    PN->addIncoming(ElseV, ElseBB);
    return PN;
}

// Output for-loop as:

```

```
//  var = alloca double
//  ...
//  start = startexpr
//  store start -> var
//  goto loop
// loop:
//  ...
//  bodyexpr
//  ...
// loopend:
//  step = stepexpr
//  endcond = endexpr
//
//  curvar = load var
//  nextvar = curvar + step
//  store nextvar -> var
//  br endcond, loop, endloop
// outloop:
Value *ForExprAST::codegen() {
    Function *TheFunction = Builder.GetInsertBlock()->getParent();

    // Create an alloca for the variable in the entry block.
    AllocaInst *Alloca = CreateEntryBlockAlloca(TheFunction, VarName);

    KSDbgInfo.emitLocation(this);

    // Emit the start code first, without 'variable' in scope.
    Value *StartVal = Start->codegen();
    if (!StartVal)
        return nullptr;

    // Store the value into the alloca.
    Builder.CreateStore(StartVal, Alloca);

    // Make the new basic block for the loop header, inserting after current
    // block.
    BasicBlock *LoopBB = BasicBlock::Create(TheContext, "loop", TheFunction);

    // Insert an explicit fall through from the current block to the LoopBB.
    Builder.CreateBr(LoopBB);

    // Start insertion in LoopBB.
    Builder.SetInsertPoint(LoopBB);

    // Within the loop, the variable is defined equal to the PHI node. If it
    // shadows an existing variable, we have to restore it, so save it now.
    AllocaInst *OldVal = NamedValues[VarName];
    NamedValues[VarName] = Alloca;

    // Emit the body of the loop. This, like any other expr, can change the
    // current BB. Note that we ignore the value computed by the body, but don't
    // allow an error.
    if (!Body->codegen())
        return nullptr;

    // Emit the step value.
    Value *StepVal = nullptr;
    if (Step) {
```

```

    StepVal = Step->codegen();
    if (!StepVal)
        return nullptr;
} else {
    // If not specified, use 1.0.
    StepVal = ConstantFP::get(TheContext, APFloat(1.0));
}

// Compute the end condition.
Value *EndCond = End->codegen();
if (!EndCond)
    return nullptr;

// Reload, increment, and restore the alloca. This handles the case where
// the body of the loop mutates the variable.
Value *CurVar = Builder.CreateLoad(Alloca, VarName.c_str());
Value *NextVar = Builder.CreateFAdd(CurVar, StepVal, "nextvar");
Builder.CreateStore(NextVar, Alloca);

// Convert condition to a bool by comparing non-equal to 0.0.
EndCond = Builder.CreateFCmpONE(
    EndCond, ConstantFP::get(TheContext, APFloat(0.0)), "loopcond");

// Create the "after loop" block and insert it.
BasicBlock *AfterBB =
    BasicBlock::Create(TheContext, "afterloop", TheFunction);

// Insert the conditional branch into the end of LoopEndBB.
Builder.CreateCondBr(EndCond, LoopBB, AfterBB);

// Any new code will be inserted in AfterBB.
Builder.SetInsertPoint(AfterBB);

// Restore the unshadowed variable.
if (OldVal)
    NamedValues[VarName] = OldVal;
else
    NamedValues.erase(VarName);

// for expr always returns 0.0.
return Constant::getNullValue(Type::getDoubleTy(TheContext));
}

Value *VarExprAST::codegen() {
    std::vector<AllocaInst *> OldBindings;

    Function *TheFunction = Builder.GetInsertBlock()->getParent();

    // Register all variables and emit their initializer.
    for (unsigned i = 0, e = VarNames.size(); i != e; ++i) {
        const std::string &VarName = VarNames[i].first;
        ExprAST *Init = VarNames[i].second.get();

        // Emit the initializer before adding the variable to scope, this prevents
        // the initializer from referencing the variable itself, and permits stuff
        // like this:
        //   var a = 1 in
        //   var a = a in ... # refers to outer 'a'.

```

```

Value *InitVal;
if (Init) {
    InitVal = Init->codegen();
    if (!InitVal)
        return nullptr;
} else { // If not specified, use 0.0.
    InitVal = ConstantFP::get(TheContext, APFloat(0.0));
}

AllocaInst *Alloca = CreateEntryBlockAlloca(TheFunction, VarName);
Builder.CreateStore(InitVal, Alloca);

// Remember the old variable binding so that we can restore the binding when
// we unrecurse.
OldBindings.push_back(NamedValues[VarName]);

// Remember this binding.
NamedValues[VarName] = Alloca;
}

KSDbgInfo.emitLocation(this);

// Codegen the body, now that all vars are in scope.
Value *BodyVal = Body->codegen();
if (!BodyVal)
    return nullptr;

// Pop all our variables from scope.
for (unsigned i = 0, e = VarNames.size(); i != e; ++i)
    NamedValues[VarNames[i].first] = OldBindings[i];

// Return the body computation.
return BodyVal;
}

Function *PrototypeAST::codegen() {
    // Make the function type: double(double, double) etc.
    std::vector<Type*> Doubles(Args.size(), Type::getDoubleTy(TheContext));
    FunctionType *FT =
        FunctionType::get(Type::getDoubleTy(TheContext), Doubles, false);

    Function *F =
        Function::Create(FT, Function::ExternalLinkage, Name, TheModule.get());

    // Set names for all arguments.
    unsigned Idx = 0;
    for (auto &Arg : F->args())
        Arg.setName(Args[Idx++]);

    return F;
}

Function *FunctionAST::codegen() {
    // Transfer ownership of the prototype to the FunctionProtos map, but keep a
    // reference to it for use below.
    auto &P = *Proto;
    FunctionProtos[Proto->getName()] = std::move(Proto);
    Function *TheFunction = getFunction(P.getName());

```



```

if (!TheFunction)
    return nullptr;

// If this is an operator, install it.
if (P.isBinaryOp())
    BinopPrecedence[P.getOperatorName()] = P.getBinaryPrecedence();

// Create a new basic block to start insertion into.
BasicBlock *BB = BasicBlock::Create(TheContext, "entry", TheFunction);
Builder.SetInsertPoint(BB);

// Create a subprogram DIE for this function.
DIFile *Unit = DBuilder->createFile(KSDBGInfo.TheCU->getFilename(),
                                   KSDBGInfo.TheCU->getDirectory());

DIScope *FContext = Unit;
unsigned LineNo = P.getLine();
unsigned ScopeLine = LineNo;
DISubprogram *SP = DBuilder->createFunction(
    FContext, P.getName(), StringRef(), Unit, LineNo,
    CreateFunctionType(TheFunction->arg_size(), Unit),
    false /* internal linkage */, true /* definition */, ScopeLine,
    DINode::FlagPrototyped, false);
TheFunction->setSubprogram(SP);

// Push the current scope.
KSDBGInfo.LexicalBlocks.push_back(SP);

// Unset the location for the prologue emission (leading instructions with no
// location in a function are considered part of the prologue and the debugger
// will run past them when breaking on a function)
KSDBGInfo.emitLocation(nullptr);

// Record the function arguments in the NamedValues map.
NamedValues.clear();
unsigned ArgIdx = 0;
for (auto &Arg : TheFunction->args()) {
    // Create an alloca for this variable.
    AllocaInst *Alloca = CreateEntryBlockAlloca(TheFunction, Arg.getName());

    // Create a debug descriptor for the variable.
    DILocalVariable *D = DBuilder->createParameterVariable(
        SP, Arg.getName(), ++ArgIdx, Unit, LineNo, KSDBGInfo.getDoubleTy(),
        true);

    DBuilder->insertDeclare(Alloca, D, DBuilder->createExpression(),
                          DebugLoc::get(LineNo, 0, SP),
                          Builder.GetInsertBlock());

    // Store the initial value into the alloca.
    Builder.CreateStore(&Arg, Alloca);

    // Add arguments to variable symbol table.
    NamedValues[Arg.getName()] = Alloca;
}

KSDBGInfo.emitLocation(Body.get());

if (Value *RetVal = Body->codegen()) {

```

```

    // Finish off the function.
    Builder.CreateRet(RetVal);

    // Pop off the lexical block for the function.
    KSDBGInfo.LexicalBlocks.pop_back();

    // Validate the generated code, checking for consistency.
    verifyFunction(*TheFunction);

    return TheFunction;
}

// Error reading body, remove function.
TheFunction->eraseFromParent();

if (P.isBinaryOp())
    BinopPrecedence.erase(Proto->getOperatorName());

// Pop off the lexical block for the function since we added it
// unconditionally.
KSDBGInfo.LexicalBlocks.pop_back();

return nullptr;
}

//===-----//
// Top-Level parsing and JIT Driver
//===-----//

static void InitializeModule() {
    // Open a new module.
    TheModule = llvm::make_unique<Module>("my cool jit", TheContext);
    TheModule->setDataLayout(TheJIT->getTargetMachine().createDataLayout());
}

static void HandleDefinition() {
    if (auto FnAST = ParseDefinition()) {
        if (!FnAST->codegen())
            fprintf(stderr, "Error reading function definition:");
    } else {
        // Skip token for error recovery.
        getNextToken();
    }
}

static void HandleExtern() {
    if (auto ProtoAST = ParseExtern()) {
        if (!ProtoAST->codegen())
            fprintf(stderr, "Error reading extern");
        else
            FunctionProtos[ProtoAST->getName()] = std::move(ProtoAST);
    } else {
        // Skip token for error recovery.
        getNextToken();
    }
}

static void HandleTopLevelExpression() {

```

```

// Evaluate a top-level expression into an anonymous function.
if (auto FnAST = ParseTopLevelExpr()) {
    if (!FnAST->codegen()) {
        fprintf(stderr, "Error generating code for top level expr");
    }
} else {
    // Skip token for error recovery.
    getNextToken();
}
}

/// top ::= definition | external | expression | ';'
static void MainLoop() {
    while (1) {
        switch (CurTok) {
            case tok_eof:
                return;
            case ';': // ignore top-level semicolons.
                getNextToken();
                break;
            case tok_def:
                HandleDefinition();
                break;
            case tokExtern:
                HandleExtern();
                break;
            default:
                HandleTopLevelExpression();
                break;
        }
    }
}

//===-----//
// "Library" functions that can be "extern'd" from user code.
//===-----//

#ifdef LLVM_ON_WIN32
#define DLLEXPORT __declspec(dllexport)
#else
#define DLLEXPORT
#endif

// putchar - putchar that takes a double and returns 0.
extern "C" DLLEXPORT double putchar(double X) {
    fputc((char)X, stderr);
    return 0;
}

// printfd - printf that takes a double prints it as "%f\n", returning 0.
extern "C" DLLEXPORT double printfd(double X) {
    fprintf(stderr, "%f\n", X);
    return 0;
}

//===-----//
// Main driver code.
//===-----//

```

```
int main() {
    InitializeNativeTarget();
    InitializeNativeTargetAsmPrinter();
    InitializeNativeTargetAsmParser();

    // Install standard binary operators.
    // 1 is lowest precedence.
    BinopPrecedence['='] = 2;
    BinopPrecedence['<'] = 10;
    BinopPrecedence['+'] = 20;
    BinopPrecedence['-'] = 20;
    BinopPrecedence['*'] = 40; // highest.

    // Prime the first token.
    getNextToken();

    TheJIT = llvm::make_unique<KaleidoscopeJIT>();

    InitializeModule();

    // Add the current debug info version into the module.
    TheModule->addModuleFlag(Module::Warning, "Debug Info Version",
                             DEBUG_METADATA_VERSION);

    // Darwin only supports dwarf2.
    if (Triple(sys::getProcessTriple()).isOSDarwin())
        TheModule->addModuleFlag(llvm::Module::Warning, "Dwarf Version", 2);

    // Construct the DIBuilder, we do this here because we need the module.
    DBuilder = llvm::make_unique<DIBuilder>(*TheModule);

    // Create the compile unit for the module.
    // Currently down as "fib.ks" as a filename since we're redirecting stdin
    // but we'd like actual source locations.
    KSDBGInfo.TheCU = DBuilder->createCompileUnit(
        dwarf::DW_LANG_C, "fib.ks", ".", "Kaleidoscope Compiler", 0, "", 0);

    // Run the main "interpreter loop" now.
    MainLoop();

    // Finalize the debug info.
    DBuilder->finalize();

    // Print out all of the generated code.
    TheModule->print(errs(), nullptr);

    return 0;
}
```

Next: Conclusion and other useful LLVM tidbits

Kaleidoscope: Conclusion and other useful LLVM tidbits

- *Tutorial Conclusion*
- *Properties of the LLVM IR*
 - *Target Independence*
 - *Safety Guarantees*
 - *Language-Specific Optimizations*
- *Tips and Tricks*
 - *Implementing portable offsetof/sizeof*
 - *Garbage Collected Stack Frames*

Tutorial Conclusion

Welcome to the final chapter of the “[Implementing a language with LLVM](#)” tutorial. In the course of this tutorial, we have grown our little Kaleidoscope language from being a useless toy, to being a semi-interesting (but probably still useless) toy. :)

It is interesting to see how far we’ve come, and how little code it has taken. We built the entire lexer, parser, AST, code generator, an interactive run-loop (with a JIT!), and emitted debug information in standalone executables - all in under 1000 lines of (non-comment/non-blank) code.

Our little language supports a couple of interesting features: it supports user defined binary and unary operators, it uses JIT compilation for immediate evaluation, and it supports a few control flow constructs with SSA construction.

Part of the idea of this tutorial was to show you how easy and fun it can be to define, build, and play with languages. Building a compiler need not be a scary or mystical process! Now that you’ve seen some of the basics, I strongly encourage you to take the code and hack on it. For example, try adding:

- **global variables** - While global variables have questional value in modern software engineering, they are often useful when putting together quick little hacks like the Kaleidoscope compiler itself. Fortunately, our current setup makes it very easy to add global variables: just have value lookup check to see if an unresolved variable is in the global variable symbol table before rejecting it. To create a new global variable, make an instance of the `LLVM GlobalVariable` class.
- **typed variables** - Kaleidoscope currently only supports variables of type double. This gives the language a very nice elegance, because only supporting one type means that you never have to specify types. Different languages have different ways of handling this. The easiest way is to require the user to specify types for every variable definition, and record the type of the variable in the symbol table along with its `Value*`.
- **arrays, structs, vectors, etc** - Once you add types, you can start extending the type system in all sorts of interesting ways. Simple arrays are very easy and are quite useful for many different applications. Adding them is mostly an exercise in learning how the LLVM [getelementptr](#) instruction works: it is so nifty/unconventional, [it has its own FAQ!](#)
- **standard runtime** - Our current language allows the user to access arbitrary external functions, and we use it for things like “`printf`” and “`putchar`”. As you extend the language to add higher-level constructs, often these constructs make the most sense if they are lowered to calls into a language-supplied runtime. For example, if you add hash tables to the language, it would probably make sense to add the routines to a runtime, instead of inlining them all the way.
- **memory management** - Currently we can only access the stack in Kaleidoscope. It would also be useful to be able to allocate heap memory, either with calls to the standard libc `malloc/free` interface or with a garbage collector. If you would like to use garbage collection, note that LLVM fully supports [Accurate Garbage Collection](#) including algorithms that move objects and need to scan/update the stack.

- **exception handling support** - LLVM supports generation of [zero cost exceptions](#) which interoperate with code compiled in other languages. You could also generate code by implicitly making every function return an error value and checking it. You could also make explicit use of `setjmp/longjmp`. There are many different ways to go here.
- **object orientation, generics, database access, complex numbers, geometric programming, ...** - Really, there is no end of crazy features that you can add to the language.
- **unusual domains** - We've been talking about applying LLVM to a domain that many people are interested in: building a compiler for a specific language. However, there are many other domains that can use compiler technology that are not typically considered. For example, LLVM has been used to implement OpenGL graphics acceleration, translate C++ code to ActionScript, and many other cute and clever things. Maybe you will be the first to JIT compile a regular expression interpreter into native code with LLVM?

Have fun - try doing something crazy and unusual. Building a language like everyone else always has, is much less fun than trying something a little crazy or off the wall and seeing how it turns out. If you get stuck or want to talk about it, feel free to email the [llvm-dev mailing list](#): it has lots of people who are interested in languages and are often willing to help out.

Before we end this tutorial, I want to talk about some “tips and tricks” for generating LLVM IR. These are some of the more subtle things that may not be obvious, but are very useful if you want to take advantage of LLVM's capabilities.

Properties of the LLVM IR

We have a couple of common questions about code in the LLVM IR form - let's just get these out of the way right now, shall we?

Target Independence

Kaleidoscope is an example of a “portable language”: any program written in Kaleidoscope will work the same way on any target that it runs on. Many other languages have this property, e.g. lisp, java, haskell, javascript, python, etc (note that while these languages are portable, not all their libraries are).

One nice aspect of LLVM is that it is often capable of preserving target independence in the IR: you can take the LLVM IR for a Kaleidoscope-compiled program and run it on any target that LLVM supports, even emitting C code and compiling that on targets that LLVM doesn't support natively. You can trivially tell that the Kaleidoscope compiler generates target-independent code because it never queries for any target-specific information when generating code.

The fact that LLVM provides a compact, target-independent, representation for code gets a lot of people excited. Unfortunately, these people are usually thinking about C or a language from the C family when they are asking questions about language portability. I say “unfortunately”, because there is really no way to make (fully general) C code portable, other than shipping the source code around (and of course, C source code is not actually portable in general either - ever port a really old application from 32- to 64-bits?).

The problem with C (again, in its full generality) is that it is heavily laden with target specific assumptions. As one simple example, the preprocessor often destructively removes target-independence from the code when it processes the input text:

```
#ifdef __i386__
    int X = 1;
#else
    int X = 42;
#endif
```

While it is possible to engineer more and more complex solutions to problems like this, it cannot be solved in full generality in a way that is better than shipping the actual source code.

That said, there are interesting subsets of C that can be made portable. If you are willing to fix primitive types to a fixed size (say `int` = 32-bits, and `long` = 64-bits), don't care about ABI compatibility with existing binaries, and are willing to give up some other minor features, you can have portable code. This can make sense for specialized domains such as an in-kernel language.

Safety Guarantees

Many of the languages above are also “safe” languages: it is impossible for a program written in Java to corrupt its address space and crash the process (assuming the JVM has no bugs). Safety is an interesting property that requires a combination of language design, runtime support, and often operating system support.

It is certainly possible to implement a safe language in LLVM, but LLVM IR does not itself guarantee safety. The LLVM IR allows unsafe pointer casts, use after free bugs, buffer over-runs, and a variety of other problems. Safety needs to be implemented as a layer on top of LLVM and, conveniently, several groups have investigated this. Ask on the [llvm-dev mailing list](#) if you are interested in more details.

Language-Specific Optimizations

One thing about LLVM that turns off many people is that it does not solve all the world's problems in one system (sorry ‘world hunger’, someone else will have to solve you some other day). One specific complaint is that people perceive LLVM as being incapable of performing high-level language-specific optimization: LLVM “loses too much information”.

Unfortunately, this is really not the place to give you a full and unified version of “Chris Lattner’s theory of compiler design”. Instead, I’ll make a few observations:

First, you’re right that LLVM does lose information. For example, as of this writing, there is no way to distinguish in the LLVM IR whether an SSA-value came from a C “`int`” or a C “`long`” on an ILP32 machine (other than debug info). Both get compiled down to an ‘i32’ value and the information about what it came from is lost. The more general issue here, is that the LLVM type system uses “structural equivalence” instead of “name equivalence”. Another place this surprises people is if you have two types in a high-level language that have the same structure (e.g. two different structs that have a single `int` field): these types will compile down into a single LLVM type and it will be impossible to tell what it came from.

Second, while LLVM does lose information, LLVM is not a fixed target: we continue to enhance and improve it in many different ways. In addition to adding new features (LLVM did not always support exceptions or debug info), we also extend the IR to capture important information for optimization (e.g. whether an argument is sign or zero extended, information about pointers aliasing, etc). Many of the enhancements are user-driven: people want LLVM to include some specific feature, so they go ahead and extend it.

Third, it is *possible and easy* to add language-specific optimizations, and you have a number of choices in how to do it. As one trivial example, it is easy to add language-specific optimization passes that “know” things about code compiled for a language. In the case of the C family, there is an optimization pass that “knows” about the standard C library functions. If you call “`exit(0)`” in `main()`, it knows that it is safe to optimize that into “`return 0;`” because C specifies what the ‘`exit`’ function does.

In addition to simple library knowledge, it is possible to embed a variety of other language-specific information into the LLVM IR. If you have a specific need and run into a wall, please bring the topic up on the [llvm-dev](#) list. At the very worst, you can always treat LLVM as if it were a “dumb code generator” and implement the high-level optimizations you desire in your front-end, on the language-specific AST.

Tips and Tricks

There is a variety of useful tips and tricks that you come to know after working on/with LLVM that aren't obvious at first glance. Instead of letting everyone rediscover them, this section talks about some of these issues.

Implementing portable `offsetof/sizeof`

One interesting thing that comes up, if you are trying to keep the code generated by your compiler “target independent”, is that you often need to know the size of some LLVM type or the offset of some field in an LLVM structure. For example, you might need to pass the size of a type into a function that allocates memory.

Unfortunately, this can vary widely across targets: for example the width of a pointer is trivially target-specific. However, there is a [clever way to use the `getelementptr` instruction](#) that allows you to compute this in a portable way.

Garbage Collected Stack Frames

Some languages want to explicitly manage their stack frames, often so that they are garbage collected or to allow easy implementation of closures. There are often better ways to implement these features than explicit stack frames, but [LLVM does support them](#), if you want. It requires your front-end to convert the code into [Continuation Passing Style](#) and the use of tail calls (which LLVM also supports).

2.17.2 Kaleidoscope: Implementing a Language with LLVM in Objective Caml

Kaleidoscope: Tutorial Introduction and the Lexer

- [Tutorial Introduction](#)
- [The Basic Language](#)
- [The Lexer](#)

Tutorial Introduction

Welcome to the “Implementing a language with LLVM” tutorial. This tutorial runs through the implementation of a simple language, showing how fun and easy it can be. This tutorial will get you up and started as well as help to build a framework you can extend to other languages. The code in this tutorial can also be used as a playground to hack on other LLVM specific things.

The goal of this tutorial is to progressively unveil our language, describing how it is built up over time. This will let us cover a fairly broad range of language design and LLVM-specific usage issues, showing and explaining the code for it all along the way, without overwhelming you with tons of details up front.

It is useful to point out ahead of time that this tutorial is really about teaching compiler techniques and LLVM specifically, *not* about teaching modern and sane software engineering principles. In practice, this means that we'll take a number of shortcuts to simplify the exposition. For example, the code leaks memory, uses global variables all over the place, doesn't use nice design patterns like [visitors](#), etc... but it is very simple. If you dig in and use the code as a basis for future projects, fixing these deficiencies shouldn't be hard.

I've tried to put this tutorial together in a way that makes chapters easy to skip over if you are already familiar with or are uninterested in the various pieces. The structure of the tutorial is:

- *Chapter #1*: Introduction to the Kaleidoscope language, and the definition of its Lexer - This shows where we are going and the basic functionality that we want it to do. In order to make this tutorial maximally understandable and hackable, we choose to implement everything in Objective Caml instead of using lexer and parser generators. LLVM obviously works just fine with such tools, feel free to use one if you prefer.
- *Chapter #2*: Implementing a Parser and AST - With the lexer in place, we can talk about parsing techniques and basic AST construction. This tutorial describes recursive descent parsing and operator precedence parsing. Nothing in Chapters 1 or 2 is LLVM-specific, the code doesn't even link in LLVM at this point. :)
- *Chapter #3*: Code generation to LLVM IR - With the AST ready, we can show off how easy generation of LLVM IR really is.
- *Chapter #4*: Adding JIT and Optimizer Support - Because a lot of people are interested in using LLVM as a JIT, we'll dive right into it and show you the 3 lines it takes to add JIT support. LLVM is also useful in many other ways, but this is one simple and "sexy" way to shows off its power. :)
- *Chapter #5*: Extending the Language: Control Flow - With the language up and running, we show how to extend it with control flow operations (if/then/else and a 'for' loop). This gives us a chance to talk about simple SSA construction and control flow.
- *Chapter #6*: Extending the Language: User-defined Operators - This is a silly but fun chapter that talks about extending the language to let the user program define their own arbitrary unary and binary operators (with assignable precedence!). This lets us build a significant piece of the "language" as library routines.
- *Chapter #7*: Extending the Language: Mutable Variables - This chapter talks about adding user-defined local variables along with an assignment operator. The interesting part about this is how easy and trivial it is to construct SSA form in LLVM: no, LLVM does *not* require your front-end to construct SSA form!
- *Chapter #8*: Conclusion and other useful LLVM tidbits - This chapter wraps up the series by talking about potential ways to extend the language, but also includes a bunch of pointers to info about "special topics" like adding garbage collection support, exceptions, debugging, support for "spaghetti stacks", and a bunch of other tips and tricks.

By the end of the tutorial, we'll have written a bit less than 700 lines of non-comment, non-blank, lines of code. With this small amount of code, we'll have built up a very reasonable compiler for a non-trivial language including a hand-written lexer, parser, AST, as well as code generation support with a JIT compiler. While other systems may have interesting "hello world" tutorials, I think the breadth of this tutorial is a great testament to the strengths of LLVM and why you should consider it if you're interested in language or compiler design.

A note about this tutorial: we expect you to extend the language and play with it on your own. Take the code and go crazy hacking away at it, compilers don't need to be scary creatures - it can be a lot of fun to play with languages!

The Basic Language

This tutorial will be illustrated with a toy language that we'll call "**Kaleidoscope**" (derived from "meaning beautiful, form, and view"). Kaleidoscope is a procedural language that allows you to define functions, use conditionals, math, etc. Over the course of the tutorial, we'll extend Kaleidoscope to support the if/then/else construct, a for loop, user defined operators, JIT compilation with a simple command line interface, etc.

Because we want to keep things simple, the only datatype in Kaleidoscope is a 64-bit floating point type (aka 'float' in OCaml parlance). As such, all values are implicitly double precision and the language doesn't require type declarations. This gives the language a very nice and simple syntax. For example, the following simple example computes Fibonacci numbers:

```
# Compute the x'th fibonacci number.
def fib(x)
  if x < 3 then
    1
```

```
else
  fib(x-1)+fib(x-2)

# This expression will compute the 40th number.
fib(40)
```

We also allow Kaleidoscope to call into standard library functions (the LLVM JIT makes this completely trivial). This means that you can use the ‘extern’ keyword to define a function before you use it (this is also useful for mutually recursive functions). For example:

```
extern sin(arg);
extern cos(arg);
extern atan2(arg1 arg2);

atan2(sin(.4), cos(.42))
```

A more interesting example is included in Chapter 6 where we write a little Kaleidoscope application that [displays a Mandelbrot Set](#) at various levels of magnification.

Lets dive into the implementation of this language!

The Lexer

When it comes to implementing a language, the first thing needed is the ability to process a text file and recognize what it says. The traditional way to do this is to use a “lexer” (aka ‘scanner’) to break the input up into “tokens”. Each token returned by the lexer includes a token code and potentially some metadata (e.g. the numeric value of a number). First, we define the possibilities:

```
(* The lexer returns these 'Kwd' if it is an unknown character, otherwise one of
 * these others for known things. *)
type token =
  (* commands *)
  | Def | Extern

  (* primary *)
  | Ident of string | Number of float

  (* unknown *)
  | Kwd of char
```

Each token returned by our lexer will be one of the token variant values. An unknown character like ‘+’ will be returned as `Token.Kwd '+'`. If the curr token is an identifier, the value will be `Token.Ident s`. If the current token is a numeric literal (like 1.0), the value will be `Token.Number 1.0`.

The actual implementation of the lexer is a collection of functions driven by a function named `Lexer.lex`. The `Lexer.lex` function is called to return the next token from standard input. We will use [Camlp4](#) to simplify the tokenization of the standard input. Its definition starts as:

```
(*=====
 * Lexer
 *=====*)

let rec lex = parser
  (* Skip any whitespace. *)
  | [< ' (' ' | '\n' | '\r' | '\t'); stream >] -> lex stream
```

`Lexer.lex` works by recursing over a `char Stream.t` to read characters one at a time from the standard input. It eats them as it recognizes them and stores them in a `Token.token` variant. The first thing that it has to do is ignore whitespace between tokens. This is accomplished with the recursive call above.

The next thing `Lexer.lex` needs to do is recognize identifiers and specific keywords like “def”. Kaleidoscope does this with a pattern match and a helper function.

```
(* identifier: [a-zA-Z][a-zA-Z0-9]* *)
| [< ' ('A' .. 'Z' | 'a' .. 'z' as c); stream >] ->
    let buffer = Buffer.create 1 in
    Buffer.add_char buffer c;
    lex_ident buffer stream

...

and lex_ident buffer = parser
| [< ' ('A' .. 'Z' | 'a' .. 'z' | '0' .. '9' as c); stream >] ->
    Buffer.add_char buffer c;
    lex_ident buffer stream
| [< stream=lex >] ->
    match Buffer.contents buffer with
    | "def" -> [< 'Token.Def; stream >]
    | "extern" -> [< 'Token.Extern; stream >]
    | id -> [< 'Token.Ident id; stream >]
```

Numeric values are similar:

```
(* number: [0-9.]+ *)
| [< ' ('0' .. '9' as c); stream >] ->
    let buffer = Buffer.create 1 in
    Buffer.add_char buffer c;
    lex_number buffer stream

...

and lex_number buffer = parser
| [< ' ('0' .. '9' | '.' as c); stream >] ->
    Buffer.add_char buffer c;
    lex_number buffer stream
| [< stream=lex >] ->
    [< 'Token.Number (float_of_string (Buffer.contents buffer)); stream >]
```

This is all pretty straight-forward code for processing input. When reading a numeric value from input, we use the ocaml `float_of_string` function to convert it to a numeric value that we store in `Token.Number`. Note that this isn’t doing sufficient error checking: it will raise `Failure` if the string “1.23.45.67”. Feel free to extend it :). Next we handle comments:

```
(* Comment until end of line. *)
| [< ' ('#'); stream >] ->
    lex_comment stream

...

and lex_comment = parser
| [< ' ('\n'); stream=lex >] -> stream
| [< 'c; e=lex_comment >] -> e
| [< >] -> [< >]
```

We handle comments by skipping to the end of the line and then return the next token. Finally, if the input doesn’t

match one of the above cases, it is either an operator character like ‘+’ or the end of the file. These are handled with this code:

```
(* Otherwise, just return the character as its ascii value. *)
| [< 'c; stream >] ->
    [< 'Token.Kwd c; lex stream >]

(* end of stream. *)
| [< >] -> [< >]
```

With this, we have the complete lexer for the basic Kaleidoscope language (the [full code listing](#) for the Lexer is available in the [next chapter](#) of the tutorial). Next we’ll build a [simple parser that uses this to build an Abstract Syntax Tree](#). When we have that, we’ll include a driver so that you can use the lexer and parser together.

Next: [Implementing a Parser and AST](#)

Kaleidoscope: Implementing a Parser and AST

- *Chapter 2 Introduction*
- *The Abstract Syntax Tree (AST)*
- *Parser Basics*
- *Basic Expression Parsing*
- *Binary Expression Parsing*
- *Parsing the Rest*
- *The Driver*
- *Conclusions*
- *Full Code Listing*

Chapter 2 Introduction

Welcome to Chapter 2 of the “[Implementing a language with LLVM in Objective Caml](#)” tutorial. This chapter shows you how to use the lexer, built in [Chapter 1](#), to build a full [parser](#) for our Kaleidoscope language. Once we have a parser, we’ll define and build an [Abstract Syntax Tree \(AST\)](#).

The parser we will build uses a combination of [Recursive Descent Parsing](#) and [Operator-Precedence Parsing](#) to parse the Kaleidoscope language (the latter for binary expressions and the former for everything else). Before we get to parsing though, lets talk about the output of the parser: the Abstract Syntax Tree.

The Abstract Syntax Tree (AST)

The AST for a program captures its behavior in such a way that it is easy for later stages of the compiler (e.g. code generation) to interpret. We basically want one object for each construct in the language, and the AST should closely model the language. In Kaleidoscope, we have expressions, a prototype, and a function object. We’ll start with expressions first:

```
(* expr - Base type for all expression nodes. *)
type expr =
  (* variant for numeric literals like "1.0". *)
  | Number of float
```

The code above shows the definition of the base ExprAST class and one subclass which we use for numeric literals. The important thing to note about this code is that the Number variant captures the numeric value of the literal as an instance variable. This allows later phases of the compiler to know what the stored numeric value is.

Right now we only create the AST, so there are no useful functions on them. It would be very easy to add a function to pretty print the code, for example. Here are the other expression AST node definitions that we'll use in the basic form of the Kaleidoscope language:

```
(* variant for referencing a variable, like "a". *)
| Variable of string

(* variant for a binary operator. *)
| Binary of char * expr * expr

(* variant for function calls. *)
| Call of string * expr array
```

This is all (intentionally) rather straight-forward: variables capture the variable name, binary operators capture their opcode (e.g. '+'), and calls capture a function name as well as a list of any argument expressions. One thing that is nice about our AST is that it captures the language features without talking about the syntax of the language. Note that there is no discussion about precedence of binary operators, lexical structure, etc.

For our basic language, these are all of the expression nodes we'll define. Because it doesn't have conditional control flow, it isn't Turing-complete; we'll fix that in a later installment. The two things we need next are a way to talk about the interface to a function, and a way to talk about functions themselves:

```
(* proto - This type represents the "prototype" for a function, which captures
 * its name, and its argument names (thus implicitly the number of arguments the
 * function takes). *)
type proto = Prototype of string * string array

(* func - This type represents a function definition itself. *)
type func = Function of proto * expr
```

In Kaleidoscope, functions are typed with just a count of their arguments. Since all values are double precision floating point, the type of each argument doesn't need to be stored anywhere. In a more aggressive and realistic language, the "expr" variants would probably have a type field.

With this scaffolding, we can now talk about parsing expressions and function bodies in Kaleidoscope.

Parser Basics

Now that we have an AST to build, we need to define the parser code to build it. The idea here is that we want to parse something like "x+y" (which is returned as three tokens by the lexer) into an AST that could be generated with calls like this:

```
let x = Variable "x" in
let y = Variable "y" in
let result = Binary ('+', x, y) in
...
```

The error handling routines make use of the builtin `Stream.Failure` and `Stream.Error`'s. `Stream.Failure` is raised when the parser is unable to find any matching token in the first position of a pattern. `Stream.Error` is raised when the first token matches, but the rest do not. The error recovery in our parser will not be the best and is not particular user-friendly, but it will be enough for our tutorial. These exceptions make it easier to handle errors in routines that have various return types.

With these basic types and exceptions, we can implement the first piece of our grammar: numeric literals.

Basic Expression Parsing

We start with numeric literals, because they are the simplest to process. For each production in our grammar, we'll define a function which parses that production. We call this class of expressions "primary" expressions, for reasons that will become more clear [later in the tutorial](#). In order to parse an arbitrary primary expression, we need to determine what sort of expression it is. For numeric literals, we have:

```
(* primary
 * ::= identifier
 * ::= numberexpr
 * ::= parenexpr *)
parse_primary = parser
  (* numberexpr ::= number *)
  | [< 'Token.Number n >] -> Ast.Number n
```

This routine is very simple: it expects to be called when the current token is a `Token.Number` token. It takes the current number value, creates a `Ast.Number` node, advances the lexer to the next token, and finally returns.

There are some interesting aspects to this. The most important one is that this routine eats all of the tokens that correspond to the production and returns the lexer buffer with the next token (which is not part of the grammar production) ready to go. This is a fairly standard way to go for recursive descent parsers. For a better example, the parenthesis operator is defined like this:

```
(* parenexpr ::= '(' expression ')' *)
| [< 'Token.Kwd '('; e=parse_expr; 'Token.Kwd ')' ?? "expected ')" >] -> e
```

This function illustrates a number of interesting things about the parser:

1) It shows how we use the `Stream.Error` exception. When called, this function expects that the current token is a '(' token, but after parsing the subexpression, it is possible that there is no ')' waiting. For example, if the user types in "(4 x" instead of "(4)", the parser should emit an error. Because errors can occur, the parser needs a way to indicate that they happened. In our parser, we use the `camlp4` shortcut syntax token `?? "parse error"`, where if the token before the `??` does not match, then `Stream.Error "parse error"` will be raised.

2) Another interesting aspect of this function is that it uses recursion by calling `Parser.parse_primary` (we will soon see that `Parser.parse_primary` can call `Parser.parse_primary`). This is powerful because it allows us to handle recursive grammars, and keeps each production very simple. Note that parentheses do not cause construction of AST nodes themselves. While we could do it this way, the most important role of parentheses are to guide the parser and provide grouping. Once the parser constructs the AST, parentheses are not needed.

The next simple production is for handling variable references and function calls:

```
(* identifierexpr
 * ::= identifier
 * ::= identifier '(' argumentexpr ')' *)
| [< 'Token.Ident id; stream >] ->
  let rec parse_args accumulator = parser
    | [< e=parse_expr; stream >] ->
      begin parser
```

```

        | [< 'Token.Kwd ','; e=parse_args (e :: accumulator) >] -> e
        | [< >] -> e :: accumulator
    end stream
    | [< >] -> accumulator
in
let rec parse_ident id = parser
  (* Call. *)
  | [< 'Token.Kwd '(';
    args=parse_args [];
    'Token.Kwd ')' ?? "expected ')">] ->
    Ast.Call (id, Array.of_list (List.rev args))

  (* Simple variable ref. *)
  | [< >] -> Ast.Variable id
in
parse_ident id stream

```

This routine follows the same style as the other routines. (It expects to be called if the current token is a `Token.Ident` token). It also has recursion and error handling. One interesting aspect of this is that it uses *look-ahead* to determine if the current identifier is a stand alone variable reference or if it is a function call expression. It handles this by checking to see if the token after the identifier is a `'('` token, constructing either a `Ast.Variable` or `Ast.Call` node as appropriate.

We finish up by raising an exception if we received a token we didn't expect:

```

| [< >] -> raise (Stream.Error "unknown token when expecting an expression.")

```

Now that basic expressions are handled, we need to handle binary expressions. They are a bit more complex.

Binary Expression Parsing

Binary expressions are significantly harder to parse because they are often ambiguous. For example, when given the string `"x+y*z"`, the parser can choose to parse it as either `"(x+y)*z"` or `"x+(y*z)"`. With common definitions from mathematics, we expect the later parse, because `"*"` (multiplication) has higher *precedence* than `"+"` (addition).

There are many ways to handle this, but an elegant and efficient way is to use [Operator-Precedence Parsing](#). This parsing technique uses the precedence of binary operators to guide recursion. To start with, we need a table of precedences:

```

(* binop_precedence - This holds the precedence for each binary operator that is
 * defined *)
let binop_precedence: (char, int) Hashtbl.t = Hashtbl.create 10

(* precedence - Get the precedence of the pending binary operator token. *)
let precedence c = try Hashtbl.find binop_precedence c with Not_found -> -1

...

let main () =
  (* Install standard binary operators.
   * 1 is the lowest precedence. *)
  Hashtbl.add Parser.binop_precedence '<' 10;
  Hashtbl.add Parser.binop_precedence '+' 20;
  Hashtbl.add Parser.binop_precedence '-' 20;
  Hashtbl.add Parser.binop_precedence '*' 40;      (* highest. *)
  ...

```

For the basic form of Kaleidoscope, we will only support 4 binary operators (this can obviously be extended by you, our brave and intrepid reader). The `Parser.precedence` function returns the precedence for the current token, or -1 if the token is not a binary operator. Having a `Hashtbl.t` makes it easy to add new operators and makes it clear that the algorithm doesn't depend on the specific operators involved, but it would be easy enough to eliminate the `Hashtbl.t` and do the comparisons in the `Parser.precedence` function. (Or just use a fixed-size array).

With the helper above defined, we can now start parsing binary expressions. The basic idea of operator precedence parsing is to break down an expression with potentially ambiguous binary operators into pieces. Consider, for example, the expression `"a+b+(c+d)*e*f+g"`. Operator precedence parsing considers this as a stream of primary expressions separated by binary operators. As such, it will first parse the leading primary expression `"a"`, then it will see the pairs `[+, b]` `[+, (c+d)]` `[*, e]` `[*, f]` and `[+, g]`. Note that because parentheses are primary expressions, the binary expression parser doesn't need to worry about nested subexpressions like `(c+d)` at all.

To start, an expression is a primary expression potentially followed by a sequence of `[binop,primaryexpr]` pairs:

```
(* expression
 *   ::= primary binoprhs *)
and parse_expr = parser
| [< lhs=parse_primary; stream >] -> parse_bin_rhs 0 lhs stream
```

`Parser.parse_bin_rhs` is the function that parses the sequence of pairs for us. It takes a precedence and a pointer to an expression for the part that has been parsed so far. Note that `"x"` is a perfectly valid expression: As such, `"binoprhs"` is allowed to be empty, in which case it returns the expression that is passed into it. In our example above, the code passes the expression for `"a"` into `Parser.parse_bin_rhs` and the current token is `"+"`.

The precedence value passed into `Parser.parse_bin_rhs` indicates the *minimal operator precedence* that the function is allowed to eat. For example, if the current pair stream is `[+, x]` and `Parser.parse_bin_rhs` is passed in a precedence of 40, it will not consume any tokens (because the precedence of `'+'` is only 20). With this in mind, `Parser.parse_bin_rhs` starts with:

```
(* binoprhs
 *   ::= ('+' primary)* *)
and parse_bin_rhs expr_prec lhs stream =
  match Stream.peek stream with
  (* If this is a binop, find its precedence. *)
  | Some (Token.Kwd c) when Hashtbl.mem binop_precedence c ->
    let token_prec = precedence c in

    (* If this is a binop that binds at least as tightly as the current binop,
     * consume it, otherwise we are done. *)
    if token_prec < expr_prec then lhs else begin
```

This code gets the precedence of the current token and checks to see if it is too low. Because we defined invalid tokens to have a precedence of -1, this check implicitly knows that the pair-stream ends when the token stream runs out of binary operators. If this check succeeds, we know that the token is a binary operator and that it will be included in this expression:

```
(* Eat the binop. *)
Stream.junk stream;

(* Parse the primary expression after the binary operator *)
let rhs = parse_primary stream in

(* Okay, we know this is a binop. *)
let rhs =
  match Stream.peek stream with
  | Some (Token.Kwd c2) ->
```


As such, this code eats (and remembers) the binary operator and then parses the primary expression that follows. This builds up the whole pair, the first of which is [+ , b] for the running example.

Now that we parsed the left-hand side of an expression and one pair of the RHS sequence, we have to decide which way the expression associates. In particular, we could have “(a+b) binop unparsed” or “a + (b binop unparsed)”. To determine this, we look ahead at “binop” to determine its precedence and compare it to BinOp’s precedence (which is ‘+’ in this case):

```
(* If BinOp binds less tightly with rhs than the operator after
 * rhs, let the pending operator take rhs as its lhs. *)
let next_prec = precedence c2 in
if token_prec < next_prec
```

If the precedence of the binop to the right of “RHS” is lower or equal to the precedence of our current operator, then we know that the parentheses associate as “(a+b) binop ...”. In our example, the current operator is “+” and the next operator is “+”, we know that they have the same precedence. In this case we’ll create the AST node for “a+b”, and then continue parsing:

```
... if body omitted ...
in

(* Merge lhs/rhs. *)
let lhs = Ast.Binary (c, lhs, rhs) in
parse_bin_rhs expr_prec lhs stream
end
```

In our example above, this will turn “a+b+” into “(a+b)” and execute the next iteration of the loop, with “+” as the current token. The code above will eat, remember, and parse “(c+d)” as the primary expression, which makes the current pair equal to [+ , (c+d)]. It will then evaluate the ‘if’ conditional above with “*” as the binop to the right of the primary. In this case, the precedence of “*” is higher than the precedence of “+” so the if condition will be entered.

The critical question left here is “how can the if condition parse the right hand side in full”? In particular, to build the AST correctly for our example, it needs to get all of “(c+d)*e*f” as the RHS expression variable. The code to do this is surprisingly simple (code from the above two blocks duplicated for context):

```
match Stream.peek stream with
| Some (Token.Kwd c2) ->
    (* If BinOp binds less tightly with rhs than the operator after
     * rhs, let the pending operator take rhs as its lhs. *)
    if token_prec < precedence c2
    then parse_bin_rhs (token_prec + 1) rhs stream
    else rhs
| _ -> rhs
in

(* Merge lhs/rhs. *)
let lhs = Ast.Binary (c, lhs, rhs) in
parse_bin_rhs expr_prec lhs stream
end
```

At this point, we know that the binary operator to the RHS of our primary has higher precedence than the binop we are currently parsing. As such, we know that any sequence of pairs whose operators are all higher precedence than “+” should be parsed together and returned as “RHS”. To do this, we recursively invoke the `Parser.parse_bin_rhs` function specifying “token_prec+1” as the minimum precedence required for it to continue. In our example above, this will cause it to return the AST node for “(c+d)*e*f” as RHS, which is then set as the RHS of the ‘+’ expression.

Finally, on the next iteration of the while loop, the “+g” piece is parsed and added to the AST. With this little bit of code (14 non-trivial lines), we correctly handle fully general binary expression parsing in a very elegant way. This was

a whirlwind tour of this code, and it is somewhat subtle. I recommend running through it with a few tough examples to see how it works.

This wraps up handling of expressions. At this point, we can point the parser at an arbitrary token stream and build an expression from it, stopping at the first token that is not part of the expression. Next up we need to handle function definitions, etc.

Parsing the Rest

The next thing missing is handling of function prototypes. In Kaleidoscope, these are used both for ‘extern’ function declarations as well as function body definitions. The code to do this is straight-forward and not very interesting (once you’ve survived expressions):

```
(* prototype
 * ::= id '(' id* ')' *)
let parse_prototype =
  let rec parse_args accumulator = parser
    | [< 'Token.Ident' id; e=parse_args (id::accumulator) >] -> e
    | [< >] -> accumulator
  in

  parser
  | [< 'Token.Ident' id;
    'Token.Kwd '(' ?? "expected '(' in prototype";
    args=parse_args [];
    'Token.Kwd ')' ?? "expected ')' in prototype" >] ->
    (* success. *)
    Ast.Prototype (id, Array.of_list (List.rev args))

  | [< >] ->
    raise (Stream.Error "expected function name in prototype")
```

Given this, a function definition is very simple, just a prototype plus an expression to implement the body:

```
(* definition ::= 'def' prototype expression *)
let parse_definition = parser
  | [< 'Token.Def'; p=parse_prototype; e=parse_expr >] ->
    Ast.Function (p, e)
```

In addition, we support ‘extern’ to declare functions like ‘sin’ and ‘cos’ as well as to support forward declaration of user functions. These ‘extern’s are just prototypes with no body:

```
(* external ::= 'extern' prototype *)
let parse_extern = parser
  | [< 'Token.Extern'; e=parse_prototype >] -> e
```

Finally, we’ll also let the user type in arbitrary top-level expressions and evaluate them on the fly. We will handle this by defining anonymous nullary (zero argument) functions for them:

```
(* toplevelexpr ::= expression *)
let parse_toplevel = parser
  | [< e=parse_expr >] ->
    (* Make an anonymous proto. *)
    Ast.Function (Ast.Prototype ("", []), e)
```

Now that we have all the pieces, let’s build a little driver that will let us actually *execute* this code we’ve built!

The Driver

The driver for this simply invokes all of the parsing pieces with a top-level dispatch loop. There isn't much interesting here, so I'll just include the top-level loop. See [below](#) for full code in the "Top-Level Parsing" section.

```
(* top ::= definition | external | expression | ';' *)
let rec main_loop stream =
  match Stream.peek stream with
  | None -> ()

  (* ignore top-level semicolons. *)
  | Some (Token.Kwd ';') ->
    Stream.junk stream;
    main_loop stream

  | Some token ->
    begin
      try match token with
      | Token.Def ->
        ignore(Parser.parse_definition stream);
        print_endline "parsed a function definition.";
      | Token.Extern ->
        ignore(Parser.parse_extern stream);
        print_endline "parsed an extern.";
      | _ ->
        (* Evaluate a top-level expression into an anonymous function. *)
        ignore(Parser.parse_toplevel stream);
        print_endline "parsed a top-level expr";
      with Stream.Error s ->
        (* Skip token for error recovery. *)
        Stream.junk stream;
        print_endline s;
    end;
    print_string "ready> "; flush stdout;
    main_loop stream
```

The most interesting part of this is that we ignore top-level semicolons. Why is this, you ask? The basic reason is that if you type "4 + 5" at the command line, the parser doesn't know whether that is the end of what you will type or not. For example, on the next line you could type "def foo..." in which case 4+5 is the end of a top-level expression. Alternatively you could type "* 6", which would continue the expression. Having top-level semicolons allows you to type "4+5;"; and the parser will know you are done.

Conclusions

With just under 300 lines of commented code (240 lines of non-comment, non-blank code), we fully defined our minimal language, including a lexer, parser, and AST builder. With this done, the executable will validate Kaleidoscope code and tell us if it is grammatically invalid. For example, here is a sample interaction:

```
$ ./toy.byte
ready> def foo(x y) x+foo(y, 4.0);
Parsed a function definition.
ready> def foo(x y) x+y y;
Parsed a function definition.
Parsed a top-level expr
ready> def foo(x y) x+y );
Parsed a function definition.
```

```
Error: unknown token when expecting an expression
ready> extern sin(a);
ready> Parsed an extern
ready> ^D
$
```

There is a lot of room for extension here. You can define new AST nodes, extend the language in many ways, etc. In the [next installment](#), we will describe how to generate LLVM Intermediate Representation (IR) from the AST.

Full Code Listing

Here is the complete code listing for this and the previous chapter. Note that it is fully self-contained: you don't need LLVM or any external libraries at all for this. (Besides the ocaml standard libraries, of course.) To build this, just compile with:

```
# Compile
ocamlbuild toy.byte
# Run
./toy.byte
```

Here is the code:

_tags:

```
<{lexer,parser}.ml>: use_camlp4, pp(camlp4of)
```

token.ml:

```
(=====)
* Lexer Tokens
*=====)

(* The lexer returns these 'Kwd' if it is an unknown character, otherwise one of
 * these others for known things. *)
type token =
  (* commands *)
  | Def | Extern

  (* primary *)
  | Ident of string | Number of float

  (* unknown *)
  | Kwd of char
```

lexer.ml:

```
(=====)
* Lexer
*=====)

let rec lex = parser
  (* Skip any whitespace. *)
  | [< ' (' | '\n' | '\r' | '\t'); stream >] -> lex stream

  (* identifier: [a-zA-Z][a-zA-Z0-9] *)
  | [< ' ('A' .. 'Z' | 'a' .. 'z' as c); stream >] ->
```

```

    let buffer = Buffer.create 1 in
    Buffer.add_char buffer c;
    lex_ident buffer stream

(* number: [0-9.]+ *)
| [< ' ('0' .. '9' as c); stream >] ->
    let buffer = Buffer.create 1 in
    Buffer.add_char buffer c;
    lex_number buffer stream

(* Comment until end of line. *)
| [< ' ('#'); stream >] ->
    lex_comment stream

(* Otherwise, just return the character as its ascii value. *)
| [< 'c; stream >] ->
    [< 'Token.Kwd c; lex stream >]

(* end of stream. *)
| [< >] -> [< >]

and lex_number buffer = parser
| [< ' ('0' .. '9' | '.' as c); stream >] ->
    Buffer.add_char buffer c;
    lex_number buffer stream
| [< stream=lex >] ->
    [< 'Token.Number (float_of_string (Buffer.contents buffer)); stream >]

and lex_ident buffer = parser
| [< ' ('A' .. 'Z' | 'a' .. 'z' | '0' .. '9' as c); stream >] ->
    Buffer.add_char buffer c;
    lex_ident buffer stream
| [< stream=lex >] ->
    match Buffer.contents buffer with
    | "def" -> [< 'Token.Def; stream >]
    | "extern" -> [< 'Token.Extern; stream >]
    | id -> [< 'Token.Ident id; stream >]

and lex_comment = parser
| [< ' ('\n'); stream=lex >] -> stream
| [< 'c; e=lex_comment >] -> e
| [< >] -> [< >]

```

ast.ml:

```

(*====-----*)
* Abstract Syntax Tree (aka Parse Tree)
*====-----*)

(* expr - Base type for all expression nodes. *)
type expr =
  (* variant for numeric literals like "1.0". *)
  | Number of float

  (* variant for referencing a variable, like "a". *)
  | Variable of string

  (* variant for a binary operator. *)

```

```

| Binary of char * expr * expr

(* variant for function calls. *)
| Call of string * expr array

(* proto - This type represents the "prototype" for a function, which captures
 * its name, and its argument names (thus implicitly the number of arguments the
 * function takes). *)
type proto = Prototype of string * string array

(* func - This type represents a function definition itself. *)
type func = Function of proto * expr

```

parser.ml:

```

(=====
 * Parser
 *=====)

(* binop_precedence - This holds the precedence for each binary operator that is
 * defined *)
let binop_precedence: (char, int) Hashtbl.t = Hashtbl.create 10

(* precedence - Get the precedence of the pending binary operator token. *)
let precedence c = try Hashtbl.find binop_precedence c with Not_found -> -1

(* primary
 * ::= identifier
 * ::= numberexpr
 * ::= parenexpr *)
let rec parse_primary = parser
  (* numberexpr ::= number *)
  | ['Token.Number' n >] -> Ast.Number n

  (* parenexpr ::= '(' expression ')' *)
  | ['Token.Kwd '('; e=parse_expr; 'Token.Kwd ')'' ?? "expected ')'" >] -> e

  (* identifierexpr
   * ::= identifier
   * ::= identifier '(' argumentexpr ')' *)
  | ['Token.Ident' id; stream >] ->
    let rec parse_args accumulator = parser
      | ['Token.Kwd ','; e=parse_expr; stream >] ->
        begin parser
          | ['Token.Kwd '('; e=parse_args (e :: accumulator) >] -> e
          | [>] -> e :: accumulator
        end stream
      | [>] -> accumulator
    in
    let rec parse_ident id = parser
      (* Call. *)
      | ['Token.Kwd '(';
        args=parse_args [];
        'Token.Kwd ')'' ?? "expected ')'" >] ->
        Ast.Call (id, Array.of_list (List.rev args))

      (* Simple variable ref. *)
      | [>] -> Ast.Variable id

```

```

    in
    parse_ident id stream

    | [< >] -> raise (Stream.Error "unknown token when expecting an expression.")

    (* binoprhs
    *  ::= ('+' primary)* *)
    and parse_bin_rhs expr_prec lhs stream =
    match Stream.peek stream with
    (* If this is a binop, find its precedence. *)
    | Some (Token.Kwd c) when Hashtbl.mem binop_precedence c ->
        let token_prec = precedence c in

        (* If this is a binop that binds at least as tightly as the current binop,
        * consume it, otherwise we are done. *)
        if token_prec < expr_prec then lhs else begin
            (* Eat the binop. *)
            Stream.junk stream;

            (* Parse the primary expression after the binary operator. *)
            let rhs = parse_primary stream in

            (* Okay, we know this is a binop. *)
            let rhs =
                match Stream.peek stream with
                | Some (Token.Kwd c2) ->
                    (* If BinOp binds less tightly with rhs than the operator after
                    * rhs, let the pending operator take rhs as its lhs. *)
                    let next_prec = precedence c2 in
                    if token_prec < next_prec
                    then parse_bin_rhs (token_prec + 1) rhs stream
                    else rhs
                | _ -> rhs
            in

            (* Merge lhs/rhs. *)
            let lhs = Ast.Binary (c, lhs, rhs) in
            parse_bin_rhs expr_prec lhs stream
        end
    | _ -> lhs

    (* expression
    *  ::= primary binoprhs *)
    and parse_expr = parser
    | [< lhs=parse_primary; stream >] -> parse_bin_rhs 0 lhs stream

    (* prototype
    *  ::= id '(' id* ')' *)
    let parse_prototype =
        let rec parse_args accumulator = parser
            | [< 'Token.Ident' id; e=parse_args (id::accumulator) >] -> e
            | [< >] -> accumulator
        in

        parser
        | [< 'Token.Ident' id;
            'Token.Kwd' '(' ?? "expected '(' in prototype";
            args=parse_args []];

```

```

    'Token.Kwd ') ' ?? "expected ')' in prototype" >] ->
    (* success. *)
    Ast.Prototype (id, Array.of_list (List.rev args))

| [< >] ->
    raise (Stream.Error "expected function name in prototype")

(* definition ::= 'def' prototype expression *)
let parse_definition = parser
| [< 'Token.Def; p=parse_prototype; e=parse_expr >] ->
    Ast.Function (p, e)

(* toplevelexpr ::= expression *)
let parse_toplevel = parser
| [< e=parse_expr >] ->
    (* Make an anonymous proto. *)
    Ast.Function (Ast.Prototype ("", []), e)

(* external ::= 'extern' prototype *)
let parse_extern = parser
| [< 'Token.Extern; e=parse_prototype >] -> e

```

toplevel.ml:

```

(=====
 * Top-Level parsing and JIT Driver
 =====)

(* top ::= definition | external | expression | ';' *)
let rec main_loop stream =
  match Stream.peek stream with
  | None -> ()

  (* ignore top-level semicolons. *)
  | Some (Token.Kwd ';') ->
      Stream.junk stream;
      main_loop stream

  | Some token ->
      begin
        try match token with
        | Token.Def ->
            ignore(Parser.parse_definition stream);
            print_endline "parsed a function definition.";
        | Token.Extern ->
            ignore(Parser.parse_extern stream);
            print_endline "parsed an extern.";
        | _ ->
            (* Evaluate a top-level expression into an anonymous function. *)
            ignore(Parser.parse_toplevel stream);
            print_endline "parsed a top-level expr";
        with Stream.Error s ->
            (* Skip token for error recovery. *)
            Stream.junk stream;
            print_endline s;
        end;
        print_string "ready> "; flush stdout;
        main_loop stream

```


toy.ml:

```
(*====-----*)
* Main driver code.
*====-----*)

let main () =
  (* Install standard binary operators.
   * 1 is the lowest precedence. *)
  Hashtbl.add Parser.binop_precedence '<' 10;
  Hashtbl.add Parser.binop_precedence '+' 20;
  Hashtbl.add Parser.binop_precedence '-' 20;
  Hashtbl.add Parser.binop_precedence '*' 40;      (* highest. *)

  (* Prime the first token. *)
  print_string "ready> "; flush stdout;
  let stream = Lexer.lex (Stream.of_channel stdin) in

  (* Run the main "interpreter loop" now. *)
  Toplevel.main_loop stream;
;;

main ()
```

Next: Implementing Code Generation to LLVM IR

Kaleidoscope: Code generation to LLVM IR

- *Chapter 3 Introduction*
- *Code Generation Setup*
- *Expression Code Generation*
- *Function Code Generation*
- *Driver Changes and Closing Thoughts*
- *Full Code Listing*

Chapter 3 Introduction

Welcome to Chapter 3 of the “[Implementing a language with LLVM](#)” tutorial. This chapter shows you how to transform the [Abstract Syntax Tree](#), built in Chapter 2, into LLVM IR. This will teach you a little bit about how LLVM does things, as well as demonstrate how easy it is to use. It’s much more work to build a lexer and parser than it is to generate LLVM IR code. :)

Please note: the code in this chapter and later require LLVM 2.3 or LLVM SVN to work. LLVM 2.2 and before will not work with it.

Code Generation Setup

In order to generate LLVM IR, we want some simple setup to get started. First we define virtual code generation (codegen) methods in each AST class:

```
let rec codegen_expr = function
| Ast.Number n -> ...
| Ast.Variable name -> ...
```

The `Codgen.codegen_expr` function says to emit IR for that AST node along with all the things it depends on, and they all return an LLVM Value object. “Value” is the class used to represent a “[Static Single Assignment \(SSA\)](#) register” or “SSA value” in LLVM. The most distinct aspect of SSA values is that their value is computed as the related instruction executes, and it does not get a new value until (and if) the instruction re-executes. In other words, there is no way to “change” an SSA value. For more information, please read up on [Static Single Assignment](#) - the concepts are really quite natural once you grok them.

The second thing we want is an “Error” exception like we used for the parser, which will be used to report errors found during code generation (for example, use of an undeclared parameter):

```
exception Error of string

let context = global_context ()
let the_module = create_module context "my cool jit"
let builder = builder context
let named_values:(string, llvalue) Hashtbl.t = Hashtbl.create 10
let double_type = double_type context
```

The static variables will be used during code generation. `Codgen.the_module` is the LLVM construct that contains all of the functions and global variables in a chunk of code. In many ways, it is the top-level structure that the LLVM IR uses to contain code.

The `Codgen.builder` object is a helper object that makes it easy to generate LLVM instructions. Instances of the `IRBuilder` class keep track of the current place to insert instructions and has methods to create new instructions.

The `Codgen.named_values` map keeps track of which values are defined in the current scope and what their LLVM representation is. (In other words, it is a symbol table for the code). In this form of Kaleidoscope, the only things that can be referenced are function parameters. As such, function parameters will be in this map when generating code for their function body.

With these basics in place, we can start talking about how to generate code for each expression. Note that this assumes that the `Codgen.builder` has been set up to generate code *into* something. For now, we’ll assume that this has already been done, and we’ll just use it to emit code.

Expression Code Generation

Generating LLVM code for expression nodes is very straightforward: less than 30 lines of commented code for all four of our expression nodes. First we’ll do numeric literals:

```
| Ast.Number n -> const_float double_type n
```

In the LLVM IR, numeric constants are represented with the `ConstantFP` class, which holds the numeric value in an `APFloat` internally (`APFloat` has the capability of holding floating point constants of Arbitrary Precision). This code basically just creates and returns a `ConstantFP`. Note that in the LLVM IR that constants are all unique together and shared. For this reason, the API uses “the `foo::get(..)`” idiom instead of “`new foo(..)`” or “`foo::Create(..)`”.

```
| Ast.Variable name ->
  (try Hashtbl.find named_values name with
   | Not_found -> raise (Error "unknown variable name"))
```

References to variables are also quite simple using LLVM. In the simple version of Kaleidoscope, we assume that the variable has already been emitted somewhere and its value is available. In practice, the only values that can be in the

Codegen.named_values map are function arguments. This code simply checks to see that the specified name is in the map (if not, an unknown variable is being referenced) and returns the value for it. In future chapters, we'll add support for [loop induction variables](#) in the symbol table, and for [local variables](#).

```
| Ast.Binary (op, lhs, rhs) ->
  let lhs_val = codegen_expr lhs in
  let rhs_val = codegen_expr rhs in
  begin
    match op with
    | '+' -> build_fadd lhs_val rhs_val "addtmp" builder
    | '-' -> build_fsub lhs_val rhs_val "subtmp" builder
    | '*' -> build_fmul lhs_val rhs_val "multmp" builder
    | '<' ->
      (* Convert bool 0/1 to double 0.0 or 1.0 *)
      let i = build_fcmp Fcmp.Ult lhs_val rhs_val "cmptmp" builder in
      build_uitofp i double_type "booltmp" builder
    | _ -> raise (Error "invalid binary operator")
  end
```

Binary operators start to get more interesting. The basic idea here is that we recursively emit code for the left-hand side of the expression, then the right-hand side, then we compute the result of the binary expression. In this code, we do a simple switch on the opcode to create the right LLVM instruction.

In the example above, the LLVM builder class is starting to show its value. IRBuilder knows where to insert the newly created instruction, all you have to do is specify what instruction to create (e.g. with `Llvm.create_add`), which operands to use (lhs and rhs here) and optionally provide a name for the generated instruction.

One nice thing about LLVM is that the name is just a hint. For instance, if the code above emits multiple “addtmp” variables, LLVM will automatically provide each one with an increasing, unique numeric suffix. Local value names for instructions are purely optional, but it makes it much easier to read the IR dumps.

[LLVM instructions](#) are constrained by strict rules: for example, the Left and Right operators of an [add instruction](#) must have the same type, and the result type of the add must match the operand types. Because all values in Kaleidoscope are doubles, this makes for very simple code for add, sub and mul.

On the other hand, LLVM specifies that the [fcmp instruction](#) always returns an ‘i1’ value (a one bit integer). The problem with this is that Kaleidoscope wants the value to be a 0.0 or 1.0 value. In order to get these semantics, we combine the fcmp instruction with a [uitofp instruction](#). This instruction converts its input integer into a floating point value by treating the input as an unsigned value. In contrast, if we used the [sitofp instruction](#), the Kaleidoscope ‘<’ operator would return 0.0 and -1.0, depending on the input value.

```
| Ast.Call (callee, args) ->
  (* Look up the name in the module table. *)
  let callee =
    match lookup_function callee the_module with
    | Some callee -> callee
    | None -> raise (Error "unknown function referenced")
  in
  let params = params callee in

  (* If argument mismatch error. *)
  if Array.length params == Array.length args then () else
    raise (Error "incorrect # arguments passed");
  let args = Array.map codegen_expr args in
  build_call callee args "calltmp" builder
```

Code generation for function calls is quite straightforward with LLVM. The code above initially does a function name lookup in the LLVM Module’s symbol table. Recall that the LLVM Module is the container that holds all of the

functions we are JIT'ing. By giving each function the same name as what the user specifies, we can use the LLVM symbol table to resolve function names for us.

Once we have the function to call, we recursively codegen each argument that is to be passed in, and create an LLVM [call instruction](#). Note that LLVM uses the native C calling conventions by default, allowing these calls to also call into standard library functions like “sin” and “cos”, with no additional effort.

This wraps up our handling of the four basic expressions that we have so far in Kaleidoscope. Feel free to go in and add some more. For example, by browsing the [LLVM language reference](#) you'll find several other interesting instructions that are really easy to plug into our basic framework.

Function Code Generation

Code generation for prototypes and functions must handle a number of details, which make their code less beautiful than expression code generation, but allows us to illustrate some important points. First, let's talk about code generation for prototypes: they are used both for function bodies and external function declarations. The code starts with:

```
let codegen_proto = function
| Ast.Prototype (name, args) ->
    (* Make the function type: double(double,double) etc. *)
    let doubles = Array.make (Array.length args) double_type in
    let ft = function_type double_type doubles in
    let f =
        match lookup_function name the_module with
```

This code packs a lot of power into a few lines. Note first that this function returns a “Function*” instead of a “Value*” (although at the moment they both are modeled by `llvalue` in `ocaml`). Because a “prototype” really talks about the external interface for a function (not the value computed by an expression), it makes sense for it to return the LLVM Function it corresponds to when codegen'd.

The call to `Llvm.function_type` creates the `Llvm.llvalue` that should be used for a given Prototype. Since all function arguments in Kaleidoscope are of type `double`, the first line creates a vector of “N” LLVM double types. It then uses the `Llvm.function_type` method to create a function type that takes “N” doubles as arguments, returns one double as a result, and that is not `vararg` (that uses the function `Llvm.var_arg_function_type`). Note that Types in LLVM are unique just like `Constant`'s are, so you don't “new” a type, you “get” it.

The final line above checks if the function has already been defined in `Codegen.the_module`. If not, we will create it.

```
| None -> declare_function name ft the_module
```

This indicates the type and name to use, as well as which module to insert into. By default we assume a function has `Llvm.Linkage.ExternalLinkage`. “external linkage” means that the function may be defined outside the current module and/or that it is callable by functions outside the module. The “name” passed in is the name the user specified: this name is registered in “`Codegen.the_module`”'s symbol table, which is used by the function call code above.

In Kaleidoscope, I choose to allow redefinitions of functions in two cases: first, we want to allow ‘extern’ing a function more than once, as long as the prototypes for the externs match (since all arguments have the same type, we just have to check that the number of arguments match). Second, we want to allow ‘extern’ing a function and then defining a body for it. This is useful when defining mutually recursive functions.

```
(* If 'f' conflicted, there was already something named 'name'. If it
 * has a body, don't allow redefinition or reextern. *)
| Some f ->
    (* If 'f' already has a body, reject this. *)
    if Array.length (basic_blocks f) == 0 then () else
```

```

    raise (Error "redefinition of function");

    (* If 'f' took a different number of arguments, reject. *)
    if Array.length (params f) == Array.length args then () else
        raise (Error "redefinition of function with different # args");
    f
in

```

In order to verify the logic above, we first check to see if the pre-existing function is “empty”. In this case, empty means that it has no basic blocks in it, which means it has no body. If it has no body, it is a forward declaration. Since we don’t allow anything after a full definition of the function, the code rejects this case. If the previous reference to a function was an ‘extern’, we simply verify that the number of arguments for that definition and this one match up. If not, we emit an error.

```

(* Set names for all arguments. *)
Array.iteri (fun i a ->
    let n = args.(i) in
    set_value_name n a;
    Hashtbl.add named_values n a;
) (params f);
f

```

The last bit of code for prototypes loops over all of the arguments in the function, setting the name of the LLVM Argument objects to match, and registering the arguments in the `Codegen.named_values` map for future use by the `Ast.Variable` variant. Once this is set up, it returns the Function object to the caller. Note that we don’t check for conflicting argument names here (e.g. “extern foo(a b a)”). Doing so would be very straight-forward with the mechanics we have already used above.

```

let codegen_func = function
| Ast.Function (proto, body) ->
    Hashtbl.clear named_values;
    let the_function = codegen_proto proto in

```

Code generation for function definitions starts out simply enough: we just codegen the prototype (Proto) and verify that it is ok. We then clear out the `Codegen.named_values` map to make sure that there isn’t anything in it from the last function we compiled. Code generation of the prototype ensures that there is an LLVM Function object that is ready to go for us.

```

(* Create a new basic block to start insertion into. *)
let bb = append_block context "entry" the_function in
position_at_end bb builder;

try
    let ret_val = codegen_expr body in

```

Now we get to the point where the `Codegen.builder` is set up. The first line creates a new basic block (named “entry”), which is inserted into `the_function`. The second line then tells the builder that new instructions should be inserted into the end of the new basic block. Basic blocks in LLVM are an important part of functions that define the Control Flow Graph. Since we don’t have any control flow, our functions will only contain one block at this point. We’ll fix this in Chapter 5 :).

```

let ret_val = codegen_expr body in

(* Finish off the function. *)
let _ = build_ret ret_val builder in

(* Validate the generated code, checking for consistency. *)

```

```
Llvm_analysis.assert_valid_function the_function;

the_function
```

Once the insertion point is set up, we call the `Codegen.codegen_func` method for the root expression of the function. If no error happens, this emits code to compute the expression into the entry block and returns the value that was computed. Assuming no error, we then create an LLVM [ret instruction](#), which completes the function. Once the function is built, we call `Llvm_analysis.assert_valid_function`, which is provided by LLVM. This function does a variety of consistency checks on the generated code, to determine if our compiler is doing everything right. Using this is important: it can catch a lot of bugs. Once the function is finished and validated, we return it.

```
with e ->
    delete_function the_function;
    raise e
```

The only piece left here is handling of the error case. For simplicity, we handle this by merely deleting the function we produced with the `Llvm.delete_function` method. This allows the user to redefine a function that they incorrectly typed in before: if we didn't delete it, it would live in the symbol table, with a body, preventing future redefinition.

This code does have a bug, though. Since the `Codegen.codegen_proto` can return a previously defined forward declaration, our code can actually delete a forward declaration. There are a number of ways to fix this bug, see what you can come up with! Here is a testcase:

```
extern foo(a b);      # ok, defines foo.
def foo(a b) c;        # error, 'c' is invalid.
def bar() foo(1, 2);  # error, unknown function "foo"
```

Driver Changes and Closing Thoughts

For now, code generation to LLVM doesn't really get us much, except that we can look at the pretty IR calls. The sample code inserts calls to `Codegen` into the `"Toplevel.main_loop"`, and then dumps out the LLVM IR. This gives a nice way to look at the LLVM IR for simple functions. For example:

```
ready> 4+5;
Read top-level expression:
define double @"()" {
entry:
    %addtmp = fadd double 4.000000e+00, 5.000000e+00
    ret double %addtmp
}
```

Note how the parser turns the top-level expression into anonymous functions for us. This will be handy when we add [JIT support](#) in the next chapter. Also note that the code is very literally transcribed, no optimizations are being performed. We will [add optimizations](#) explicitly in the next chapter.

```
ready> def foo(a b) a*a + 2*a*b + b*b;
Read function definition:
define double @foo(double %a, double %b) {
entry:
    %multmp = fmul double %a, %a
    %multmp1 = fmul double 2.000000e+00, %a
    %multmp2 = fmul double %multmp1, %b
    %addtmp = fadd double %multmp, %multmp2
    %multmp3 = fmul double %b, %b
```

```

    %addtmp4 = fadd double %addtmp, %multmp3
    ret double %addtmp4
}

```

This shows some simple arithmetic. Notice the striking similarity to the LLVM builder calls that we use to create the instructions.

```

ready> def bar(a) foo(a, 4.0) + bar(31337);
Read function definition:
define double @bar(double %a) {
entry:
    %calltmp = call double @foo(double %a, double 4.000000e+00)
    %calltmp1 = call double @bar(double 3.133700e+04)
    %addtmp = fadd double %calltmp, %calltmp1
    ret double %addtmp
}

```

This shows some function calls. Note that this function will take a long time to execute if you call it. In the future we'll add conditional control flow to actually make recursion useful :).

```

ready> extern cos(x);
Read extern:
declare double @cos(double)

ready> cos(1.234);
Read top-level expression:
define double @"()" {
entry:
    %calltmp = call double @cos(double 1.234000e+00)
    ret double %calltmp
}

```

This shows an extern for the libm “cos” function, and a call to it.

```

ready> ^D
; ModuleID = 'my cool jit'

define double @"()" {
entry:
    %addtmp = fadd double 4.000000e+00, 5.000000e+00
    ret double %addtmp
}

define double @foo(double %a, double %b) {
entry:
    %multmp = fmul double %a, %a
    %multmp1 = fmul double 2.000000e+00, %a
    %multmp2 = fmul double %multmp1, %b
    %addtmp = fadd double %multmp, %multmp2
    %multmp3 = fmul double %b, %b
    %addtmp4 = fadd double %addtmp, %multmp3
    ret double %addtmp4
}

define double @bar(double %a) {
entry:
    %calltmp = call double @foo(double %a, double 4.000000e+00)
    %calltmp1 = call double @bar(double 3.133700e+04)
}

```

```
        %addtmp = fadd double %calltmp, %calltmp1
        ret double %addtmp
    }

declare double @cos(double)

define double @"()"() {
entry:
    %calltmp = call double @cos(double 1.234000e+00)
    ret double %calltmp
}
```

When you quit the current demo, it dumps out the IR for the entire module generated. Here you can see the big picture with all the functions referencing each other.

This wraps up the third chapter of the Kaleidoscope tutorial. Up next, we'll describe how to [add JIT codegen and optimizer support](#) to this so we can actually start running code!

Full Code Listing

Here is the complete code listing for our running example, enhanced with the LLVM code generator. Because this uses the LLVM libraries, we need to link them in. To do this, we use the [llvm-config](#) tool to inform our makefile/command line about which options to use:

```
# Compile
ocamlbuild toy.byte
# Run
./toy.byte
```

Here is the code:

_tags:

```
<{lexer,parser}.ml>: use_camlp4, pp(camlp4of)
<*. {byte,native}>: g++, use_llvm, use_llvm_analysis
```

myocamlbuild.ml:

```
open Ocamlbuild_plugin;;

ocaml_lib ~extern:true "llvm";;
ocaml_lib ~extern:true "llvm_analysis";;

flag ["link"; "ocaml"; "g++"] (S[A"-cc"; A"g++"]);;
```

token.ml:

```
(=====
 * Lexer Tokens
 *=====)

(* The lexer returns these 'Kwd' if it is an unknown character, otherwise one of
 * these others for known things. *)
type token =
  (* commands *)
  | Def | Extern
```



```

(* primary *)
| Ident of string | Number of float

(* unknown *)
| Kwd of char

```

lexer.ml:

```

(=====
* Lexer
=====*)

let rec lex = parser
  (* Skip any whitespace. *)
  | [< ' (' | '\n' | '\r' | '\t'); stream >] -> lex stream

  (* identifier: [a-zA-Z][a-zA-Z0-9] *)
  | [< ' ('A' .. 'Z' | 'a' .. 'z' as c); stream >] ->
    let buffer = Buffer.create 1 in
    Buffer.add_char buffer c;
    lex_ident buffer stream

  (* number: [0-9.]+ *)
  | [< ' ('0' .. '9' as c); stream >] ->
    let buffer = Buffer.create 1 in
    Buffer.add_char buffer c;
    lex_number buffer stream

  (* Comment until end of line. *)
  | [< ' ('#'); stream >] ->
    lex_comment stream

  (* Otherwise, just return the character as its ascii value. *)
  | [< 'c; stream >] ->
    [< 'Token.Kwd c; lex stream >]

  (* end of stream. *)
  | [< >] -> [< >]

and lex_number buffer = parser
  | [< ' ('0' .. '9' | '.' as c); stream >] ->
    Buffer.add_char buffer c;
    lex_number buffer stream
  | [< stream=lex >] ->
    [< 'Token.Number (float_of_string (Buffer.contents buffer)); stream >]

and lex_ident buffer = parser
  | [< ' ('A' .. 'Z' | 'a' .. 'z' | '0' .. '9' as c); stream >] ->
    Buffer.add_char buffer c;
    lex_ident buffer stream
  | [< stream=lex >] ->
    match Buffer.contents buffer with
    | "def" -> [< 'Token.Def; stream >]
    | "extern" -> [< 'Token.Extern; stream >]
    | id -> [< 'Token.Ident id; stream >]

and lex_comment = parser
  | [< ' ('\n'); stream=lex >] -> stream

```

```
| [< 'c; e=lex_comment >] -> e
| [< >] -> [< >]
```

ast.ml:

```
(=====
 * Abstract Syntax Tree (aka Parse Tree)
 *=====)

(* expr - Base type for all expression nodes. *)
type expr =
  (* variant for numeric literals like "1.0". *)
  | Number of float

  (* variant for referencing a variable, like "a". *)
  | Variable of string

  (* variant for a binary operator. *)
  | Binary of char * expr * expr

  (* variant for function calls. *)
  | Call of string * expr array

(* proto - This type represents the "prototype" for a function, which captures
 * its name, and its argument names (thus implicitly the number of arguments the
 * function takes). *)
type proto = Prototype of string * string array

(* func - This type represents a function definition itself. *)
type func = Function of proto * expr
```

parser.ml:

```
(=====
 * Parser
 *=====)

(* binop_precedence - This holds the precedence for each binary operator that is
 * defined *)
let binop_precedence: (char, int) Hashtbl.t = Hashtbl.create 10

(* precedence - Get the precedence of the pending binary operator token. *)
let precedence c = try Hashtbl.find binop_precedence c with Not_found -> -1

(* primary
 * ::= identifier
 * ::= numberexpr
 * ::= parenexpr *)
let rec parse_primary = parser
  (* numberexpr ::= number *)
  | [< 'Token.Number n >] -> Ast.Number n

  (* parenexpr ::= '(' expression ')' *)
  | [< 'Token.Kwd '('; e=parse_expr; 'Token.Kwd ')' ?? "expected ')" >] -> e

  (* identifierexpr
   * ::= identifier
   * ::= identifier '(' argumentexpr ')' *)
```

```

| [< 'Token.Ident id; stream >] ->
  let rec parse_args accumulator = parser
    | [< e=parse_expr; stream >] ->
      begin parser
        | [< 'Token.Kwd ',', e=parse_args (e :: accumulator) >] -> e
        | [< >] -> e :: accumulator
      end stream
    | [< >] -> accumulator
  in
  let rec parse_ident id = parser
    (* Call. *)
    | [< 'Token.Kwd '(';
      args=parse_args [];
      'Token.Kwd ')' ?? "expected ')'">] ->
      Ast.Call (id, Array.of_list (List.rev args))

    (* Simple variable ref. *)
    | [< >] -> Ast.Variable id
  in
  parse_ident id stream

| [< >] -> raise (Stream.Error "unknown token when expecting an expression.")

(* binoprhs
 * ::= ('+' primary)* *)
and parse_bin_rhs expr_prec lhs stream =
  match Stream.peek stream with
  (* If this is a binop, find its precedence. *)
  | Some (Token.Kwd c) when Hashtbl.mem binop_precedence c ->
    let token_prec = precedence c in
    (* If this is a binop that binds at least as tightly as the current binop,
     * consume it, otherwise we are done. *)
    if token_prec <= expr_prec then lhs else begin
      (* Eat the binop. *)
      Stream.junk stream;

      (* Parse the primary expression after the binary operator. *)
      let rhs = parse_primary stream in

      (* Okay, we know this is a binop. *)
      let rhs =
        match Stream.peek stream with
        | Some (Token.Kwd c2) ->
          (* If BinOp binds less tightly with rhs than the operator after
           * rhs, let the pending operator take rhs as its lhs. *)
          let next_prec = precedence c2 in
          if token_prec < next_prec
          then parse_bin_rhs (token_prec + 1) rhs stream
          else rhs
        | _ -> rhs
      in

      (* Merge lhs/rhs. *)
      let lhs = Ast.Binary (c, lhs, rhs) in
      parse_bin_rhs expr_prec lhs stream
    end
  | _ -> lhs

```

```

(* expression
 * ::= primary binoprhs *)
and parse_expr = parser
| [< lhs=parse_primary; stream >] -> parse_bin_rhs 0 lhs stream

(* prototype
 * ::= id '(' id* ')' *)
let parse_prototype =
  let rec parse_args accumulator = parser
    | [< 'Token.Ident' id; e=parse_args (id::accumulator) >] -> e
    | [< >] -> accumulator
  in

  parser
  | [< 'Token.Ident' id;
    'Token.Kwd' '(' ?? "expected '(' in prototype";
    args=parse_args [];
    'Token.Kwd' ')' ?? "expected ')' in prototype" >] ->
    (* success. *)
    Ast.Prototype (id, Array.of_list (List.rev args))

  | [< >] ->
    raise (Stream.Error "expected function name in prototype")

(* definition ::= 'def' prototype expression *)
let parse_definition = parser
| [< 'Token.Def'; p=parse_prototype; e=parse_expr >] ->
  Ast.Function (p, e)

(* toplevelexpr ::= expression *)
let parse_toplevel = parser
| [< e=parse_expr >] ->
  (* Make an anonymous proto. *)
  Ast.Function (Ast.Prototype ("", []), e)

(* external ::= 'extern' prototype *)
let parse_extern = parser
| [< 'Token.Extern'; e=parse_prototype >] -> e

```

codegen.ml:

```

(=====
 * Code Generation
 *=====)

open Llvm

exception Error of string

let context = global_context ()
let the_module = create_module context "my cool jit"
let builder = builder context
let named_values:(string, llvalue) Hashtbl.t = Hashtbl.create 10
let double_type = double_type context

let rec codegen_expr = function
| Ast.Number n -> const_float double_type n

```

```

| Ast.Variable name ->
  (try Hashtbl.find named_values name with
   | Not_found -> raise (Error "unknown variable name"))
| Ast.Binary (op, lhs, rhs) ->
  let lhs_val = codegen_expr lhs in
  let rhs_val = codegen_expr rhs in
  begin
    match op with
    | '+' -> build_add lhs_val rhs_val "addtmp" builder
    | '-' -> build_sub lhs_val rhs_val "subtmp" builder
    | '*' -> build_mul lhs_val rhs_val "multmp" builder
    | '<' ->
      (* Convert bool 0/1 to double 0.0 or 1.0 *)
      let i = build_fcmp Fcmp.Ult lhs_val rhs_val "cmptmp" builder in
      build_uitofp i double_type "booltmp" builder
    | _ -> raise (Error "invalid binary operator")
  end
| Ast.Call (callee, args) ->
  (* Look up the name in the module table. *)
  let callee =
    match lookup_function callee the_module with
    | Some callee -> callee
    | None -> raise (Error "unknown function referenced")
  in
  let params = params callee in

  (* If argument mismatch error. *)
  if Array.length params == Array.length args then () else
    raise (Error "incorrect # arguments passed");
  let args = Array.map codegen_expr args in
  build_call callee args "calltmp" builder

let codegen_proto = function
| Ast.Prototype (name, args) ->
  (* Make the function type: double(double,double) etc. *)
  let doubles = Array.make (Array.length args) double_type in
  let ft = function_type double_type doubles in
  let f =
    match lookup_function name the_module with
    | None -> declare_function name ft the_module

    (* If 'f' conflicted, there was already something named 'name'. If it
     * has a body, don't allow redefinition or reextern. *)
    | Some f ->
      (* If 'f' already has a body, reject this. *)
      if block_begin f <> At_end f then
        raise (Error "redefinition of function");

      (* If 'f' took a different number of arguments, reject. *)
      if element_type (type_of f) <> ft then
        raise (Error "redefinition of function with different # args");
      f
  in

  (* Set names for all arguments. *)
  Array.iteri (fun i a ->
    let n = args.(i) in
    set_value_name n a;

```

```

    Hashtbl.add named_values n a;
  ) (params f);
  f

let codegen_func = function
| Ast.Function (proto, body) ->
  Hashtbl.clear named_values;
  let the_function = codegen_proto proto in

  (* Create a new basic block to start insertion into. *)
  let bb = append_block context "entry" the_function in
  position_at_end bb builder;

  try
    let ret_val = codegen_expr body in

    (* Finish off the function. *)
    let _ = build_ret ret_val builder in

    (* Validate the generated code, checking for consistency. *)
    Llvm_analysis.assert_valid_function the_function;

    the_function
  with e ->
    delete_function the_function;
    raise e

```

toplevel.ml:

```

(=====
 * Top-Level parsing and JIT Driver
 *=====*)

open Llvm

(* top ::= definition | external | expression | ';' *)
let rec main_loop stream =
  match Stream.peek stream with
  | None -> ()

  (* ignore top-level semicolons. *)
  | Some (Token.Kwd ';') ->
    Stream.junk stream;
    main_loop stream

  | Some token ->
    begin
      try match token with
      | Token.Def ->
        let e = Parser.parse_definition stream in
        print_endline "parsed a function definition.";
        dump_value (Codegen.codegen_func e);
      | Token.Extern ->
        let e = Parser.parse_extern stream in
        print_endline "parsed an extern.";
        dump_value (Codegen.codegen_proto e);
      | _ ->
        (* Evaluate a top-level expression into an anonymous function. *)

```

```

    let e = Parser.parse_toplevel stream in
    print_endline "parsed a top-level expr";
    dump_value (Codegen.codegen_func e);
    with Stream.Error s | Codegen.Error s ->
      (* Skip token for error recovery. *)
      Stream.junk stream;
      print_endline s;
  end;
  print_string "ready> "; flush stdout;
  main_loop stream

```

toy.ml:

```

(*=====*)
* Main driver code.
*=====*)

open Llvm

let main () =
  (* Install standard binary operators.
   * 1 is the lowest precedence. *)
  Hashtbl.add Parser.binop_precedence '<' 10;
  Hashtbl.add Parser.binop_precedence '+' 20;
  Hashtbl.add Parser.binop_precedence '-' 20;
  Hashtbl.add Parser.binop_precedence '*' 40;      (* highest. *)

  (* Prime the first token. *)
  print_string "ready> "; flush stdout;
  let stream = Lexer.lex (Stream.of_channel stdin) in

  (* Run the main "interpreter loop" now. *)
  Toplevel.main_loop stream;

  (* Print out all the generated code. *)
  dump_module Codegen.the_module
;;

main ()

```

Next: Adding JIT and Optimizer Support

Kaleidoscope: Adding JIT and Optimizer Support

- *Chapter 4 Introduction*
- *Trivial Constant Folding*
- *LLVM Optimization Passes*
- *Adding a JIT Compiler*
- *Full Code Listing*

Chapter 4 Introduction

Welcome to Chapter 4 of the “[Implementing a language with LLVM](#)” tutorial. Chapters 1-3 described the implementation of a simple language and added support for generating LLVM IR. This chapter describes two new techniques: adding optimizer support to your language, and adding JIT compiler support. These additions will demonstrate how to get nice, efficient code for the Kaleidoscope language.

Trivial Constant Folding

Note: the default `IRBuilder` now always includes the constant folding optimisations below.

Our demonstration for Chapter 3 is elegant and easy to extend. Unfortunately, it does not produce wonderful code. For example, when compiling simple code, we don’t get obvious optimizations:

```
ready> def test(x) 1+2+x;
Read function definition:
define double @test(double %x) {
entry:
    %addtmp = fadd double 1.000000e+00, 2.000000e+00
    %addtmp1 = fadd double %addtmp, %x
    ret double %addtmp1
}
```

This code is a very, very literal transcription of the AST built by parsing the input. As such, this transcription lacks optimizations like constant folding (we’d like to get “add x, 3.0” in the example above) as well as other more important optimizations. Constant folding, in particular, is a very common and very important optimization: so much so that many language implementors implement constant folding support in their AST representation.

With LLVM, you don’t need this support in the AST. Since all calls to build LLVM IR go through the LLVM builder, it would be nice if the builder itself checked to see if there was a constant folding opportunity when you call it. If so, it could just do the constant fold and return the constant instead of creating an instruction. This is exactly what the `LLVMFoldingBuilder` class does.

All we did was switch from `LLVMBuilder` to `LLVMFoldingBuilder`. Though we change no other code, we now have all of our instructions implicitly constant folded without us having to do anything about it. For example, the input above now compiles to:

```
ready> def test(x) 1+2+x;
Read function definition:
define double @test(double %x) {
entry:
    %addtmp = fadd double 3.000000e+00, %x
    ret double %addtmp
}
```

Well, that was easy :). In practice, we recommend always using `LLVMFoldingBuilder` when generating code like this. It has no “syntactic overhead” for its use (you don’t have to uglify your compiler with constant checks everywhere) and it can dramatically reduce the amount of LLVM IR that is generated in some cases (particular for languages with a macro preprocessor or that use a lot of constants).

On the other hand, the `LLVMFoldingBuilder` is limited by the fact that it does all of its analysis inline with the code as it is built. If you take a slightly more complex example:

```
ready> def test(x) (1+2+x) * (x + (1+2));
ready> Read function definition:
define double @test(double %x) {
entry:
```



```

    %addtmp = fadd double 3.000000e+00, %x
    %addtmp1 = fadd double %x, 3.000000e+00
    %multmp = fmul double %addtmp, %addtmp1
    ret double %multmp
}

```

In this case, the LHS and RHS of the multiplication are the same value. We’d really like to see this generate “tmp = x+3; result = tmp*tmp;” instead of computing “x*3” twice.

Unfortunately, no amount of local analysis will be able to detect and correct this. This requires two transformations: reassociation of expressions (to make the add’s lexically identical) and Common Subexpression Elimination (CSE) to delete the redundant add instruction. Fortunately, LLVM provides a broad range of optimizations that you can use, in the form of “passes”.

LLVM Optimization Passes

LLVM provides many optimization passes, which do many different sorts of things and have different tradeoffs. Unlike other systems, LLVM doesn’t hold to the mistaken notion that one set of optimizations is right for all languages and for all situations. LLVM allows a compiler implementor to make complete decisions about what optimizations to use, in which order, and in what situation.

As a concrete example, LLVM supports both “whole module” passes, which look across as large of body of code as they can (often a whole file, but if run at link time, this can be a substantial portion of the whole program). It also supports and includes “per-function” passes which just operate on a single function at a time, without looking at other functions. For more information on passes and how they are run, see the [How to Write a Pass](#) document and the [List of LLVM Passes](#).

For Kaleidoscope, we are currently generating functions on the fly, one at a time, as the user types them in. We aren’t shooting for the ultimate optimization experience in this setting, but we also want to catch the easy and quick stuff where possible. As such, we will choose to run a few per-function optimizations as the user types the function in. If we wanted to make a “static Kaleidoscope compiler”, we would use exactly the code we have now, except that we would defer running the optimizer until the entire file has been parsed.

In order to get per-function optimizations going, we need to set up a `Llvm.PassManager` to hold and organize the LLVM optimizations that we want to run. Once we have that, we can add a set of optimizations to run. The code looks like this:

```

(* Create the JIT. *)
let the_execution_engine = ExecutionEngine.create Codegen.the_module in
let the_fpm = PassManager.create_function Codegen.the_module in

(* Set up the optimizer pipeline. Start with registering info about how the
 * target lays out data structures. *)
DataLayout.add (ExecutionEngine.target_data the_execution_engine) the_fpm;

(* Do simple "peephole" optimizations and bit-twiddling optzn. *)
add_instruction_combining the_fpm;

(* reassociate expressions. *)
add_reassociation the_fpm;

(* Eliminate Common SubExpressions. *)
add_gvn the_fpm;

(* Simplify the control flow graph (deleting unreachable blocks, etc). *)
add_cfg_simplification the_fpm;

```

```
ignore (PassManager.initialize the_fpm);

(* Run the main "interpreter loop" now. *)
Toplevel.main_loop the_fpm the_execution_engine stream;
```

The meat of the matter here, is the definition of “the_fpm”. It requires a pointer to the the_module to construct itself. Once it is set up, we use a series of “add” calls to add a bunch of LLVM passes. The first pass is basically boilerplate, it adds a pass so that later optimizations know how the data structures in the program are laid out. The “the_execution_engine” variable is related to the JIT, which we will get to in the next section.

In this case, we choose to add 4 optimization passes. The passes we chose here are a pretty standard set of “cleanup” optimizations that are useful for a wide variety of code. I won’t delve into what they do but, believe me, they are a good starting place :).

Once the `Llvm.PassManager.` is set up, we need to make use of it. We do this by running it after our newly created function is constructed (in `Codegen.codegen_func`), but before it is returned to the client:

```
let codegen_func the_fpm = function
  ...
  try
    let ret_val = codegen_expr body in

    (* Finish off the function. *)
    let _ = build_ret ret_val builder in

    (* Validate the generated code, checking for consistency. *)
    Llvm_analysis.assert_valid_function the_function;

    (* Optimize the function. *)
    let _ = PassManager.run_function the_function the_fpm in

    the_function
```

As you can see, this is pretty straightforward. The `the_fpm` optimizes and updates the LLVM Function* in place, improving (hopefully) its body. With this in place, we can try our test above again:

```
ready> def test(x) (1+2+x)*(x+(1+2));
ready> Read function definition:
define double @test(double %x) {
entry:
    %addtmp = fadd double %x, 3.000000e+00
    %multmp = fmul double %addtmp, %addtmp
    ret double %multmp
}
```

As expected, we now get our nicely optimized code, saving a floating point add instruction from every execution of this function.

LLVM provides a wide variety of optimizations that can be used in certain circumstances. Some [documentation about the various passes](#) is available, but it isn’t very complete. Another good source of ideas can come from looking at the passes that Clang runs to get started. The “opt” tool allows you to experiment with passes from the command line, so you can see if they do anything.

Now that we have reasonable code coming out of our front-end, lets talk about executing it!

Adding a JIT Compiler

Code that is available in LLVM IR can have a wide variety of tools applied to it. For example, you can run optimizations on it (as we did above), you can dump it out in textual or binary forms, you can compile the code to an assembly file (.s) for some target, or you can JIT compile it. The nice thing about the LLVM IR representation is that it is the “common currency” between many different parts of the compiler.

In this section, we’ll add JIT compiler support to our interpreter. The basic idea that we want for Kaleidoscope is to have the user enter function bodies as they do now, but immediately evaluate the top-level expressions they type in. For example, if they type in “1 + 2;”, we should evaluate and print out 3. If they define a function, they should be able to call it from the command line.

In order to do this, we first declare and initialize the JIT. This is done by adding a global variable and a call in `main`:

```
...
let main () =
  ...
  (* Create the JIT. *)
  let the_execution_engine = ExecutionEngine.create Codegen.the_module in
  ...
```

This creates an abstract “Execution Engine” which can be either a JIT compiler or the LLVM interpreter. LLVM will automatically pick a JIT compiler for you if one is available for your platform, otherwise it will fall back to the interpreter.

Once the `Llvm_executionengine.ExecutionEngine.t` is created, the JIT is ready to be used. There are a variety of APIs that are useful, but the simplest one is the “`Llvm_executionengine.ExecutionEngine.run_function`” function. This method JIT compiles the specified LLVM Function and returns a function pointer to the generated machine code. In our case, this means that we can change the code that parses a top-level expression to look like this:

```
(* Evaluate a top-level expression into an anonymous function. *)
let e = Parser.parse_toplevel stream in
print_endline "parsed a top-level expr";
let the_function = Codegen.codegen_func the_fpm e in
dump_value the_function;

(* JIT the function, returning a function pointer. *)
let result = ExecutionEngine.run_function the_function [||]
  the_execution_engine in

print_string "Evaluated to ";
print_float (GenericValue.as_float Codegen.double_type result);
print_newline ();
```

Recall that we compile top-level expressions into a self-contained LLVM function that takes no arguments and returns the computed double. Because the LLVM JIT compiler matches the native platform ABI, this means that you can just cast the result pointer to a function pointer of that type and call it directly. This means, there is no difference between JIT compiled code and native machine code that is statically linked into your application.

With just these two changes, lets see how Kaleidoscope works now!

```
ready> 4+5;
define double @"()" {
entry:
    ret double 9.000000e+00
}
```

```
Evaluated to 9.000000
```

Well this looks like it is basically working. The dump of the function shows the “no argument function that always returns double” that we synthesize for each top level expression that is typed in. This demonstrates very basic functionality, but can we do more?

```
ready> def testfunc(x y) x + y*2;
Read function definition:
define double @testfunc(double %x, double %y) {
entry:
    %multmp = fmul double %y, 2.000000e+00
    %addtmp = fadd double %multmp, %x
    ret double %addtmp
}

ready> testfunc(4, 10);
define double @"()" {
entry:
    %calltmp = call double @testfunc(double 4.000000e+00, double 1.000000e+01)
    ret double %calltmp
}

Evaluated to 24.000000
```

This illustrates that we can now call user code, but there is something a bit subtle going on here. Note that we only invoke the JIT on the anonymous functions that *call testfunc*, but we never invoked it on *testfunc* itself. What actually happened here is that the JIT scanned for all non-JIT'd functions transitively called from the anonymous function and compiled all of them before returning from *run_function*.

The JIT provides a number of other more advanced interfaces for things like freeing allocated machine code, rejit'ing functions to update them, etc. However, even with this simple code, we get some surprisingly powerful capabilities - check this out (I removed the dump of the anonymous functions, you should get the idea by now :) :

```
ready> extern sin(x);
Read extern:
declare double @sin(double)

ready> extern cos(x);
Read extern:
declare double @cos(double)

ready> sin(1.0);
Evaluated to 0.841471

ready> def foo(x) sin(x)*sin(x) + cos(x)*cos(x);
Read function definition:
define double @foo(double %x) {
entry:
    %calltmp = call double @sin(double %x)
    %multmp = fmul double %calltmp, %calltmp
    %calltmp2 = call double @cos(double %x)
    %multmp4 = fmul double %calltmp2, %calltmp2
    %addtmp = fadd double %multmp, %multmp4
    ret double %addtmp
}

ready> foo(4.0);
```

```
Evaluated to 1.000000
```

Whoa, how does the JIT know about `sin` and `cos`? The answer is surprisingly simple: in this example, the JIT started execution of a function and got to a function call. It realized that the function was not yet JIT compiled and invoked the standard set of routines to resolve the function. In this case, there is no body defined for the function, so the JIT ended up calling “`dlsym("sin")`” on the Kaleidoscope process itself. Since “`sin`” is defined within the JIT’s address space, it simply patches up calls in the module to call the `libm` version of `sin` directly.

The LLVM JIT provides a number of interfaces (look in the `llvm_executionengine.mli` file) for controlling how unknown functions get resolved. It allows you to establish explicit mappings between IR objects and addresses (useful for LLVM global variables that you want to map to static tables, for example), allows you to dynamically decide on the fly based on the function name, and even allows you to have the JIT compile functions lazily the first time they’re called.

One interesting application of this is that we can now extend the language by writing arbitrary C code to implement operations. For example, if we add:

```
/* putchar - putchar that takes a double and returns 0. */
extern "C"
double putchar(double X) {
    putchar((char)X);
    return 0;
}
```

Now we can produce simple output to the console by using things like: “`extern putchar(x); putchar(120);`”, which prints a lowercase ‘x’ on the console (120 is the ASCII code for ‘x’). Similar code could be used to implement file I/O, console input, and many other capabilities in Kaleidoscope.

This completes the JIT and optimizer chapter of the Kaleidoscope tutorial. At this point, we can compile a non-Turing-complete programming language, optimize and JIT compile it in a user-driven way. Next up we’ll look into [extending the language with control flow constructs](#), tackling some interesting LLVM IR issues along the way.

Full Code Listing

Here is the complete code listing for our running example, enhanced with the LLVM JIT and optimizer. To build this example, use:

```
# Compile
ocamlbuild toy.byte
# Run
./toy.byte
```

Here is the code:

`_tags:`

```
<{lexer,parser}.ml>: use_camlp4, pp(camlp4of)
<*. {byte,native}>: g++, use_llvm, use_llvm_analysis
<*. {byte,native}>: use_llvm_executionengine, use_llvm_target
<*. {byte,native}>: use_llvm_scalar_opts, use_bindings
```

`myocamlbuild.ml:`

```
open Ocamlbuild_plugin;;

ocaml_lib ~extern:true "llvm";;
ocaml_lib ~extern:true "llvm_analysis";;
```

```
ocaml_lib ~extern:true "llvm_executionengine";;
ocaml_lib ~extern:true "llvm_target";;
ocaml_lib ~extern:true "llvm_scalar_opts";;

flag ["link"; "ocaml"; "g++"] (S[A"-cc"; A"g++"]);;
dep ["link"; "ocaml"; "use_bindings"] ["bindings.o"];;
```

token.ml:

```
(=====)
* Lexer Tokens
*=====*)

(* The lexer returns these 'Kwd' if it is an unknown character, otherwise one of
 * these others for known things. *)
type token =
  (* commands *)
  | Def | Extern

  (* primary *)
  | Ident of string | Number of float

  (* unknown *)
  | Kwd of char
```

lexer.ml:

```
(=====)
* Lexer
*=====*)

let rec lex = parser
  (* Skip any whitespace. *)
  | [< ' (' | '\n' | '\r' | '\t'); stream >] -> lex stream

  (* identifier: [a-zA-Z][a-zA-Z0-9] *)
  | [< ' ('A' .. 'Z' | 'a' .. 'z' as c); stream >] ->
    let buffer = Buffer.create 1 in
    Buffer.add_char buffer c;
    lex_ident buffer stream

  (* number: [0-9.]+ *)
  | [< ' ('0' .. '9' as c); stream >] ->
    let buffer = Buffer.create 1 in
    Buffer.add_char buffer c;
    lex_number buffer stream

  (* Comment until end of line. *)
  | [< ' ('#'); stream >] ->
    lex_comment stream

  (* Otherwise, just return the character as its ascii value. *)
  | [< 'c; stream >] ->
    [< 'Token.Kwd c; lex stream >]

  (* end of stream. *)
  | [< >] -> [< >]
```

```

and lex_number buffer = parser
| [< ' ('0' .. '9' | '.' as c); stream >] ->
    Buffer.add_char buffer c;
    lex_number buffer stream
| [< stream=lex >] ->
    [< 'Token.Number (float_of_string (Buffer.contents buffer)); stream >]

and lex_ident buffer = parser
| [< ' ('A' .. 'Z' | 'a' .. 'z' | '0' .. '9' as c); stream >] ->
    Buffer.add_char buffer c;
    lex_ident buffer stream
| [< stream=lex >] ->
    match Buffer.contents buffer with
    | "def" -> [< 'Token.Def; stream >]
    | "extern" -> [< 'Token.Extern; stream >]
    | id -> [< 'Token.Ident id; stream >]

and lex_comment = parser
| [< ' ('\n'); stream=lex >] -> stream
| [< 'c; e=lex_comment >] -> e
| [< >] -> [< >]

```

ast.ml:

```

(*=====
 * Abstract Syntax Tree (aka Parse Tree)
 *=====*)

(* expr - Base type for all expression nodes. *)
type expr =
  (* variant for numeric literals like "1.0". *)
  | Number of float

  (* variant for referencing a variable, like "a". *)
  | Variable of string

  (* variant for a binary operator. *)
  | Binary of char * expr * expr

  (* variant for function calls. *)
  | Call of string * expr array

  (* proto - This type represents the "prototype" for a function, which captures
   * its name, and its argument names (thus implicitly the number of arguments the
   * function takes). *)
  type proto = Prototype of string * string array

  (* func - This type represents a function definition itself. *)
  type func = Function of proto * expr

```

parser.ml:

```

(*=====
 * Parser
 *=====*)

(* binop_precedence - This holds the precedence for each binary operator that is
 * defined *)

```

```

let binop_precedence: (char, int) Hashtbl.t = Hashtbl.create 10

(* precedence - Get the precedence of the pending binary operator token. *)
let precedence c = try Hashtbl.find binop_precedence c with Not_found -> -1

(* primary
 * ::= identifier
 * ::= numberexpr
 * ::= parenexpr *)
let rec parse_primary = parser
  (* numberexpr ::= number *)
  | [< 'Token.Number n >] -> Ast.Number n

  (* parenexpr ::= '(' expression ')' *)
  | [< 'Token.Kwd '('; e=parse_expr; 'Token.Kwd ')' ?? "expected ')'" >] -> e

  (* identifierexpr
   * ::= identifier
   * ::= identifier '(' argumentexpr ')' *)
  | [< 'Token.Ident id; stream >] ->
    let rec parse_args accumulator = parser
      | [< e=parse_expr; stream >] ->
        begin parser
          | [< 'Token.Kwd ','; e=parse_args (e :: accumulator) >] -> e
          | [< >] -> e :: accumulator
        end stream
      | [< >] -> accumulator
    in
    let rec parse_ident id = parser
      (* Call. *)
      | [< 'Token.Kwd '(';
        args=parse_args [];
        'Token.Kwd ')' ?? "expected ')'" >] ->
        Ast.Call (id, Array.of_list (List.rev args))

      (* Simple variable ref. *)
      | [< >] -> Ast.Variable id
    in
    parse_ident id stream

  | [< >] -> raise (Stream.Error "unknown token when expecting an expression.")

(* binoprhs
 * ::= ('+' primary)* *)
and parse_bin_rhs expr_prec lhs stream =
  match Stream.peek stream with
  (* If this is a binop, find its precedence. *)
  | Some (Token.Kwd c) when Hashtbl.mem binop_precedence c ->
    let token_prec = precedence c in
    (* If this is a binop that binds at least as tightly as the current binop,
     * consume it, otherwise we are done. *)
    if token_prec < expr_prec then lhs else begin
      (* Eat the binop. *)
      Stream.junk stream;

      (* Parse the primary expression after the binary operator. *)
      let rhs = parse_primary stream in

```



```

    (* Okay, we know this is a binop. *)
    let rhs =
      match Stream.peek stream with
      | Some (Token.Kwd c2) ->
        (* If BinOp binds less tightly with rhs than the operator after
         * rhs, let the pending operator take rhs as its lhs. *)
        let next_prec = precedence c2 in
        if token_prec < next_prec
        then parse_bin_rhs (token_prec + 1) rhs stream
        else rhs
      | _ -> rhs
    in

    (* Merge lhs/rhs. *)
    let lhs = Ast.Binary (c, lhs, rhs) in
    parse_bin_rhs expr_prec lhs stream
  end
| _ -> lhs

(* expression
 * ::= primary binoprhs *)
and parse_expr = parser
| [< lhs=parse_primary; stream >] -> parse_bin_rhs 0 lhs stream

(* prototype
 * ::= id '(' id* ')' *)
let parse_prototype =
  let rec parse_args accumulator = parser
    | [< 'Token.Ident id; e=parse_args (id::accumulator) >] -> e
    | [< >] -> accumulator
  in

  parser
  | [< 'Token.Ident id;
    'Token.Kwd '(' ?? "expected '(' in prototype";
    args=parse_args [];
    'Token.Kwd ')' ?? "expected ')' in prototype" >] ->
    (* success. *)
    Ast.Prototype (id, Array.of_list (List.rev args))

  | [< >] ->
    raise (Stream.Error "expected function name in prototype")

(* definition ::= 'def' prototype expression *)
let parse_definition = parser
| [< 'Token.Def; p=parse_prototype; e=parse_expr >] ->
  Ast.Function (p, e)

(* toplevelexpr ::= expression *)
let parse_toplevel = parser
| [< e=parse_expr >] ->
  (* Make an anonymous proto. *)
  Ast.Function (Ast.Prototype ("", []), e)

(* external ::= 'extern' prototype *)
let parse_extern = parser
| [< 'Token.Extern; e=parse_prototype >] -> e

```

codegen.ml:

```

(*=====*)
* Code Generation
*=====*)

open Llvml

exception Error of string

let context = global_context ()
let the_module = create_module context "my cool jit"
let builder = builder context
let named_values:(string, llvalue) Hashtbl.t = Hashtbl.create 10
let double_type = double_type context

let rec codegen_expr = function
| Ast.Number n -> const_float double_type n
| Ast.Variable name ->
    (try Hashtbl.find named_values name with
    | Not_found -> raise (Error "unknown variable name"))
| Ast.Binary (op, lhs, rhs) ->
    let lhs_val = codegen_expr lhs in
    let rhs_val = codegen_expr rhs in
    begin
        match op with
        | '+' -> build_add lhs_val rhs_val "addtmp" builder
        | '-' -> build_sub lhs_val rhs_val "subtmp" builder
        | '*' -> build_mul lhs_val rhs_val "multmp" builder
        | '<' ->
            (* Convert bool 0/1 to double 0.0 or 1.0 *)
            let i = build_fcmp Fcmp.Ult lhs_val rhs_val "cmptmp" builder in
            build_uitofp i double_type "booltmp" builder
        | _ -> raise (Error "invalid binary operator")
    end
| Ast.Call (callee, args) ->
    (* Look up the name in the module table. *)
    let callee =
        match lookup_function callee the_module with
        | Some callee -> callee
        | None -> raise (Error "unknown function referenced")
    in
    let params = params callee in

    (* If argument mismatch error. *)
    if Array.length params == Array.length args then () else
        raise (Error "incorrect # arguments passed");
    let args = Array.map codegen_expr args in
    build_call callee args "calltmp" builder

let codegen_proto = function
| Ast.Prototype (name, args) ->
    (* Make the function type: double(double,double) etc. *)
    let doubles = Array.make (Array.length args) double_type in
    let ft = function_type double_type doubles in
    let f =
        match lookup_function name the_module with
        | None -> declare_function name ft the_module

```

```

    (* If 'f' conflicted, there was already something named 'name'. If it
    * has a body, don't allow redefinition or reextern. *)
  | Some f ->
    (* If 'f' already has a body, reject this. *)
    if block_begin f <> At_end f then
      raise (Error "redefinition of function");

    (* If 'f' took a different number of arguments, reject. *)
    if element_type (type_of f) <> ft then
      raise (Error "redefinition of function with different # args");
    f
in

  (* Set names for all arguments. *)
  Array.iter1 (fun i a ->
    let n = args.(i) in
    set_value_name n a;
    Hashtbl.add named_values n a;
  ) (params f);
  f

let codegen_func the_fpm = function
| Ast.Function (proto, body) ->
  Hashtbl.clear named_values;
  let the_function = codegen_proto proto in

  (* Create a new basic block to start insertion into. *)
  let bb = append_block context "entry" the_function in
  position_at_end bb builder;

  try
    let ret_val = codegen_expr body in

    (* Finish off the function. *)
    let _ = build_ret ret_val builder in

    (* Validate the generated code, checking for consistency. *)
    Llvm_analysis.assert_valid_function the_function;

    (* Optimize the function. *)
    let _ = PassManager.run_function the_function the_fpm in

    the_function
  with e ->
    delete_function the_function;
    raise e

```

toplevel.ml:

```

(=====
 * Top-Level parsing and JIT Driver
 *=====)

open Llvm
open Llvm_executionengine

(* top ::= definition | external | expression | ';' *)
let rec main_loop the_fpm the_execution_engine stream =

```

```

match Stream.peek stream with
| None -> ()

(* ignore top-level semicolons. *)
| Some (Token.Kwd ';') ->
    Stream.junk stream;
    main_loop the_fpm the_execution_engine stream

| Some token ->
    begin
        try match token with
        | Token.Def ->
            let e = Parser.parse_definition stream in
            print_endline "parsed a function definition.";
            dump_value (Codegen.codegen_func the_fpm e);
        | Token.Extern ->
            let e = Parser.parse_extern stream in
            print_endline "parsed an extern.";
            dump_value (Codegen.codegen_proto e);
        | _ ->
            (* Evaluate a top-level expression into an anonymous function. *)
            let e = Parser.parse_toplevel stream in
            print_endline "parsed a top-level expr";
            let the_function = Codegen.codegen_func the_fpm e in
            dump_value the_function;

            (* JIT the function, returning a function pointer. *)
            let result = ExecutionEngine.run_function the_function [||]
                the_execution_engine in

            print_string "Evaluated to ";
            print_float (GenericValue.as_float Codegen.double_type result);
            print_newline ();
            with Stream.Error s | Codegen.Error s ->
                (* Skip token for error recovery. *)
                Stream.junk stream;
                print_endline s;
        end;
        print_string "ready> "; flush stdout;
        main_loop the_fpm the_execution_engine stream

```

toy.ml:

```

(*=====*)
* Main driver code.
*=====*)

open Llm
open Llm_executionengine
open Llm_target
open Llm_scalar_opts

let main () =
    ignore (initialize_native_target ());

    (* Install standard binary operators.
       * 1 is the lowest precedence. *)
    Hashtbl.add Parser.binop_precedence '<' 10;

```

```

Hashtbl.add Parser.binop_precedence '+' 20;
Hashtbl.add Parser.binop_precedence '-' 20;
Hashtbl.add Parser.binop_precedence '*' 40;      (* highest. *)

(* Prime the first token. *)
print_string "ready> "; flush stdout;
let stream = Lexer.lex (Stream.of_channel stdin) in

(* Create the JIT. *)
let the_execution_engine = ExecutionEngine.create Codegen.the_module in
let the_fpm = PassManager.create_function Codegen.the_module in

(* Set up the optimizer pipeline. Start with registering info about how the
 * target lays out data structures. *)
DataLayout.add (ExecutionEngine.target_data the_execution_engine) the_fpm;

(* Do simple "peephole" optimizations and bit-twiddling optzn. *)
add_instruction_combination the_fpm;

(* reassociate expressions. *)
add_reassociation the_fpm;

(* Eliminate Common SubExpressions. *)
add_gvn the_fpm;

(* Simplify the control flow graph (deleting unreachable blocks, etc). *)
add_cfg_simplification the_fpm;

ignore (PassManager.initialize the_fpm);

(* Run the main "interpreter loop" now. *)
Toplevel.main_loop the_fpm the_execution_engine stream;

(* Print out all the generated code. *)
dump_module Codegen.the_module
;;

main ()

```

bindings.c

```

#include <stdio.h>

/* putchar - putchar that takes a double and returns 0. */
extern double putchar(double X) {
    putchar((char)X);
    return 0;
}

```

Next: Extending the language: control flow

Kaleidoscope: Extending the Language: Control Flow

- *Chapter 5 Introduction*
- *If/Then/Else*
 - *Lexer Extensions for If/Then/Else*
 - *AST Extensions for If/Then/Else*
 - *Parser Extensions for If/Then/Else*
 - *LLVM IR for If/Then/Else*
 - *Code Generation for If/Then/Else*
- *‘for’ Loop Expression*
 - *Lexer Extensions for the ‘for’ Loop*
 - *AST Extensions for the ‘for’ Loop*
 - *Parser Extensions for the ‘for’ Loop*
 - *LLVM IR for the ‘for’ Loop*
 - *Code Generation for the ‘for’ Loop*
- *Full Code Listing*

Chapter 5 Introduction

Welcome to Chapter 5 of the “[Implementing a language with LLVM](#)” tutorial. Parts 1-4 described the implementation of the simple Kaleidoscope language and included support for generating LLVM IR, followed by optimizations and a JIT compiler. Unfortunately, as presented, Kaleidoscope is mostly useless: it has no control flow other than call and return. This means that you can’t have conditional branches in the code, significantly limiting its power. In this episode of “build that compiler”, we’ll extend Kaleidoscope to have an if/then/else expression plus a simple ‘for’ loop.

If/Then/Else

Extending Kaleidoscope to support if/then/else is quite straightforward. It basically requires adding lexer support for this “new” concept to the lexer, parser, AST, and LLVM code emitter. This example is nice, because it shows how easy it is to “grow” a language over time, incrementally extending it as new ideas are discovered.

Before we get going on “how” we add this extension, let’s talk about “what” we want. The basic idea is that we want to be able to write this sort of thing:

```
def fib(x)
  if x < 3 then
    1
  else
    fib(x-1)+fib(x-2);
```

In Kaleidoscope, every construct is an expression: there are no statements. As such, the if/then/else expression needs to return a value like any other. Since we’re using a mostly functional form, we’ll have it evaluate its conditional, then return the ‘then’ or ‘else’ value based on how the condition was resolved. This is very similar to the C “?:” expression.

The semantics of the if/then/else expression is that it evaluates the condition to a boolean equality value: 0.0 is considered to be false and everything else is considered to be true. If the condition is true, the first subexpression is evaluated and returned, if the condition is false, the second subexpression is evaluated and returned. Since Kaleidoscope allows side-effects, this behavior is important to nail down.

Now that we know what we “want”, let's break this down into its constituent pieces.

Lexer Extensions for If/Then/Else

The lexer extensions are straightforward. First we add new variants for the relevant tokens:

```
(* control *)
| If | Then | Else | For | In
```

Once we have that, we recognize the new keywords in the lexer. This is pretty simple stuff:

```
...
match Buffer.contents buffer with
| "def" -> [< 'Token.Def; stream >]
| "extern" -> [< 'Token.Extern; stream >]
| "if" -> [< 'Token.If; stream >]
| "then" -> [< 'Token.Then; stream >]
| "else" -> [< 'Token.Else; stream >]
| "for" -> [< 'Token.For; stream >]
| "in" -> [< 'Token.In; stream >]
| id -> [< 'Token.Ident id; stream >]
```

AST Extensions for If/Then/Else

To represent the new expression we add a new AST variant for it:

```
type expr =
...
(* variant for if/then/else. *)
| If of expr * expr * expr
```

The AST variant just has pointers to the various subexpressions.

Parser Extensions for If/Then/Else

Now that we have the relevant tokens coming from the lexer and we have the AST node to build, our parsing logic is relatively straightforward. Next we add a new case for parsing a if-expression as a primary expression:

```
let rec parse_primary = parser
...
(* ifexpr ::= 'if' expr 'then' expr 'else' expr *)
| [< 'Token.If; c=parse_expr;
   'Token.Then ?? "expected 'then'"; t=parse_expr;
   'Token.Else ?? "expected 'else'"; e=parse_expr >] ->
   Ast.If (c, t, e)
```

LLVM IR for If/Then/Else

Now that we have it parsing and building the AST, the final piece is adding LLVM code generation support. This is the most interesting part of the if/then/else example, because this is where it starts to introduce new concepts. All of the code above has been thoroughly described in previous chapters.

To motivate the code we want to produce, let's take a look at a simple example. Consider:

```
extern foo();
extern bar();
def baz(x) if x then foo() else bar();
```

If you disable optimizations, the code you'll (soon) get from Kaleidoscope looks like this:

```
declare double @foo()

declare double @bar()

define double @baz(double %x) {
entry:
    %ifcond = fcmp one double %x, 0.000000e+00
    br i1 %ifcond, label %then, label %else

then:    ; preds = %entry
    %calltmp = call double @foo()
    br label %ifcont

else:    ; preds = %entry
    %calltmp1 = call double @bar()
    br label %ifcont

ifcont:  ; preds = %else, %then
    %iftmp = phi double [ %calltmp, %then ], [ %calltmp1, %else ]
    ret double %iftmp
}
```

To visualize the control flow graph, you can use a nifty feature of the LLVM 'opt' tool. If you put this LLVM IR into "t.ll" and run "llvm-as < t.ll | opt -analyze -view-cfg", a window will pop up and you'll see this graph:

Another way to get this is to call "Llvm_analysis.view_function_cfg f" or "Llvm_analysis.view_function_cfg_only f" (where f is a "Function") either by inserting actual calls into the code and recompiling or by calling these in the debugger. LLVM has many nice features for visualizing various graphs.

Getting back to the generated code, it is fairly simple: the entry block evaluates the conditional expression ("x" in our case here) and compares the result to 0.0 with the "fcmp one" instruction ('one' is "Ordered and Not Equal"). Based on the result of this expression, the code jumps to either the "then" or "else" blocks, which contain the expressions for the true/false cases.

Once the then/else blocks are finished executing, they both branch back to the 'ifcont' block to execute the code that happens after the if/then/else. In this case the only thing left to do is to return to the caller of the function. The question then becomes: how does the code know which expression to return?

The answer to this question involves an important SSA operation: the [Phi operation](#). If you're not familiar with SSA, [the wikipedia article](#) is a good introduction and there are various other introductions to it available on your favorite search engine. The short version is that "execution" of the Phi operation requires "remembering" which block control came from. The Phi operation takes on the value corresponding to the input control block. In this case, if control comes in from the "then" block, it gets the value of "calltmp". If control comes from the "else" block, it gets the value of "calltmp1".

At this point, you are probably starting to think "Oh no! This means my simple and elegant front-end will have to start generating SSA form in order to use LLVM!". Fortunately, this is not the case, and we strongly advise *not* implementing an SSA construction algorithm in your front-end unless there is an amazingly good reason to do so. In practice, there are two sorts of values that float around in code written for your average imperative programming language that might need Phi nodes:

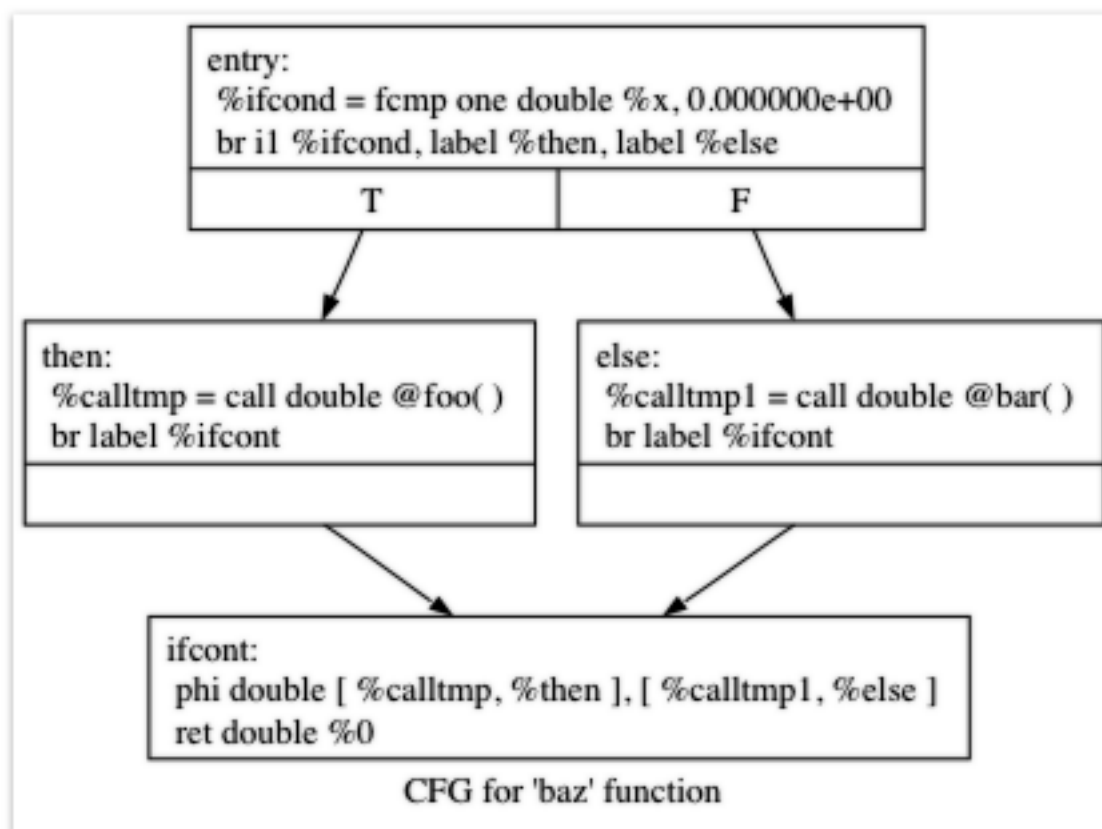


Fig. 2.2: Example CFG

1. Code that involves user variables: `x = 1; x = x + 1;`
2. Values that are implicit in the structure of your AST, such as the Phi node in this case.

In [Chapter 7](#) of this tutorial (“mutable variables”), we’ll talk about #1 in depth. For now, just believe me that you don’t need SSA construction to handle this case. For #2, you have the choice of using the techniques that we will describe for #1, or you can insert Phi nodes directly, if convenient. In this case, it is really really easy to generate the Phi node, so we choose to do it directly.

Okay, enough of the motivation and overview, lets generate code!

Code Generation for If/Then/Else

In order to generate code for this, we implement the Codegen method for `IfExprAST`:

```
let rec codegen_expr = function
...
| Ast.If (cond, then_, else_) ->
    let cond = codegen_expr cond in

    (Convert condition to a bool by comparing equal to 0.0 *)
    let zero = const_float double_type 0.0 in
    let cond_val = build_fcmp Fcmp.One cond zero "ifcond" builder in
```

This code is straightforward and similar to what we saw before. We emit the expression for the condition, then compare that value to zero to get a truth value as a 1-bit (bool) value.

```
(Grab the first block so that we might later add the conditional branch
 * to it at the end of the function. *)
let start_bb = insertion_block builder in
let the_function = block_parent start_bb in

let then_bb = append_block context "then" the_function in
position_at_end then_bb builder;
```

As opposed to the [C++ tutorial](#), we have to build our basic blocks bottom up since we can’t have dangling BasicBlocks. We start off by saving a pointer to the first block (which might not be the entry block), which we’ll need to build a conditional branch later. We do this by asking the builder for the current BasicBlock. The fourth line gets the current Function object that is being built. It gets this by the `start_bb` for its “parent” (the function it is currently embedded into).

Once it has that, it creates one block. It is automatically appended into the function’s list of blocks.

```
(Emit 'then' value. *)
position_at_end then_bb builder;
let then_val = codegen_expr then_ in

(Codegen of 'then' can change the current block, update then_bb for the
 * phi. We create a new name because one is used for the phi node, and the
 * other is used for the conditional branch. *)
let new_then_bb = insertion_block builder in
```

We move the builder to start inserting into the “then” block. Strictly speaking, this call moves the insertion point to be at the end of the specified block. However, since the “then” block is empty, it also starts out by inserting at the beginning of the block. :)

Once the insertion point is set, we recursively codegen the “then” expression from the AST.

The final line here is quite subtle, but is very important. The basic issue is that when we create the Phi node in the merge block, we need to set up the block/value pairs that indicate how the Phi will work. Importantly, the Phi node expects to have an entry for each predecessor of the block in the CFG. Why then, are we getting the current block when we just set it to ThenBB 5 lines above? The problem is that the “Then” expression may actually itself change the block that the Builder is emitting into if, for example, it contains a nested “if/then/else” expression. Because calling Codegen recursively could arbitrarily change the notion of the current block, we are required to get an up-to-date value for code that will set up the Phi node.

```
(* Emit 'else' value. *)
let else_bb = append_block context "else" the_function in
position_at_end else_bb builder;
let else_val = codegen_expr else_ in

(* Codegen of 'else' can change the current block, update else_bb for the
   * phi. *)
let new_else_bb = insertion_block builder in
```

Code generation for the ‘else’ block is basically identical to codegen for the ‘then’ block.

```
(* Emit merge block. *)
let merge_bb = append_block context "ifcont" the_function in
position_at_end merge_bb builder;
let incoming = [(then_val, new_then_bb); (else_val, new_else_bb)] in
let phi = build_phi incoming "iftmp" builder in
```

The first two lines here are now familiar: the first adds the “merge” block to the Function object. The second changes the insertion point so that newly created code will go into the “merge” block. Once that is done, we need to create the PHI node and set up the block/value pairs for the PHI.

```
(* Return to the start block to add the conditional branch. *)
position_at_end start_bb builder;
ignore (build_cond_br cond_val then_bb else_bb builder);
```

Once the blocks are created, we can emit the conditional branch that chooses between them. Note that creating new blocks does not implicitly affect the IRBuilder, so it is still inserting into the block that the condition went into. This is why we needed to save the “start” block.

```
(* Set a unconditional branch at the end of the 'then' block and the
   * 'else' block to the 'merge' block. *)
position_at_end new_then_bb builder; ignore (build_br merge_bb builder);
position_at_end new_else_bb builder; ignore (build_br merge_bb builder);

(* Finally, set the builder to the end of the merge block. *)
position_at_end merge_bb builder;

phi
```

To finish off the blocks, we create an unconditional branch to the merge block. One interesting (and very important) aspect of the LLVM IR is that it **requires all basic blocks to be “terminated”** with a **control flow instruction** such as return or branch. This means that all control flow, *including fall throughs* must be made explicit in the LLVM IR. If you violate this rule, the verifier will emit an error.

Finally, the CodeGen function returns the phi node as the value computed by the if/then/else expression. In our example above, this returned value will feed into the code for the top-level function, which will create the return instruction.

Overall, we now have the ability to execute conditional code in Kaleidoscope. With this extension, Kaleidoscope is a fairly complete language that can calculate a wide variety of numeric functions. Next up we’ll add another useful

expression that is familiar from non-functional languages...

‘for’ Loop Expression

Now that we know how to add basic control flow constructs to the language, we have the tools to add more powerful things. Lets add something more aggressive, a ‘for’ expression:

```
extern putchar(char);
def printstar(n)
  for i = 1, i < n, 1.0 in
    putchar(42); # ascii 42 = '*'

# print 100 '*' characters
printstar(100);
```

This expression defines a new variable (“i” in this case) which iterates from a starting value, while the condition (“i < n” in this case) is true, incrementing by an optional step value (“1.0” in this case). If the step value is omitted, it defaults to 1.0. While the loop is true, it executes its body expression. Because we don’t have anything better to return, we’ll just define the loop as always returning 0.0. In the future when we have mutable variables, it will get more useful.

As before, lets talk about the changes that we need to Kaleidoscope to support this.

Lexer Extensions for the ‘for’ Loop

The lexer extensions are the same sort of thing as for if/then/else:

```
... in Token.token ...
(* control *)
| If | Then | Else
| For | In

... in Lexer.lex_ident...
match Buffer.contents buffer with
| "def" -> [< 'Token.Def; stream >]
| "extern" -> [< 'Token.Extern; stream >]
| "if" -> [< 'Token.If; stream >]
| "then" -> [< 'Token.Then; stream >]
| "else" -> [< 'Token.Else; stream >]
| "for" -> [< 'Token.For; stream >]
| "in" -> [< 'Token.In; stream >]
| id -> [< 'Token.Ident id; stream >]
```

AST Extensions for the ‘for’ Loop

The AST variant is just as simple. It basically boils down to capturing the variable name and the constituent expressions in the node.

```
type expr =
...
(* variant for for/in. *)
| For of string * expr * expr * expr option * expr
```

Parser Extensions for the ‘for’ Loop

The parser code is also fairly standard. The only interesting thing here is handling of the optional step value. The parser code handles it by checking to see if the second comma is present. If not, it sets the step value to null in the AST node:

```
let rec parse_primary = parser
...
(* forexpr
   ::= 'for' identifier '=' expr ',' expr (',' expr)? 'in' expression *)
| [< 'Token.For;
   'Token.Ident id ?? "expected identifier after for";
   'Token.Kwd '=' ?? "expected '=' after for";
   stream >] ->
begin parser
  | [<
      start=parse_expr;
      'Token.Kwd ',' ?? "expected ',' after for";
      end_=parse_expr;
      stream >] ->
let step =
  begin parser
    | [< 'Token.Kwd ','; step=parse_expr >] -> Some step
    | [< >] -> None
  end stream
in
begin parser
  | [< 'Token.In; body=parse_expr >] ->
    Ast.For (id, start, end_, step, body)
  | [< >] ->
    raise (Stream.Error "expected 'in' after for")
  end stream
  | [< >] ->
    raise (Stream.Error "expected '=' after for")
end stream
```

LLVM IR for the ‘for’ Loop

Now we get to the good part: the LLVM IR we want to generate for this thing. With the simple example above, we get this LLVM IR (note that this dump is generated with optimizations disabled for clarity):

```
declare double @putchard(double)

define double @printstar(double %n) {
entry:
    ; initial value = 1.0 (inlined into phi)
    br label %loop

loop:    ; preds = %loop, %entry
    %i = phi double [ 1.000000e+00, %entry ], [ %nextvar, %loop ]
    ; body
    %calltmp = call double @putchard(double 4.200000e+01)
    ; increment
    %nextvar = fadd double %i, 1.000000e+00

    ; termination test
```

```
%cmptmp = fcmp ult double %i, %n
%booltmp = uitofp i1 %cmptmp to double
%loopcond = fcmp one double %booltmp, 0.000000e+00
br i1 %loopcond, label %loop, label %afterloop

afterloop:    ; preds = %loop
              ; loop always returns 0.0
  ret double 0.000000e+00
}
```

This loop contains all the same constructs we saw before: a phi node, several expressions, and some basic blocks. Lets see how this fits together.

Code Generation for the ‘for’ Loop

The first part of Codegen is very simple: we just output the start expression for the loop value:

```
let rec codegen_expr = function
...
| Ast.For (var_name, start, end_, step, body) ->
  (* Emit the start code first, without 'variable' in scope. *)
  let start_val = codegen_expr start in
```

With this out of the way, the next step is to set up the LLVM basic block for the start of the loop body. In the case above, the whole loop body is one block, but remember that the body code itself could consist of multiple blocks (e.g. if it contains an if/then/else or a for/in expression).

```
(* Make the new basic block for the loop header, inserting after current
 * block. *)
let preheader_bb = insertion_block builder in
let the_function = block_parent preheader_bb in
let loop_bb = append_block context "loop" the_function in

(* Insert an explicit fall through from the current block to the
 * loop_bb. *)
ignore (build_br loop_bb builder);
```

This code is similar to what we saw for if/then/else. Because we will need it to create the Phi node, we remember the block that falls through into the loop. Once we have that, we create the actual block that starts the loop and create an unconditional branch for the fall-through between the two blocks.

```
(* Start insertion in loop_bb. *)
position_at_end loop_bb builder;

(* Start the PHI node with an entry for start. *)
let variable = build_phi [(start_val, preheader_bb)] var_name builder in
```

Now that the “preheader” for the loop is set up, we switch to emitting code for the loop body. To begin with, we move the insertion point and create the PHI node for the loop induction variable. Since we already know the incoming value for the starting value, we add it to the Phi node. Note that the Phi will eventually get a second value for the backedge, but we can’t set it up yet (because it doesn’t exist!).

```
(* Within the loop, the variable is defined equal to the PHI node. If it
 * shadows an existing variable, we have to restore it, so save it
 * now. *)
let old_val =
```

```

    try Some (Hashtbl.find named_values var_name) with Not_found -> None
in
Hashtbl.add named_values var_name variable;

(* Emit the body of the loop. This, like any other expr, can change the
 * current BB. Note that we ignore the value computed by the body, but
 * don't allow an error *)
ignore (codegen_expr body);

```

Now the code starts to get more interesting. Our ‘for’ loop introduces a new variable to the symbol table. This means that our symbol table can now contain either function arguments or loop variables. To handle this, before we codegen the body of the loop, we add the loop variable as the current value for its name. Note that it is possible that there is a variable of the same name in the outer scope. It would be easy to make this an error (emit an error and return null if there is already an entry for VarName) but we choose to allow shadowing of variables. In order to handle this correctly, we remember the Value that we are potentially shadowing in `old_val` (which will be None if there is no shadowed variable).

Once the loop variable is set into the symbol table, the code recursively codegen’s the body. This allows the body to use the loop variable: any references to it will naturally find it in the symbol table.

```

(* Emit the step value. *)
let step_val =
  match step with
  | Some step -> codegen_expr step
  (* If not specified, use 1.0. *)
  | None -> const_float double_type 1.0
in
let next_var = build_add variable step_val "nextvar" builder in

```

Now that the body is emitted, we compute the next value of the iteration variable by adding the step value, or 1.0 if it isn’t present. ‘next_var’ will be the value of the loop variable on the next iteration of the loop.

```

(* Compute the end condition. *)
let end_cond = codegen_expr end_ in

(* Convert condition to a bool by comparing equal to 0.0. *)
let zero = const_float double_type 0.0 in
let end_cond = build_fcmp Fcmp.One end_cond zero "loopcond" builder in

```

Finally, we evaluate the exit value of the loop, to determine whether the loop should exit. This mirrors the condition evaluation for the if/then/else statement.

```

(* Create the "after loop" block and insert it. *)
let loop_end_bb = insertion_block builder in
let after_bb = append_block context "afterloop" the_function in

(* Insert the conditional branch into the end of loop_end_bb. *)
ignore (build_cond_br end_cond loop_bb after_bb builder);

(* Any new code will be inserted in after_bb. *)
position_at_end after_bb builder;

```

With the code for the body of the loop complete, we just need to finish up the control flow for it. This code remembers the end block (for the phi node), then creates the block for the loop exit (“afterloop”). Based on the value of the exit condition, it creates a conditional branch that chooses between executing the loop again and exiting the loop. Any future code is emitted in the “afterloop” block, so it sets the insertion position to it.

```
(* Add a new entry to the PHI node for the backedge. *)
add_incoming (next_var, loop_end_bb) variable;

(* Restore the unshadowed variable. *)
begin match old_val with
| Some old_val -> Hashtbl.add named_values var_name old_val
| None -> ()
end;

(* for expr always returns 0.0. *)
const_null double_type
```

The final code handles various cleanups: now that we have the “next_var” value, we can add the incoming value to the loop PHI node. After that, we remove the loop variable from the symbol table, so that it isn’t in scope after the for loop. Finally, code generation of the for loop always returns 0.0, so that is what we return from `Codegen.codegen_expr`.

With this, we conclude the “adding control flow to Kaleidoscope” chapter of the tutorial. In this chapter we added two control flow constructs, and used them to motivate a couple of aspects of the LLVM IR that are important for front-end implementors to know. In the next chapter of our saga, we will get a bit crazier and add [user-defined operators](#) to our poor innocent language.

Full Code Listing

Here is the complete code listing for our running example, enhanced with the if/then/else and for expressions.. To build this example, use:

```
# Compile
ocamlbuild toy.byte
# Run
./toy.byte
```

Here is the code:

_tags:

```
<{lexer,parser}.ml>: use_camlp4, pp(camlp4of)
<*. {byte,native}>: g++, use_llvm, use_llvm_analysis
<*. {byte,native}>: use_llvm_executionengine, use_llvm_target
<*. {byte,native}>: use_llvm_scalar_opts, use_bindings
```

myocamlbuild.ml:

```
open Ocamlbuild_plugin;;

ocaml_lib ~extern:true "llvm";;
ocaml_lib ~extern:true "llvm_analysis";;
ocaml_lib ~extern:true "llvm_executionengine";;
ocaml_lib ~extern:true "llvm_target";;
ocaml_lib ~extern:true "llvm_scalar_opts";;

flag ["link"; "ocaml"; "g++"] (S[A"-cc"; A"g++"]);;
dep ["link"; "ocaml"; "use_bindings"] ["bindings.o"];;
```

token.ml:


```

(=====
* Lexer Tokens
*=====)

(* The lexer returns these 'Kwd' if it is an unknown character, otherwise one of
* these others for known things. *)
type token =
  (* commands *)
  | Def | Extern

  (* primary *)
  | Ident of string | Number of float

  (* unknown *)
  | Kwd of char

  (* control *)
  | If | Then | Else
  | For | In

```

lexer.ml:

```

(=====
* Lexer
*=====)

let rec lex = parser
  (* Skip any whitespace. *)
  | [< ' (' ' ' | '\n' | '\r' | '\t'); stream >] -> lex stream

  (* identifier: [a-zA-Z][a-zA-Z0-9]* *)
  | [< ' ('A' .. 'Z' | 'a' .. 'z' as c); stream >] ->
    let buffer = Buffer.create 1 in
    Buffer.add_char buffer c;
    lex_ident buffer stream

  (* number: [0-9.]+ *)
  | [< ' ('0' .. '9' as c); stream >] ->
    let buffer = Buffer.create 1 in
    Buffer.add_char buffer c;
    lex_number buffer stream

  (* Comment until end of line. *)
  | [< ' ('#'); stream >] ->
    lex_comment stream

  (* Otherwise, just return the character as its ascii value. *)
  | [< 'c; stream >] ->
    [< 'Token.Kwd c; lex stream >]

  (* end of stream. *)
  | [< >] -> [< >]

and lex_number buffer = parser
  | [< ' ('0' .. '9' | '.' as c); stream >] ->
    Buffer.add_char buffer c;
    lex_number buffer stream
  | [< stream=lex >] ->

```

```

    [< 'Token.Number (float_of_string (Buffer.contents buffer)); stream >]

and lex_ident buffer = parser
| [< ' ('A' .. 'Z' | 'a' .. 'z' | '0' .. '9' as c); stream >] ->
    Buffer.add_char buffer c;
    lex_ident buffer stream
| [< stream=lex >] ->
    match Buffer.contents buffer with
    | "def" -> [< 'Token.Def; stream >]
    | "extern" -> [< 'Token.Extern; stream >]
    | "if" -> [< 'Token.If; stream >]
    | "then" -> [< 'Token.Then; stream >]
    | "else" -> [< 'Token.Else; stream >]
    | "for" -> [< 'Token.For; stream >]
    | "in" -> [< 'Token.In; stream >]
    | id -> [< 'Token.Ident id; stream >]

and lex_comment = parser
| [< ' ('\n'); stream=lex >] -> stream
| [< 'c; e=lex_comment >] -> e
| [< >] -> [< >]

```

ast.ml:

```

(*=====
 * Abstract Syntax Tree (aka Parse Tree)
 *=====*)

(* expr - Base type for all expression nodes. *)
type expr =
  (* variant for numeric literals like "1.0". *)
  | Number of float

  (* variant for referencing a variable, like "a". *)
  | Variable of string

  (* variant for a binary operator. *)
  | Binary of char * expr * expr

  (* variant for function calls. *)
  | Call of string * expr array

  (* variant for if/then/else. *)
  | If of expr * expr * expr

  (* variant for for/in. *)
  | For of string * expr * expr * expr option * expr

(* proto - This type represents the "prototype" for a function, which captures
 * its name, and its argument names (thus implicitly the number of arguments the
 * function takes). *)
type proto = Prototype of string * string array

(* func - This type represents a function definition itself. *)
type func = Function of proto * expr

```

parser.ml:

```

(=====
* Parser
*=====*)

(* binop_precedence - This holds the precedence for each binary operator that is
* defined *)
let binop_precedence: (char, int) Hashtbl.t = Hashtbl.create 10

(* precedence - Get the precedence of the pending binary operator token. *)
let precedence c = try Hashtbl.find binop_precedence c with Not_found -> -1

(* primary
* ::= identifier
* ::= numberexpr
* ::= parenexpr
* ::= ifexpr
* ::= forexpr *)
let rec parse_primary = parser
  (* numberexpr ::= number *)
  | [< 'Token.Number n >] -> Ast.Number n

  (* parenexpr ::= '(' expression ')' *)
  | [< 'Token.Kwd '('; e=parse_expr; 'Token.Kwd ')' ?? "expected ')'" >] -> e

  (* identifierexpr
  * ::= identifier
  * ::= identifier '(' argumentexpr ')' *)
  | [< 'Token.Ident id; stream >] ->
    let rec parse_args accumulator = parser
      | [< e=parse_expr; stream >] ->
        begin parser
          | [< 'Token.Kwd ','; e=parse_args (e :: accumulator) >] -> e
          | [< >] -> e :: accumulator
        end stream
      | [< >] -> accumulator
    in
    let rec parse_ident id = parser
      (* Call. *)
      | [< 'Token.Kwd '(';
        args=parse_args [];
        'Token.Kwd ')' ?? "expected ')'" >] ->
        Ast.Call (id, Array.of_list (List.rev args))

      (* Simple variable ref. *)
      | [< >] -> Ast.Variable id
    in
    parse_ident id stream

  (* ifexpr ::= 'if' expr 'then' expr 'else' expr *)
  | [< 'Token.If; c=parse_expr;
    'Token.Then ?? "expected 'then'"; t=parse_expr;
    'Token.Else ?? "expected 'else'"; e=parse_expr >] ->
    Ast.If (c, t, e)

  (* forexpr
  ::= 'for' identifier '=' expr ',' expr (',' expr)? 'in' expression *)
  | [< 'Token.For;
    'Token.Ident id ?? "expected identifier after for";

```

```

    'Token.Kwd '=' ?? "expected '=' after for";
    stream >] ->
begin parser
  | [<
    start=parse_expr;
    'Token.Kwd ',' ?? "expected ',' after for";
    end_=parse_expr;
    stream >] ->
    let step =
      begin parser
        | [< 'Token.Kwd ','; step=parse_expr >] -> Some step
        | [< >] -> None
      end stream
    in
    begin parser
      | [< 'Token.In; body=parse_expr >] ->
        Ast.For (id, start, end_, step, body)
      | [< >] ->
        raise (Stream.Error "expected 'in' after for")
    end stream
  | [< >] ->
    raise (Stream.Error "expected '=' after for")
end stream

| [< >] -> raise (Stream.Error "unknown token when expecting an expression.")

(* binoprhs
 * ::= ('+' primary)* *)
and parse_bin_rhs expr_prec lhs stream =
  match Stream.peek stream with
  (* If this is a binop, find its precedence. *)
  | Some (Token.Kwd c) when Hashtbl.mem binop_precedence c ->
    let token_prec = precedence c in
      (* If this is a binop that binds at least as tightly as the current binop,
       * consume it, otherwise we are done. *)
      if token_prec < expr_prec then lhs else begin
        (* Eat the binop. *)
        Stream.junk stream;

        (* Parse the primary expression after the binary operator. *)
        let rhs = parse_primary stream in

        (* Okay, we know this is a binop. *)
        let rhs =
          match Stream.peek stream with
          | Some (Token.Kwd c2) ->
            (* If BinOp binds less tightly with rhs than the operator after
             * rhs, let the pending operator take rhs as its lhs. *)
            let next_prec = precedence c2 in
              if token_prec < next_prec
              then parse_bin_rhs (token_prec + 1) rhs stream
              else rhs
          | _ -> rhs
        in

        (* Merge lhs/rhs. *)
        let lhs = Ast.Binary (c, lhs, rhs) in

```

```

    parse_bin_rhs expr_prec lhs stream
  end
  | _ -> lhs

(* expression
 * ::= primary binoprhs *)
and parse_expr = parser
  | [< lhs=parse_primary; stream >] -> parse_bin_rhs 0 lhs stream

(* prototype
 * ::= id '(' id* ')' *)
let parse_prototype =
  let rec parse_args accumulator = parser
    | [< 'Token.Ident' id; e=parse_args (id::accumulator) >] -> e
    | [< >] -> accumulator
  in

  parser
  | [< 'Token.Ident' id;
    'Token.Kwd' '(' ?? "expected '(' in prototype";
    args=parse_args [];
    'Token.Kwd ')' ?? "expected ')' in prototype" >] ->
    (* success. *)
    Ast.Prototype (id, Array.of_list (List.rev args))

  | [< >] ->
    raise (Stream.Error "expected function name in prototype")

(* definition ::= 'def' prototype expression *)
let parse_definition = parser
  | [< 'Token.Def'; p=parse_prototype; e=parse_expr >] ->
    Ast.Function (p, e)

(* toplevelexpr ::= expression *)
let parse_toplevel = parser
  | [< e=parse_expr >] ->
    (* Make an anonymous proto. *)
    Ast.Function (Ast.Prototype ("", []), e)

(* external ::= 'extern' prototype *)
let parse_extern = parser
  | [< 'Token.Extern'; e=parse_prototype >] -> e

```

codegen.ml:

```

(*=====*)
* Code Generation
(*=====*)

open Llvm

exception Error of string

let context = global_context ()
let the_module = create_module context "my cool jit"
let builder = builder context
let named_values:(string, llvalue) Hashtbl.t = Hashtbl.create 10
let double_type = double_type context

```

```

let rec codegen_expr = function
| Ast.Number n -> const_float double_type n
| Ast.Variable name ->
    (try Hashtbl.find named_values name with
     | Not_found -> raise (Error "unknown variable name"))
| Ast.Binary (op, lhs, rhs) ->
    let lhs_val = codegen_expr lhs in
    let rhs_val = codegen_expr rhs in
    begin
        match op with
        | '+' -> build_add lhs_val rhs_val "addtmp" builder
        | '-' -> build_sub lhs_val rhs_val "subtmp" builder
        | '*' -> build_mul lhs_val rhs_val "multmp" builder
        | '<' ->
            (* Convert bool 0/1 to double 0.0 or 1.0 *)
            let i = build_fcmp Fcmp.Ult lhs_val rhs_val "cmptmp" builder in
            build_uitofp i double_type "booltmp" builder
        | _ -> raise (Error "invalid binary operator")
    end
| Ast.Call (callee, args) ->
    (* Look up the name in the module table. *)
    let callee =
        match lookup_function callee the_module with
        | Some callee -> callee
        | None -> raise (Error "unknown function referenced")
    in
    let params = params callee in

    (* If argument mismatch error. *)
    if Array.length params == Array.length args then () else
        raise (Error "incorrect # arguments passed");
    let args = Array.map codegen_expr args in
    build_call callee args "calltmp" builder
| Ast.If (cond, then_, else_) ->
    let cond = codegen_expr cond in

    (* Convert condition to a bool by comparing equal to 0.0 *)
    let zero = const_float double_type 0.0 in
    let cond_val = build_fcmp Fcmp.One cond zero "ifcond" builder in

    (* Grab the first block so that we might later add the conditional branch
     * to it at the end of the function. *)
    let start_bb = insertion_block builder in
    let the_function = block_parent start_bb in

    let then_bb = append_block context "then" the_function in

    (* Emit 'then' value. *)
    position_at_end then_bb builder;
    let then_val = codegen_expr then_ in

    (* Codegen of 'then' can change the current block, update then_bb for the
     * phi. We create a new name because one is used for the phi node, and the
     * other is used for the conditional branch. *)
    let new_then_bb = insertion_block builder in

    (* Emit 'else' value. *)

```

```

let else_bb = append_block context "else" the_function in
position_at_end else_bb builder;
let else_val = codegen_expr else_ in

(* Codegen of 'else' can change the current block, update else_bb for the
 * phi. *)
let new_else_bb = insertion_block builder in

(* Emit merge block. *)
let merge_bb = append_block context "ifcont" the_function in
position_at_end merge_bb builder;
let incoming = [(then_val, new_then_bb); (else_val, new_else_bb)] in
let phi = build_phi incoming "iftmp" builder in

(* Return to the start block to add the conditional branch. *)
position_at_end start_bb builder;
ignore (build_cond_br cond_val then_bb else_bb builder);

(* Set a unconditional branch at the end of the 'then' block and the
 * 'else' block to the 'merge' block. *)
position_at_end new_then_bb builder; ignore (build_br merge_bb builder);
position_at_end new_else_bb builder; ignore (build_br merge_bb builder);

(* Finally, set the builder to the end of the merge block. *)
position_at_end merge_bb builder;

phi
| Ast.For (var_name, start, end_, step, body) ->
  (* Emit the start code first, without 'variable' in scope. *)
  let start_val = codegen_expr start in

  (* Make the new basic block for the loop header, inserting after current
   * block. *)
  let preheader_bb = insertion_block builder in
  let the_function = block_parent preheader_bb in
  let loop_bb = append_block context "loop" the_function in

  (* Insert an explicit fall through from the current block to the
   * loop_bb. *)
  ignore (build_br loop_bb builder);

  (* Start insertion in loop_bb. *)
  position_at_end loop_bb builder;

  (* Start the PHI node with an entry for start. *)
  let variable = build_phi [(start_val, preheader_bb)] var_name builder in

  (* Within the loop, the variable is defined equal to the PHI node. If it
   * shadows an existing variable, we have to restore it, so save it
   * now. *)
  let old_val =
    try Some (Hashtbl.find named_values var_name) with Not_found -> None
  in
  Hashtbl.add named_values var_name variable;

  (* Emit the body of the loop. This, like any other expr, can change the
   * current BB. Note that we ignore the value computed by the body, but
   * don't allow an error *)

```

```

ignore (codegen_expr body);

(* Emit the step value. *)
let step_val =
  match step with
  | Some step -> codegen_expr step
  (* If not specified, use 1.0. *)
  | None -> const_float double_type 1.0
in

let next_var = build_add variable step_val "nextvar" builder in

(* Compute the end condition. *)
let end_cond = codegen_expr end_ in

(* Convert condition to a bool by comparing equal to 0.0. *)
let zero = const_float double_type 0.0 in
let end_cond = build_fcmp Fcmp.One end_cond zero "loopcond" builder in

(* Create the "after loop" block and insert it. *)
let loop_end_bb = insertion_block builder in
let after_bb = append_block context "afterloop" the_function in

(* Insert the conditional branch into the end of loop_end_bb. *)
ignore (build_cond_br end_cond loop_bb after_bb builder);

(* Any new code will be inserted in after_bb. *)
position_at_end after_bb builder;

(* Add a new entry to the PHI node for the backedge. *)
add_incoming (next_var, loop_end_bb) variable;

(* Restore the unshadowed variable. *)
begin match old_val with
| Some old_val -> Hashtbl.add named_values var_name old_val
| None -> ()
end;

(* for expr always returns 0.0. *)
const_null double_type

let codegen_proto = function
| Ast.Prototype (name, args) ->
  (* Make the function type: double(double,double) etc. *)
  let doubles = Array.make (Array.length args) double_type in
  let ft = function_type double_type doubles in
  let f =
    match lookup_function name the_module with
    | None -> declare_function name ft the_module

  (* If 'f' conflicted, there was already something named 'name'. If it
  * has a body, don't allow redefinition or reextern. *)
  | Some f ->
    (* If 'f' already has a body, reject this. *)
    if block_begin f <> At_end f then
      raise (Error "redefinition of function");

    (* If 'f' took a different number of arguments, reject. *)

```



```

        if element_type (type_of f) <> ft then
            raise (Error "redefinition of function with different # args");
        f
    in

    (* Set names for all arguments. *)
    Array.iteri (fun i a ->
        let n = args.(i) in
        set_value_name n a;
        Hashtbl.add named_values n a;
    ) (params f);
    f

let codegen_func the_fpm = function
| Ast.Function (proto, body) ->
    Hashtbl.clear named_values;
    let the_function = codegen_proto proto in

    (* Create a new basic block to start insertion into. *)
    let bb = append_block context "entry" the_function in
    position_at_end bb builder;

    try
        let ret_val = codegen_expr body in

        (* Finish off the function. *)
        let _ = build_ret ret_val builder in

        (* Validate the generated code, checking for consistency. *)
        Llvm_analysis.assert_valid_function the_function;

        (* Optimize the function. *)
        let _ = PassManager.run_function the_function the_fpm in

        the_function
    with e ->
        delete_function the_function;
        raise e

```

oplevel.ml:

```

(*=====*)
* Top-Level parsing and JIT Driver
*=====*)

open Llvm
open Llvm_executionengine

(* top ::= definition | external | expression | ';' *)
let rec main_loop the_fpm the_execution_engine stream =
    match Stream.peek stream with
    | None -> ()

    (* ignore top-level semicolons. *)
    | Some (Token.Kwd ';') ->
        Stream.junk stream;
        main_loop the_fpm the_execution_engine stream

```

```

| Some token ->
  begin
    try match token with
    | Token.Def ->
      let e = Parser.parse_definition stream in
      print_endline "parsed a function definition.";
      dump_value (Codegen.codegen_func the_fpm e);
    | Token.Extern ->
      let e = Parser.parse_extern stream in
      print_endline "parsed an extern.";
      dump_value (Codegen.codegen_proto e);
    | _ ->
      (* Evaluate a top-level expression into an anonymous function. *)
      let e = Parser.parse_toplevel stream in
      print_endline "parsed a top-level expr";
      let the_function = Codegen.codegen_func the_fpm e in
      dump_value the_function;

      (* JIT the function, returning a function pointer. *)
      let result = ExecutionEngine.run_function the_function [||]
        the_execution_engine in

      print_string "Evaluated to ";
      print_float (GenericValue.as_float Codegen.double_type result);
      print_newline ();
      with Stream.Error s | Codegen.Error s ->
        (* Skip token for error recovery. *)
        Stream.junk stream;
        print_endline s;
    end;
    print_string "ready> "; flush stdout;
    main_loop the_fpm the_execution_engine stream

```

toy.ml:

```

(*=====
 * Main driver code.
 *=====*)

open Llvm
open Llvm_executionengine
open Llvm_target
open Llvm_scalar_opts

let main () =
  ignore (initialize_native_target ());

  (* Install standard binary operators.
   * 1 is the lowest precedence. *)
  Hashtbl.add Parser.binop_precedence '<' 10;
  Hashtbl.add Parser.binop_precedence '+' 20;
  Hashtbl.add Parser.binop_precedence '-' 20;
  Hashtbl.add Parser.binop_precedence '*' 40;      (* highest. *)

  (* Prime the first token. *)
  print_string "ready> "; flush stdout;
  let stream = Lexer.lex (Stream.of_channel stdin) in

```

```

(* Create the JIT. *)
let the_execution_engine = ExecutionEngine.create Codegen.the_module in
let the_fpm = PassManager.create_function Codegen.the_module in

(* Set up the optimizer pipeline. Start with registering info about how the
 * target lays out data structures. *)
DataLayout.add (ExecutionEngine.target_data the_execution_engine) the_fpm;

(* Do simple "peephole" optimizations and bit-twiddling optzn. *)
add_instruction_combination the_fpm;

(* reassociate expressions. *)
add_reassociation the_fpm;

(* Eliminate Common SubExpressions. *)
add_gvn the_fpm;

(* Simplify the control flow graph (deleting unreachable blocks, etc). *)
add_cfg_simplification the_fpm;

ignore (PassManager.initialize the_fpm);

(* Run the main "interpreter loop" now. *)
Toplevel.main_loop the_fpm the_execution_engine stream;

(* Print out all the generated code. *)
dump_module Codegen.the_module
;;

main ()

```

bindings.c

```

#include <stdio.h>

/* putchar - putchar that takes a double and returns 0. */
extern double putchar(double X) {
    putchar((char)X);
    return 0;
}

```

Next: Extending the language: user-defined operators

Kaleidoscope: Extending the Language: User-defined Operators

- *Chapter 6 Introduction*
- *User-defined Operators: the Idea*
- *User-defined Binary Operators*
- *User-defined Unary Operators*
- *Kicking the Tires*
- *Full Code Listing*

Chapter 6 Introduction

Welcome to Chapter 6 of the “[Implementing a language with LLVM](#)” tutorial. At this point in our tutorial, we now have a fully functional language that is fairly minimal, but also useful. There is still one big problem with it, however. Our language doesn’t have many useful operators (like division, logical negation, or even any comparisons besides less-than).

This chapter of the tutorial takes a wild digression into adding user-defined operators to the simple and beautiful Kaleidoscope language. This digression now gives us a simple and ugly language in some ways, but also a powerful one at the same time. One of the great things about creating your own language is that you get to decide what is good or bad. In this tutorial we’ll assume that it is okay to use this as a way to show some interesting parsing techniques.

At the end of this tutorial, we’ll run through an example Kaleidoscope application that *renders the Mandelbrot set*. This gives an example of what you can build with Kaleidoscope and its feature set.

User-defined Operators: the Idea

The “operator overloading” that we will add to Kaleidoscope is more general than languages like C++. In C++, you are only allowed to redefine existing operators: you can’t programmatically change the grammar, introduce new operators, change precedence levels, etc. In this chapter, we will add this capability to Kaleidoscope, which will let the user round out the set of operators that are supported.

The point of going into user-defined operators in a tutorial like this is to show the power and flexibility of using a hand-written parser. Thus far, the parser we have been implementing uses recursive descent for most parts of the grammar and operator precedence parsing for the expressions. See [Chapter 2](#) for details. Without using operator precedence parsing, it would be very difficult to allow the programmer to introduce new operators into the grammar: the grammar is dynamically extensible as the JIT runs.

The two specific features we’ll add are programmable unary operators (right now, Kaleidoscope has no unary operators at all) as well as binary operators. An example of this is:

```
# Logical unary not.
def unary! (v)
  if v then
    0
  else
    1;

# Define > with the same precedence as <.
def binary> 10 (LHS RHS)
  RHS < LHS;

# Binary "logical or", (note that it does not "short circuit")
def binary| 5 (LHS RHS)
  if LHS then
    1
  else if RHS then
    1
  else
    0;

# Define = with slightly lower precedence than relationals.
def binary= 9 (LHS RHS)
  !(LHS < RHS | LHS > RHS);
```

Many languages aspire to being able to implement their standard runtime library in the language itself. In Kaleidoscope, we can implement significant parts of the language in the library!

We will break down implementation of these features into two parts: implementing support for user-defined binary operators and adding unary operators.

User-defined Binary Operators

Adding support for user-defined binary operators is pretty simple with our current framework. We'll first add support for the unary/binary keywords:

```
type token =
  ...
  (* operators *)
  | Binary | Unary
  ...

and lex_ident buffer = parser
  ...
  | "for" -> [< 'Token.For; stream >]
  | "in" -> [< 'Token.In; stream >]
  | "binary" -> [< 'Token.Binary; stream >]
  | "unary" -> [< 'Token.Unary; stream >]
```

This just adds lexer support for the unary and binary keywords, like we did in [previous chapters](#). One nice thing about our current AST, is that we represent binary operators with full generalisation by using their ASCII code as the opcode. For our extended operators, we'll use this same representation, so we don't need any new AST or parser support.

On the other hand, we have to be able to represent the definitions of these new operators, in the “def binary! 5” part of the function definition. In our grammar so far, the “name” for the function definition is parsed as the “prototype” production and into the `Ast.Prototype` AST node. To represent our new user-defined operators as prototypes, we have to extend the `Ast.Prototype` AST node like this:

```
(* proto - This type represents the "prototype" for a function, which captures
 * its name, and its argument names (thus implicitly the number of arguments the
 * function takes). *)
type proto =
  | Prototype of string * string array
  | BinOpPrototype of string * string array * int
```

Basically, in addition to knowing a name for the prototype, we now keep track of whether it was an operator, and if it was, what precedence level the operator is at. The precedence is only used for binary operators (as you'll see below, it just doesn't apply for unary operators). Now that we have a way to represent the prototype for a user-defined operator, we need to parse it:

```
(* prototype
 * ::= id '(' id* ')'
 * ::= binary LETTER number? (id, id)
 * ::= unary LETTER number? (id) *)
let parse_prototype =
  let rec parse_args accumulator = parser
    | [< 'Token.Ident id; e=parse_args (id::accumulator) >] -> e
    | [< >] -> accumulator
  in
  let parse_operator = parser
    | [< 'Token.Unary >] -> "unary", 1
    | [< 'Token.Binary >] -> "binary", 2
  in
  let parse_binary_precedence = parser
```

```

| [< 'Token.Number n >] -> int_of_float n
| [< >] -> 30
in
parser
| [< 'Token.Ident id;
  'Token.Kwd '(' ?? "expected '(' in prototype";
  args=parse_args [];
  'Token.Kwd ')' ?? "expected ')' in prototype" >] ->
  (* success. *)
  Ast.Prototype (id, Array.of_list (List.rev args))
| [< (prefix, kind)=parse_operator;
  'Token.Kwd op ?? "expected an operator";
  (* Read the precedence if present. *)
  binary_precedence=parse_binary_precedence;
  'Token.Kwd '(' ?? "expected '(' in prototype";
  args=parse_args [];
  'Token.Kwd ')' ?? "expected ')' in prototype" >] ->
let name = prefix ^ (String.make 1 op) in
let args = Array.of_list (List.rev args) in

(* Verify right number of arguments for operator. *)
if Array.length args != kind
then raise (Stream.Error "invalid number of operands for operator")
else
  if kind == 1 then
    Ast.Prototype (name, args)
  else
    Ast.BinOpPrototype (name, args, binary_precedence)
| [< >] ->
  raise (Stream.Error "expected function name in prototype")

```

This is all fairly straightforward parsing code, and we have already seen a lot of similar code in the past. One interesting part about the code above is the couple lines that set up name for binary operators. This builds names like “binary@” for a newly defined “@” operator. This then takes advantage of the fact that symbol names in the LLVM symbol table are allowed to have any character in them, including embedded nul characters.

The next interesting thing to add, is codegen support for these binary operators. Given our current structure, this is a simple addition of a default case for our existing binary operator node:

```

let codegen_expr = function
...
| Ast.Binary (op, lhs, rhs) ->
  let lhs_val = codegen_expr lhs in
  let rhs_val = codegen_expr rhs in
  begin
    match op with
    | '+' -> build_add lhs_val rhs_val "addtmp" builder
    | '-' -> build_sub lhs_val rhs_val "subtmp" builder
    | '*' -> build_mul lhs_val rhs_val "multmp" builder
    | '<' ->
      (* Convert bool 0/1 to double 0.0 or 1.0 *)
      let i = build_fcmp Fcmp.Ult lhs_val rhs_val "cmptmp" builder in
      build_uitofp i double_type "booltmp" builder
    | _ ->
      (* If it wasn't a builtin binary operator, it must be a user defined
       * one. Emit a call to it. *)
      let callee = "binary" ^ (String.make 1 op) in
      let callee =

```

```

    match lookup_function callee the_module with
    | Some callee -> callee
    | None -> raise (Error "binary operator not found!")
  in
    build_call callee [|lhs_val; rhs_val|] "binop" builder
end

```

As you can see above, the new code is actually really simple. It just does a lookup for the appropriate operator in the symbol table and generates a function call to it. Since user-defined operators are just built as normal functions (because the “prototype” boils down to a function with the right name) everything falls into place.

The final piece of code we are missing, is a bit of top level magic:

```

let codegen_func the_fpm = function
| Ast.Function (proto, body) ->
  Hashtbl.clear named_values;
  let the_function = codegen_proto proto in

  (* If this is an operator, install it. *)
  begin match proto with
  | Ast.BinOpPrototype (name, args, prec) ->
    let op = name.[String.length name - 1] in
    Hashtbl.add Parser.binop_precedence op prec;
  | _ -> ()
  end;

  (* Create a new basic block to start insertion into. *)
  let bb = append_block context "entry" the_function in
  position_at_end bb builder;
  ...

```

Basically, before codegening a function, if it is a user-defined operator, we register it in the precedence table. This allows the binary operator parsing logic we already have in place to handle it. Since we are working on a fully-general operator precedence parser, this is all we need to do to “extend the grammar”.

Now we have useful user-defined binary operators. This builds a lot on the previous framework we built for other operators. Adding unary operators is a bit more challenging, because we don’t have any framework for it yet - lets see what it takes.

User-defined Unary Operators

Since we don’t currently support unary operators in the Kaleidoscope language, we’ll need to add everything to support them. Above, we added simple support for the ‘unary’ keyword to the lexer. In addition to that, we need an AST node:

```

type expr =
...
(* variant for a unary operator. *)
| Unary of char * expr
...

```

This AST node is very simple and obvious by now. It directly mirrors the binary operator AST node, except that it only has one child. With this, we need to add the parsing logic. Parsing a unary operator is pretty simple: we’ll add a new function to do it:

```

(* unary
 * ::= primary

```

```

*   ::= '!' unary *)
and parse_unary = parser
  (* If this is a unary operator, read it. *)
  | [< 'Token.Kwd op when op != '(' && op != ')'; operand=parse_expr >] ->
    Ast.Unary (op, operand)

  (* If the current token is not an operator, it must be a primary expr. *)
  | [< stream >] -> parse_primary stream

```

The grammar we add is pretty straightforward here. If we see a unary operator when parsing a primary operator, we eat the operator as a prefix and parse the remaining piece as another unary operator. This allows us to handle multiple unary operators (e.g. “!!x”). Note that unary operators can’t have ambiguous parses like binary operators can, so there is no need for precedence information.

The problem with this function, is that we need to call ParseUnary from somewhere. To do this, we change previous callers of ParsePrimary to call parse_unary instead:

```

(* binoprhs
*   ::= ('+' primary)* *)
and parse_bin_rhs expr_prec lhs stream =
  ...
  (* Parse the unary expression after the binary operator. *)
  let rhs = parse_unary stream in
  ...

...

(* expression
*   ::= primary binoprhs *)
and parse_expr = parser
  | [< lhs=parse_unary; stream >] -> parse_bin_rhs 0 lhs stream

```

With these two simple changes, we are now able to parse unary operators and build the AST for them. Next up, we need to add parser support for prototypes, to parse the unary operator prototype. We extend the binary operator code above with:

```

(* prototype
*   ::= id '(' id* ')'
*   ::= binary LETTER number? (id, id)
*   ::= unary LETTER number? (id) *)
let parse_prototype =
  let rec parse_args accumulator = parser
    | [< 'Token.Ident id; e=parse_args (id::accumulator) >] -> e
    | [< >] -> accumulator
  in
  let parse_operator = parser
    | [< 'Token.Unary >] -> "unary", 1
    | [< 'Token.Binary >] -> "binary", 2
  in
  let parse_binary_precedence = parser
    | [< 'Token.Number n >] -> int_of_float n
    | [< >] -> 30
  in
  parser
  | [< 'Token.Ident id;
    'Token.Kwd '(' ?? "expected '(' in prototype";
    args=parse_args [];
    'Token.Kwd ')' ?? "expected ')' in prototype" >] ->

```



```

(* success. *)
Ast.Prototype (id, Array.of_list (List.rev args))
| [< (prefix, kind)=parse_operator;
  'Token.Kwd op ?? "expected an operator";
  (* Read the precedence if present. *)
  binary_precedence=parse_binary_precedence;
  'Token.Kwd '(' ?? "expected '(' in prototype";
  args=parse_args [];
  'Token.Kwd ')' ?? "expected ')' in prototype" >] ->
let name = prefix ^ (String.make 1 op) in
let args = Array.of_list (List.rev args) in

(* Verify right number of arguments for operator. *)
if Array.length args != kind
then raise (Stream.Error "invalid number of operands for operator")
else
  if kind == 1 then
    Ast.Prototype (name, args)
  else
    Ast.BinOpPrototype (name, args, binary_precedence)
| [>] ->
  raise (Stream.Error "expected function name in prototype")

```

As with binary operators, we name unary operators with a name that includes the operator character. This assists us at code generation time. Speaking of, the final piece we need to add is codegen support for unary operators. It looks like this:

```

let rec codegen_expr = function
...
| Ast.Unary (op, operand) ->
  let operand = codegen_expr operand in
  let callee = "unary" ^ (String.make 1 op) in
  let callee =
    match lookup_function callee the_module with
    | Some callee -> callee
    | None -> raise (Error "unknown unary operator")
  in
  build_call callee [|operand|] "unop" builder

```

This code is similar to, but simpler than, the code for binary operators. It is simpler primarily because it doesn't need to handle any predefined operators.

Kicking the Tires

It is somewhat hard to believe, but with a few simple extensions we've covered in the last chapters, we have grown a real-ish language. With this, we can do a lot of interesting things, including I/O, math, and a bunch of other things. For example, we can now add a nice sequencing operator (printd is defined to print out the specified value and a newline):

```

ready> extern printd(x);
Read extern: declare double @printd(double)
ready> def binary : 1 (x y) 0; # Low-precedence operator that ignores operands.
..
ready> printd(123) : printd(456) : printd(789);
123.000000
456.000000

```

```
789.000000
Evaluated to 0.000000
```

We can also define a bunch of other “primitive” operations, such as:

```
# Logical unary not.
def unary!(v)
  if v then
    0
  else
    1;

# Unary negate.
def unary-(v)
  0-v;

# Define > with the same precedence as <.
def binary> 10 (LHS RHS)
  RHS < LHS;

# Binary logical or, which does not short circuit.
def binary| 5 (LHS RHS)
  if LHS then
    1
  else if RHS then
    1
  else
    0;

# Binary logical and, which does not short circuit.
def binary& 6 (LHS RHS)
  if !LHS then
    0
  else
    !!RHS;

# Define = with slightly lower precedence than relationals.
def binary = 9 (LHS RHS)
  !(LHS < RHS | LHS > RHS);
```

Given the previous if/then/else support, we can also define interesting functions for I/O. For example, the following prints out a character whose “density” reflects the value passed in: the lower the value, the denser the character:

```
ready>

extern putchar(char)
def printdensity(d)
  if d > 8 then
    putchar(32) # ' '
  else if d > 4 then
    putchar(46) # '.'
  else if d > 2 then
    putchar(43) # '+'
  else
    putchar(42); # '*'
...
ready> printdensity(1): printdensity(2): printdensity(3) :
      printdensity(4): printdensity(5): printdensity(9): putchar(10);
```

```
# determine whether the specific location diverges.
# Solve for  $z = z^2 + c$  in the complex plane.
def mandelconverger(real imag iters creal cimag)
    if iters > 255 | (real*real + imag*imag > 4) then
        iters
    else
        mandelconverger(real*real - imag*imag + creal,
                        2*real*imag + cimag,
                        iters+1, creal, cimag);

# return the number of iterations required for the iteration to escape
def mandelconverge(real imag)
    mandelconverger(real, imag, 0, real, imag);
```

```
# compute and plot the mandelbrot set with the specified 2 dimensional range
# info.
def mandelhelp(xmin xmax xstep   ymin ymax ystep)
    for y = ymin, y < ymax, ystep in (
        (for x = xmin, x < xmax, xstep in
            printdensity(mandelconverge(x,y))
        : putchard(10)
    )

# mandel - This is a convenient helper function for plotting the mandelbrot set
# from the specified position with the specified Magnification.
def mandel(realstart imagstart realmag imagmag)
    mandelhelp(realstart, realstart+realmag*78, realmag,
                imagstart, imagstart+imagmag*40, imagmag);
```

[illegible]

[illegible]

```
Evaluated to 0.000000
```

```
ready> mandel(-2, -1, 0.02, 0.04);
```

[illegible]

2.17. LLVM Tutorial: Table of Contents 671

At this point, you may be starting to realize that Kaleidoscope is a real and powerful language. It may not be self-similar :), but it can be used to plot things that are!

With this, we conclude the “adding user-defined operators” chapter of the tutorial. We have successfully augmented our language, adding the ability to extend the language in the library, and we have shown how this can be used to build a simple but interesting end-user application in Kaleidoscope. At this point, Kaleidoscope can build a variety of applications that are functional and can call functions with side-effects, but it can’t actually define and mutate a variable itself.

Strikingly, variable mutation is an important feature of some languages, and it is not at all obvious how to [add support for mutable variables](#) without having to add an “SSA construction” phase to your front-end. In the next chapter, we will describe how you can add variable mutation without building SSA in your front-end.

Full Code Listing

Here is the complete code listing for our running example, enhanced with the if/then/else and for expressions.. To build this example, use:

```
# Compile
ocamlbuild toy.byte
# Run
./toy.byte
```

Here is the code:

_tags:

```
<{lexer,parser}.ml>: use_camlp4, pp(camlp4of)
<*. {byte,native}>: g++, use_llvm, use_llvm_analysis
<*. {byte,native}>: use_llvm_executionengine, use_llvm_target
<*. {byte,native}>: use_llvm_scalar_opts, use_bindings
```

myocamlbuild.ml:

```
open Ocamlbuild_plugin;;

ocaml_lib ~extern:true "llvm";;
ocaml_lib ~extern:true "llvm_analysis";;
ocaml_lib ~extern:true "llvm_executionengine";;
ocaml_lib ~extern:true "llvm_target";;
ocaml_lib ~extern:true "llvm_scalar_opts";;

flag ["link"; "ocaml"; "g++"] (S[A"-cc"; A"g++"; A"-cclib"; A"-rdynamic"]);;
dep ["link"; "ocaml"; "use_bindings"] ["bindings.o"];;
```

token.ml:

```
(=====
 * Lexer Tokens
 *=====)

(* The lexer returns these 'Kwd' if it is an unknown character, otherwise one of
 * these others for known things. *)
type token =
  (* commands *)
  | Def | Extern
```

```

(* primary *)
| Ident of string | Number of float

(* unknown *)
| Kwd of char

(* control *)
| If | Then | Else
| For | In

(* operators *)
| Binary | Unary

```

lexer.ml:

```

(=====)
* Lexer
(=====)

let rec lex = parser
  (* Skip any whitespace. *)
  | [< ' (' | '\n' | '\r' | '\t'); stream >] -> lex stream

  (* identifier: [a-zA-Z][a-zA-Z0-9] *)
  | [< ' ('A' .. 'Z' | 'a' .. 'z' as c); stream >] ->
    let buffer = Buffer.create 1 in
    Buffer.add_char buffer c;
    lex_ident buffer stream

  (* number: [0-9.]+ *)
  | [< ' ('0' .. '9' as c); stream >] ->
    let buffer = Buffer.create 1 in
    Buffer.add_char buffer c;
    lex_number buffer stream

  (* Comment until end of line. *)
  | [< ' ('#'); stream >] ->
    lex_comment stream

  (* Otherwise, just return the character as its ascii value. *)
  | [< 'c; stream >] ->
    [< 'Token.Kwd c; lex stream >]

  (* end of stream. *)
  | [< >] -> [< >]

and lex_number buffer = parser
  | [< ' ('0' .. '9' | '.' as c); stream >] ->
    Buffer.add_char buffer c;
    lex_number buffer stream
  | [< stream=lex >] ->
    [< 'Token.Number (float_of_string (Buffer.contents buffer)); stream >]

and lex_ident buffer = parser
  | [< ' ('A' .. 'Z' | 'a' .. 'z' | '0' .. '9' as c); stream >] ->
    Buffer.add_char buffer c;
    lex_ident buffer stream
  | [< stream=lex >] ->

```

```
match Buffer.contents buffer with
| "def" -> [< 'Token.Def; stream >]
| "extern" -> [< 'Token.Extern; stream >]
| "if" -> [< 'Token.If; stream >]
| "then" -> [< 'Token.Then; stream >]
| "else" -> [< 'Token.Else; stream >]
| "for" -> [< 'Token.For; stream >]
| "in" -> [< 'Token.In; stream >]
| "binary" -> [< 'Token.Binary; stream >]
| "unary" -> [< 'Token.Unary; stream >]
| id -> [< 'Token.Ident id; stream >]

and lex_comment = parser
| [< ' ('\n'); stream=lex >] -> stream
| [< 'c; e=lex_comment >] -> e
| [< >] -> [< >]
```

ast.ml:

```
(=====)
* Abstract Syntax Tree (aka Parse Tree)
=====*)

(* expr - Base type for all expression nodes. *)
type expr =
  (* variant for numeric literals like "1.0". *)
  | Number of float

  (* variant for referencing a variable, like "a". *)
  | Variable of string

  (* variant for a unary operator. *)
  | Unary of char * expr

  (* variant for a binary operator. *)
  | Binary of char * expr * expr

  (* variant for function calls. *)
  | Call of string * expr array

  (* variant for if/then/else. *)
  | If of expr * expr * expr

  (* variant for for/in. *)
  | For of string * expr * expr * expr option * expr

(* proto - This type represents the "prototype" for a function, which captures
 * its name, and its argument names (thus implicitly the number of arguments the
 * function takes). *)
type proto =
  | Prototype of string * string array
  | BinOpPrototype of string * string array * int

(* func - This type represents a function definition itself. *)
type func = Function of proto * expr
```

parser.ml:


```

(=====
* Parser
*=====*)

(* binop_precedence - This holds the precedence for each binary operator that is
* defined *)
let binop_precedence: (char, int) Hashtbl.t = Hashtbl.create 10

(* precedence - Get the precedence of the pending binary operator token. *)
let precedence c = try Hashtbl.find binop_precedence c with Not_found -> -1

(* primary
* ::= identifier
* ::= numberexpr
* ::= parenexpr
* ::= ifexpr
* ::= forexpr *)
let rec parse_primary = parser
  (* numberexpr ::= number *)
  | [< 'Token.Number n >] -> Ast.Number n

  (* parenexpr ::= '(' expression ')' *)
  | [< 'Token.Kwd '('; e=parse_expr; 'Token.Kwd ')' ?? "expected ')" ">] -> e

  (* identifierexpr
  * ::= identifier
  * ::= identifier '(' argumentexpr ')' *)
  | [< 'Token.Ident id; stream >] ->
    let rec parse_args accumulator = parser
      | [< e=parse_expr; stream >] ->
        begin parser
          | [< 'Token.Kwd ','; e=parse_args (e :: accumulator) >] -> e
          | [< >] -> e :: accumulator
        end stream
      | [< >] -> accumulator
    in
    let rec parse_ident id = parser
      (* Call. *)
      | [< 'Token.Kwd '(';
        args=parse_args [];
        'Token.Kwd ')' ?? "expected ')" ">] ->
        Ast.Call (id, Array.of_list (List.rev args))

      (* Simple variable ref. *)
      | [< >] -> Ast.Variable id
    in
    parse_ident id stream

  (* ifexpr ::= 'if' expr 'then' expr 'else' expr *)
  | [< 'Token.If; c=parse_expr;
    'Token.Then ?? "expected 'then'"; t=parse_expr;
    'Token.Else ?? "expected 'else'"; e=parse_expr >] ->
    Ast.If (c, t, e)

  (* forexpr
  ::= 'for' identifier '=' expr ',' expr (',' expr)? 'in' expression *)
  | [< 'Token.For;
    'Token.Ident id ?? "expected identifier after for";

```

```

    'Token.Kwd '=' ?? "expected '=' after for";
    stream >] ->
begin parser
  | [<
    start=parse_expr;
    'Token.Kwd ',' ?? "expected ',' after for";
    end_=parse_expr;
    stream >] ->
    let step =
      begin parser
        | [< 'Token.Kwd ','; step=parse_expr >] -> Some step
        | [< >] -> None
      end stream
    in
    begin parser
      | [< 'Token.In; body=parse_expr >] ->
        Ast.For (id, start, end_, step, body)
      | [< >] ->
        raise (Stream.Error "expected 'in' after for")
      end stream
    | [< >] ->
      raise (Stream.Error "expected '=' after for")
    end stream

  | [< >] -> raise (Stream.Error "unknown token when expecting an expression.")

(* unary
 * ::= primary
 * ::= '!' unary *)
and parse_unary = parser
  (* If this is a unary operator, read it. *)
  | [< 'Token.Kwd op when op != '(' && op != ')'; operand=parse_expr >] ->
    Ast.Unary (op, operand)

  (* If the current token is not an operator, it must be a primary expr. *)
  | [< stream >] -> parse_primary stream

(* binoprhs
 * ::= ('+' primary)* *)
and parse_bin_rhs expr_prec lhs stream =
  match Stream.peek stream with
  (* If this is a binop, find its precedence. *)
  | Some (Token.Kwd c) when Hashtbl.mem binop_precedence c ->
    let token_prec = precedence c in
    (* If this is a binop that binds at least as tightly as the current binop,
     * consume it, otherwise we are done. *)
    if token_prec < expr_prec then lhs else begin
      (* Eat the binop. *)
      Stream.junk stream;

      (* Parse the unary expression after the binary operator. *)
      let rhs = parse_unary stream in

      (* Okay, we know this is a binop. *)
      let rhs =
        match Stream.peek stream with
        | Some (Token.Kwd c2) ->

```

```

        (* If BinOp binds less tightly with rhs than the operator after
        * rhs, let the pending operator take rhs as its lhs. *)
        let next_prec = precedence c2 in
        if token_prec < next_prec
        then parse_bin_rhs (token_prec + 1) rhs stream
        else rhs
    | _ -> rhs
in

    (* Merge lhs/rhs. *)
    let lhs = Ast.Binary (c, lhs, rhs) in
    parse_bin_rhs expr_prec lhs stream
end
| _ -> lhs

(* expression
 * ::= primary binoprhs *)
and parse_expr = parser
| [< lhs=parse_unary; stream >] -> parse_bin_rhs 0 lhs stream

(* prototype
 * ::= id '(' id* ')'
 * ::= binary LETTER number? (id, id)
 * ::= unary LETTER number? (id) *)
let parse_prototype =
    let rec parse_args accumulator = parser
        | [< 'Token.Ident' id; e=parse_args (id::accumulator) >] -> e
        | [< >] -> accumulator
    in
    let parse_operator = parser
        | [< 'Token.Unary' >] -> "unary", 1
        | [< 'Token.Binary' >] -> "binary", 2
    in
    let parse_binary_precedence = parser
        | [< 'Token.Number' n >] -> int_of_float n
        | [< >] -> 30
    in
    parser
    | [< 'Token.Ident' id;
        'Token.Kwd' '(' ?? "expected '(' in prototype";
        args=parse_args [];
        'Token.Kwd' ')' ?? "expected ')' in prototype" >] ->
        (* success. *)
        Ast.Prototype (id, Array.of_list (List.rev args))
    | [< (prefix, kind)=parse_operator;
        'Token.Kwd' op ?? "expected an operator";
        (* Read the precedence if present. *)
        binary_precedence=parse_binary_precedence;
        'Token.Kwd' '(' ?? "expected '(' in prototype";
        args=parse_args [];
        'Token.Kwd' ')' ?? "expected ')' in prototype" >] ->
        let name = prefix ^ (String.make 1 op) in
        let args = Array.of_list (List.rev args) in

        (* Verify right number of arguments for operator. *)
        if Array.length args != kind
        then raise (Stream.Error "invalid number of operands for operator")
        else

```

```

    if kind == 1 then
        Ast.Prototype (name, args)
    else
        Ast.BinOpPrototype (name, args, binary_precedence)
| [< >] ->
    raise (Stream.Error "expected function name in prototype")

(* definition ::= 'def' prototype expression *)
let parse_definition = parser
| [< 'Token.Def; p=parse_prototype; e=parse_expr >] ->
    Ast.Function (p, e)

(* toplevelexpr ::= expression *)
let parse_toplevel = parser
| [< e=parse_expr >] ->
    (* Make an anonymous proto. *)
    Ast.Function (Ast.Prototype ("", [[]]), e)

(* external ::= 'extern' prototype *)
let parse_extern = parser
| [< 'Token.Extern; e=parse_prototype >] -> e

```

codegen.ml:

```

(=====
 * Code Generation
 *=====)

open Llvml

exception Error of string

let context = global_context ()
let the_module = create_module context "my cool jit"
let builder = builder context
let named_values:(string, llvalue) Hashtbl.t = Hashtbl.create 10
let double_type = double_type context

let rec codegen_expr = function
| Ast.Number n -> const_float double_type n
| Ast.Variable name ->
    (try Hashtbl.find named_values name with
    | Not_found -> raise (Error "unknown variable name"))
| Ast.Unary (op, operand) ->
    let operand = codegen_expr operand in
    let callee = "unary" ^ (String.make 1 op) in
    let callee =
        match lookup_function callee the_module with
        | Some callee -> callee
        | None -> raise (Error "unknown unary operator")
    in
    build_call callee [|operand|] "unop" builder
| Ast.Binary (op, lhs, rhs) ->
    let lhs_val = codegen_expr lhs in
    let rhs_val = codegen_expr rhs in
    begin
        match op with
        | '+' -> build_add lhs_val rhs_val "addtmp" builder

```

```

| '-' -> build_sub lhs_val rhs_val "subtmp" builder
| '*' -> build_mul lhs_val rhs_val "multmp" builder
| '<' ->
    (* Convert bool 0/1 to double 0.0 or 1.0 *)
    let i = build_fcmp Fcmp.Ult lhs_val rhs_val "cmptmp" builder in
    build_uitofp i double_type "booltmp" builder
| _ ->
    (* If it wasn't a builtin binary operator, it must be a user defined
    * one. Emit a call to it. *)
    let callee = "binary" ^ (String.make 1 op) in
    let callee =
        match lookup_function callee the_module with
        | Some callee -> callee
        | None -> raise (Error "binary operator not found!")
    in
    build_call callee [|lhs_val; rhs_val|] "binop" builder
end
| Ast.Call (callee, args) ->
    (* Look up the name in the module table. *)
    let callee =
        match lookup_function callee the_module with
        | Some callee -> callee
        | None -> raise (Error "unknown function referenced")
    in
    let params = params callee in

    (* If argument mismatch error. *)
    if Array.length params == Array.length args then () else
        raise (Error "incorrect # arguments passed");
    let args = Array.map codegen_expr args in
    build_call callee args "calltmp" builder
| Ast.If (cond, then_, else_) ->
    let cond = codegen_expr cond in

    (* Convert condition to a bool by comparing equal to 0.0 *)
    let zero = const_float double_type 0.0 in
    let cond_val = build_fcmp Fcmp.One cond zero "ifcond" builder in

    (* Grab the first block so that we might later add the conditional branch
    * to it at the end of the function. *)
    let start_bb = insertion_block builder in
    let the_function = block_parent start_bb in

    let then_bb = append_block context "then" the_function in

    (* Emit 'then' value. *)
    position_at_end then_bb builder;
    let then_val = codegen_expr then_ in

    (* Codegen of 'then' can change the current block, update then_bb for the
    * phi. We create a new name because one is used for the phi node, and the
    * other is used for the conditional branch. *)
    let new_then_bb = insertion_block builder in

    (* Emit 'else' value. *)
    let else_bb = append_block context "else" the_function in
    position_at_end else_bb builder;
    let else_val = codegen_expr else_ in

```

```

(* Codegen of 'else' can change the current block, update else_bb for the
 * phi. *)
let new_else_bb = insertion_block builder in

(* Emit merge block. *)
let merge_bb = append_block context "ifcont" the_function in
position_at_end merge_bb builder;
let incoming = [(then_val, new_then_bb); (else_val, new_else_bb)] in
let phi = build_phi incoming "iftmp" builder in

(* Return to the start block to add the conditional branch. *)
position_at_end start_bb builder;
ignore (build_cond_br cond_val then_bb else_bb builder);

(* Set a unconditional branch at the end of the 'then' block and the
 * 'else' block to the 'merge' block. *)
position_at_end new_then_bb builder; ignore (build_br merge_bb builder);
position_at_end new_else_bb builder; ignore (build_br merge_bb builder);

(* Finally, set the builder to the end of the merge block. *)
position_at_end merge_bb builder;

phi
| Ast.For (var_name, start, end_, step, body) ->
  (* Emit the start code first, without 'variable' in scope. *)
  let start_val = codegen_expr start in

  (* Make the new basic block for the loop header, inserting after current
   * block. *)
  let preheader_bb = insertion_block builder in
  let the_function = block_parent preheader_bb in
  let loop_bb = append_block context "loop" the_function in

  (* Insert an explicit fall through from the current block to the
   * loop_bb. *)
  ignore (build_br loop_bb builder);

  (* Start insertion in loop_bb. *)
  position_at_end loop_bb builder;

  (* Start the PHI node with an entry for start. *)
  let variable = build_phi [(start_val, preheader_bb)] var_name builder in

  (* Within the loop, the variable is defined equal to the PHI node. If it
   * shadows an existing variable, we have to restore it, so save it
   * now. *)
  let old_val =
    try Some (Hashtbl.find named_values var_name) with Not_found -> None
  in
  Hashtbl.add named_values var_name variable;

  (* Emit the body of the loop. This, like any other expr, can change the
   * current BB. Note that we ignore the value computed by the body, but
   * don't allow an error *)
  ignore (codegen_expr body);

  (* Emit the step value. *)

```

```

let step_val =
  match step with
  | Some step -> codegen_expr step
  (* If not specified, use 1.0. *)
  | None -> const_float double_type 1.0
in

let next_var = build_add variable step_val "nextvar" builder in

(* Compute the end condition. *)
let end_cond = codegen_expr end_ in

(* Convert condition to a bool by comparing equal to 0.0. *)
let zero = const_float double_type 0.0 in
let end_cond = build_fcmp Fcmp.One end_cond zero "loopcond" builder in

(* Create the "after loop" block and insert it. *)
let loop_end_bb = insertion_block builder in
let after_bb = append_block context "afterloop" the_function in

(* Insert the conditional branch into the end of loop_end_bb. *)
ignore (build_cond_br end_cond loop_bb after_bb builder);

(* Any new code will be inserted in after_bb. *)
position_at_end after_bb builder;

(* Add a new entry to the PHI node for the backedge. *)
add_incoming (next_var, loop_end_bb) variable;

(* Restore the unshadowed variable. *)
begin match old_val with
| Some old_val -> Hashtbl.add named_values var_name old_val
| None -> ()
end;

(* for expr always returns 0.0. *)
const_null double_type

let codegen_proto = function
| Ast.Prototype (name, args) | Ast.BinOpPrototype (name, args, _) ->
  (* Make the function type: double(double,double) etc. *)
  let doubles = Array.make (Array.length args) double_type in
  let ft = function_type double_type doubles in
  let f =
    match lookup_function name the_module with
    | None -> declare_function name ft the_module

  (* If 'f' conflicted, there was already something named 'name'. If it
  * has a body, don't allow redefinition or reextern. *)
  | Some f ->
    (* If 'f' already has a body, reject this. *)
    if block_begin f <> At_end f then
      raise (Error "redefinition of function");

    (* If 'f' took a different number of arguments, reject. *)
    if element_type (type_of f) <> ft then
      raise (Error "redefinition of function with different # args");
    f

```

```

    in

    (* Set names for all arguments. *)
    Array.iteri (fun i a ->
      let n = args.(i) in
      set_value_name n a;
      Hashtbl.add named_values n a;
    ) (params f);
    f

let codegen_func the_fpm = function
| Ast.Function (proto, body) ->
  Hashtbl.clear named_values;
  let the_function = codegen_proto proto in

  (* If this is an operator, install it. *)
  begin match proto with
  | Ast.BinOpPrototype (name, args, prec) ->
    let op = name.[String.length name - 1] in
    Hashtbl.add Parser.binop_precedence op prec;
  | _ -> ()
  end;

  (* Create a new basic block to start insertion into. *)
  let bb = append_block context "entry" the_function in
  position_at_end bb builder;

  try
    let ret_val = codegen_expr body in

    (* Finish off the function. *)
    let _ = build_ret ret_val builder in

    (* Validate the generated code, checking for consistency. *)
    Llvm_analysis.assert_valid_function the_function;

    (* Optimize the function. *)
    let _ = PassManager.run_function the_function the_fpm in

    the_function
  with e ->
    delete_function the_function;
    raise e

```

toplevel.ml:

```

(*=====
 * Top-Level parsing and JIT Driver
 *=====*)

open Llvm
open Llvm_executionengine

(* top ::= definition | external | expression | ';' *)
let rec main_loop the_fpm the_execution_engine stream =
  match Stream.peek stream with
  | None -> ()

```



```

(* ignore top-level semicolons. *)
| Some (Token.Kwd ';') ->
    Stream.junk stream;
    main_loop the_fpm the_execution_engine stream

| Some token ->
    begin
        try match token with
        | Token.Def ->
            let e = Parser.parse_definition stream in
            print_endline "parsed a function definition.";
            dump_value (Codegen.codegen_func the_fpm e);
        | Token.Extern ->
            let e = Parser.parse_extern stream in
            print_endline "parsed an extern.";
            dump_value (Codegen.codegen_proto e);
        | _ ->
            (* Evaluate a top-level expression into an anonymous function. *)
            let e = Parser.parse_toplevel stream in
            print_endline "parsed a top-level expr";
            let the_function = Codegen.codegen_func the_fpm e in
            dump_value the_function;

            (* JIT the function, returning a function pointer. *)
            let result = ExecutionEngine.run_function the_function [||]
                the_execution_engine in

            print_string "Evaluated to ";
            print_float (GenericValue.as_float Codegen.double_type result);
            print_newline ();
            with Stream.Error s | Codegen.Error s ->
                (* Skip token for error recovery. *)
                Stream.junk stream;
                print_endline s;
        end;
    print_string "ready> "; flush stdout;
    main_loop the_fpm the_execution_engine stream

```

toy.ml:

```

(=====)
* Main driver code.
(=====*)

open Llvm
open Llvm_executionengine
open Llvm_target
open Llvm_scalar_opts

let main () =
    ignore (initialize_native_target ());

    (* Install standard binary operators.
       * 1 is the lowest precedence. *)
    Hashtbl.add Parser.binop_precedence '<' 10;
    Hashtbl.add Parser.binop_precedence '+' 20;
    Hashtbl.add Parser.binop_precedence '-' 20;
    Hashtbl.add Parser.binop_precedence '*' 40;      (* highest. *)

```

```
(* Prime the first token. *)
print_string "ready> "; flush stdout;
let stream = Lexer.lex (Stream.of_channel stdin) in

(* Create the JIT. *)
let the_execution_engine = ExecutionEngine.create Codegen.the_module in
let the_fpm = PassManager.create_function Codegen.the_module in

(* Set up the optimizer pipeline. Start with registering info about how the
 * target lays out data structures. *)
DataLayout.add (ExecutionEngine.target_data the_execution_engine) the_fpm;

(* Do simple "peephole" optimizations and bit-twiddling optzn. *)
add_instruction_combination the_fpm;

(* reassociate expressions. *)
add_reassociation the_fpm;

(* Eliminate Common SubExpressions. *)
add_gvn the_fpm;

(* Simplify the control flow graph (deleting unreachable blocks, etc). *)
add_cfg_simplification the_fpm;

ignore (PassManager.initialize the_fpm);

(* Run the main "interpreter loop" now. *)
Toplevel.main_loop the_fpm the_execution_engine stream;

(* Print out all the generated code. *)
dump_module Codegen.the_module
;;

main ()
```

bindings.c

```
#include <stdio.h>

/* putchar - putchar that takes a double and returns 0. */
extern double putchar(double X) {
    putchar((char)X);
    return 0;
}

/* printd - printf that takes a double prints it as "%f\n", returning 0. */
extern double printd(double X) {
    printf("%f\n", X);
    return 0;
}
```

Next: Extending the language: mutable variables / SSA construction

Kaleidoscope: Extending the Language: Mutable Variables

- *Chapter 7 Introduction*
- *Why is this a hard problem?*
- *Memory in LLVM*
- *Mutable Variables in Kaleidoscope*
- *Adjusting Existing Variables for Mutation*
- *New Assignment Operator*
- *User-defined Local Variables*
- *Full Code Listing*

Chapter 7 Introduction

Welcome to Chapter 7 of the “Implementing a language with LLVM” tutorial. In chapters 1 through 6, we’ve built a very respectable, albeit simple, [functional programming language](#). In our journey, we learned some parsing techniques, how to build and represent an AST, how to build LLVM IR, and how to optimize the resultant code as well as JIT compile it.

While Kaleidoscope is interesting as a functional language, the fact that it is functional makes it “too easy” to generate LLVM IR for it. In particular, a functional language makes it very easy to build LLVM IR directly in [SSA form](#). Since LLVM requires that the input code be in SSA form, this is a very nice property and it is often unclear to newcomers how to generate code for an imperative language with mutable variables.

The short (and happy) summary of this chapter is that there is no need for your front-end to build SSA form: LLVM provides highly tuned and well tested support for this, though the way it works is a bit unexpected for some.

Why is this a hard problem?

To understand why mutable variables cause complexities in SSA construction, consider this extremely simple C example:

```
int G, H;
int test(_Bool Condition) {
    int X;
    if (Condition)
        X = G;
    else
        X = H;
    return X;
}
```

In this case, we have the variable “X”, whose value depends on the path executed in the program. Because there are two different possible values for X before the return instruction, a PHI node is inserted to merge the two values. The LLVM IR that we want for this example looks like this:

```
@G = weak global i32 0 ; type of @G is i32*
@H = weak global i32 0 ; type of @H is i32*

define i32 @test(i1 %Condition) {
entry:
    br i1 %Condition, label %cond_true, label %cond_false
```

```
cond_true:
    %X.0 = load i32* @G
    br label %cond_next

cond_false:
    %X.1 = load i32* @H
    br label %cond_next

cond_next:
    %X.2 = phi i32 [ %X.1, %cond_false ], [ %X.0, %cond_true ]
    ret i32 %X.2
}
```

In this example, the loads from the G and H global variables are explicit in the LLVM IR, and they live in the then/else branches of the if statement (cond_true/cond_false). In order to merge the incoming values, the X.2 phi node in the cond_next block selects the right value to use based on where control flow is coming from: if control flow comes from the cond_false block, X.2 gets the value of X.1. Alternatively, if control flow comes from cond_true, it gets the value of X.0. The intent of this chapter is not to explain the details of SSA form. For more information, see one of the many [online references](#).

The question for this article is “who places the phi nodes when lowering assignments to mutable variables?”. The issue here is that LLVM *requires* that its IR be in SSA form: there is no “non-ssa” mode for it. However, SSA construction requires non-trivial algorithms and data structures, so it is inconvenient and wasteful for every front-end to have to reproduce this logic.

Memory in LLVM

The ‘trick’ here is that while LLVM does require all register values to be in SSA form, it does not require (or permit) memory objects to be in SSA form. In the example above, note that the loads from G and H are direct accesses to G and H: they are not renamed or versioned. This differs from some other compiler systems, which do try to version memory objects. In LLVM, instead of encoding dataflow analysis of memory into the LLVM IR, it is handled with [Analysis Passes](#) which are computed on demand.

With this in mind, the high-level idea is that we want to make a stack variable (which lives in memory, because it is on the stack) for each mutable object in a function. To take advantage of this trick, we need to talk about how LLVM represents stack variables.

In LLVM, all memory accesses are explicit with load/store instructions, and it is carefully designed not to have (or need) an “address-of” operator. Notice how the type of the @G/@H global variables is actually “i32*” even though the variable is defined as “i32”. What this means is that @G defines *space* for an i32 in the global data area, but its *name* actually refers to the address for that space. Stack variables work the same way, except that instead of being declared with global variable definitions, they are declared with the [LLVM alloca instruction](#):

```
define i32 @example() {
entry:
    %X = alloca i32           ; type of %X is i32*.
    ...
    %tmp = load i32* %X       ; load the stack value %X from the stack.
    %tmp2 = add i32 %tmp, 1    ; increment it
    store i32 %tmp2, i32* %X  ; store it back
    ...
}
```

This code shows an example of how you can declare and manipulate a stack variable in the LLVM IR. Stack memory allocated with the alloca instruction is fully general: you can pass the address of the stack slot to functions, you can

store it in other variables, etc. In our example above, we could rewrite the example to use the `alloca` technique to avoid using a PHI node:

```
@G = weak global i32 0    ; type of @G is i32*
@H = weak global i32 0    ; type of @H is i32*

define i32 @test(i1 %Condition) {
entry:
    %X = alloca i32        ; type of %X is i32*.
    br i1 %Condition, label %cond_true, label %cond_false

cond_true:
    %X.0 = load i32* @G
    store i32 %X.0, i32* %X    ; Update X
    br label %cond_next

cond_false:
    %X.1 = load i32* @H
    store i32 %X.1, i32* %X    ; Update X
    br label %cond_next

cond_next:
    %X.2 = load i32* %X        ; Read X
    ret i32 %X.2
}
```

With this, we have discovered a way to handle arbitrary mutable variables without the need to create Phi nodes at all:

1. Each mutable variable becomes a stack allocation.
2. Each read of the variable becomes a load from the stack.
3. Each update of the variable becomes a store to the stack.
4. Taking the address of a variable just uses the stack address directly.

While this solution has solved our immediate problem, it introduced another one: we have now apparently introduced a lot of stack traffic for very simple and common operations, a major performance problem. Fortunately for us, the LLVM optimizer has a highly-tuned optimization pass named “`mem2reg`” that handles this case, promoting allocas like this into SSA registers, inserting Phi nodes as appropriate. If you run this example through the pass, for example, you’ll get:

```
$ llvm-as < example.ll | opt -mem2reg | llvm-dis
@G = weak global i32 0
@H = weak global i32 0

define i32 @test(i1 %Condition) {
entry:
    br i1 %Condition, label %cond_true, label %cond_false

cond_true:
    %X.0 = load i32* @G
    br label %cond_next

cond_false:
    %X.1 = load i32* @H
    br label %cond_next

cond_next:
    %X.01 = phi i32 [ %X.1, %cond_false ], [ %X.0, %cond_true ]
    ret i32 %X.01
}
```

```
ret i32 %X.01
}
```

The mem2reg pass implements the standard “iterated dominance frontier” algorithm for constructing SSA form and has a number of optimizations that speed up (very common) degenerate cases. The mem2reg optimization pass is the answer to dealing with mutable variables, and we highly recommend that you depend on it. Note that mem2reg only works on variables in certain circumstances:

1. mem2reg is alloca-driven: it looks for allocas and if it can handle them, it promotes them. It does not apply to global variables or heap allocations.
2. mem2reg only looks for alloca instructions in the entry block of the function. Being in the entry block guarantees that the alloca is only executed once, which makes analysis simpler.
3. mem2reg only promotes allocas whose uses are direct loads and stores. If the address of the stack object is passed to a function, or if any funny pointer arithmetic is involved, the alloca will not be promoted.
4. mem2reg only works on allocas of [first class](#) values (such as pointers, scalars and vectors), and only if the array size of the allocation is 1 (or missing in the .ll file). mem2reg is not capable of promoting structs or arrays to registers. Note that the “sroa” pass is more powerful and can promote structs, “unions”, and arrays in many cases.

All of these properties are easy to satisfy for most imperative languages, and we’ll illustrate it below with Kaleidoscope. The final question you may be asking is: should I bother with this nonsense for my front-end? Wouldn’t it be better if I just did SSA construction directly, avoiding use of the mem2reg optimization pass? In short, we strongly recommend that you use this technique for building SSA form, unless there is an extremely good reason not to. Using this technique is:

- Proven and well tested: clang uses this technique for local mutable variables. As such, the most common clients of LLVM are using this to handle a bulk of their variables. You can be sure that bugs are found fast and fixed early.
- Extremely Fast: mem2reg has a number of special cases that make it fast in common cases as well as fully general. For example, it has fast-paths for variables that are only used in a single block, variables that only have one assignment point, good heuristics to avoid insertion of unneeded phi nodes, etc.
- Needed for debug info generation: [Debug information in LLVM](#) relies on having the address of the variable exposed so that debug info can be attached to it. This technique dovetails very naturally with this style of debug info.

If nothing else, this makes it much easier to get your front-end up and running, and is very simple to implement. Lets extend Kaleidoscope with mutable variables now!

Mutable Variables in Kaleidoscope

Now that we know the sort of problem we want to tackle, lets see what this looks like in the context of our little Kaleidoscope language. We’re going to add two features:

1. The ability to mutate variables with the ‘=’ operator.
2. The ability to define new variables.

While the first item is really what this is about, we only have variables for incoming arguments as well as for induction variables, and redefining those only goes so far :). Also, the ability to define new variables is a useful thing regardless of whether you will be mutating them. Here’s a motivating example that shows how we could use these:

```
# Define ':' for sequencing: as a low-precedence operator that ignores operands
# and just returns the RHS.
def binary : 1 (x y) y;
```

```

# Recursive fib, we could do this before.
def fib(x)
  if (x < 3) then
    1
  else
    fib(x-1)+fib(x-2);

# Iterative fib.
def fibi(x)
  var a = 1, b = 1, c in
  (for i = 3, i < x in
    c = a + b :
    a = b :
    b = c) :
  b;

# Call it.
fibi(10);

```

In order to mutate variables, we have to change our existing variables to use the “alloca trick”. Once we have that, we’ll add our new operator, then extend Kaleidoscope to support new variable definitions.

Adjusting Existing Variables for Mutation

The symbol table in Kaleidoscope is managed at code generation time by the ‘named_values’ map. This map currently keeps track of the LLVM “Value*” that holds the double value for the named variable. In order to support mutation, we need to change this slightly, so that it named_values holds the *memory location* of the variable in question. Note that this change is a refactoring: it changes the structure of the code, but does not (by itself) change the behavior of the compiler. All of these changes are isolated in the Kaleidoscope code generator.

At this point in Kaleidoscope’s development, it only supports variables for two things: incoming arguments to functions and the induction variable of ‘for’ loops. For consistency, we’ll allow mutation of these variables in addition to other user-defined variables. This means that these will both need memory locations.

To start our transformation of Kaleidoscope, we’ll change the named_values map so that it maps to AllocInst* instead of Value*. Once we do this, the C++ compiler will tell us what parts of the code we need to update:

Note: the ocaml bindings currently model both Value*’s and AllocInst*’s as Llvm.llvalue’s, but this may change in the future to be more type safe.

```

let named_values:(string, llvalue) Hashtbl.t = Hashtbl.create 10

```

Also, since we will need to create these alloca’s, we’ll use a helper function that ensures that the allocas are created in the entry block of the function:

```

(* Create an alloca instruction in the entry block of the function. This
 * is used for mutable variables etc. *)
let create_entry_block_alloca the_function var_name =
  let builder = builder_at (instr_begin (entry_block the_function)) in
  build_alloca double_type var_name builder

```

This funny looking code creates an Llvm.llbuilder object that is pointing at the first instruction of the entry block. It then creates an alloca with the expected name and returns it. Because all values in Kaleidoscope are doubles, there is no need to pass in a type to use.

With this in place, the first functionality change we want to make is to variable references. In our new scheme, variables live on the stack, so code generating a reference to them actually needs to produce a load from the stack slot:

```
let rec codegen_expr = function
...
| Ast.Variable name ->
    let v = try Hashtbl.find named_values name with
    | Not_found -> raise (Error "unknown variable name")
    in
    (* Load the value. *)
    build_load v name builder
```

As you can see, this is pretty straightforward. Now we need to update the things that define the variables to set up the alloca. We'll start with `codegen_expr Ast.For ...` (see the *full code listing* for the unabridged code):

```
| Ast.For (var_name, start, end_, step, body) ->
    let the_function = block_parent (insertion_block builder) in

    (* Create an alloca for the variable in the entry block. *)
    let alloca = create_entry_block_alloca the_function var_name in

    (* Emit the start code first, without 'variable' in scope. *)
    let start_val = codegen_expr start in

    (* Store the value into the alloca. *)
    ignore(build_store start_val alloca builder);

    ...

    (* Within the loop, the variable is defined equal to the PHI node. If it
     * shadows an existing variable, we have to restore it, so save it
     * now. *)
    let old_val =
        try Some (Hashtbl.find named_values var_name) with Not_found -> None
    in
    Hashtbl.add named_values var_name alloca;

    ...

    (* Compute the end condition. *)
    let end_cond = codegen_expr end_ in

    (* Reload, increment, and restore the alloca. This handles the case where
     * the body of the loop mutates the variable. *)
    let cur_var = build_load alloca var_name builder in
    let next_var = build_add cur_var step_val "nextvar" builder in
    ignore(build_store next_var alloca builder);

    ...
```

This code is virtually identical to the code before we allowed mutable variables. The big difference is that we no longer have to construct a PHI node, and we use load/store to access the variable as needed.

To support mutable argument variables, we need to also make allocas for them. The code for this is also pretty simple:

```
(* Create an alloca for each argument and register the argument in the symbol
 * table so that references to it will succeed. *)
let create_argument_allocas the_function proto =
    let args = match proto with
    | Ast.Prototype (_, args) | Ast.BinOpPrototype (_, args, _) -> args
```



```

in
Array.iteri (fun i ai ->
  let var_name = args.(i) in
  (* Create an alloca for this variable. *)
  let alloca = create_entry_block_alloca the_function var_name in

  (* Store the initial value into the alloca. *)
  ignore(build_store ai alloca builder);

  (* Add arguments to variable symbol table. *)
  Hashtbl.add named_values var_name alloca;
) (params the_function)

```

For each argument, we make an alloca, store the input value to the function into the alloca, and register the alloca as the memory location for the argument. This method gets invoked by `Codegen.codegen_func` right after it sets up the entry block for the function.

The final missing piece is adding the `mem2reg` pass, which allows us to get good codegen once again:

```

let main () =
  ...
  let the_fpm = PassManager.create_function Codegen.the_module in

  (* Set up the optimizer pipeline. Start with registering info about how the
   * target lays out data structures. *)
  DataLayout.add (ExecutionEngine.target_data the_execution_engine) the_fpm;

  (* Promote allocas to registers. *)
  add_memory_to_register_promotion the_fpm;

  (* Do simple "peephole" optimizations and bit-twiddling optzn. *)
  add_instruction_combining the_fpm;

  (* reassociate expressions. *)
  add_reassociation the_fpm;

```

It is interesting to see what the code looks like before and after the `mem2reg` optimization runs. For example, this is the before/after code for our recursive `fib` function. Before the optimization:

```

define double @fib(double %x) {
entry:
  %x1 = alloca double
  store double %x, double* %x1
  %x2 = load double* %x1
  %cmptmp = fcmp ult double %x2, 3.000000e+00
  %booltmp = uitofp i1 %cmptmp to double
  %ifcond = fcmp one double %booltmp, 0.000000e+00
  br i1 %ifcond, label %then, label %else

then:    ; preds = %entry
  br label %ifcont

else:    ; preds = %entry
  %x3 = load double* %x1
  %subtmp = fsub double %x3, 1.000000e+00
  %calltmp = call double @fib(double %subtmp)
  %x4 = load double* %x1
  %subtmp5 = fsub double %x4, 2.000000e+00

```

```
%calltmp6 = call double @fib(double %subtmp5)
%addtmp = fadd double %calltmp, %calltmp6
br label %ifcont

ifcont:    ; preds = %else, %then
%iftmp = phi double [ 1.000000e+00, %then ], [ %addtmp, %else ]
ret double %iftmp
}
```

Here there is only one variable (`x`, the input argument) but you can still see the extremely simple-minded code generation strategy we are using. In the entry block, an alloca is created, and the initial input value is stored into it. Each reference to the variable does a reload from the stack. Also, note that we didn't modify the if/then/else expression, so it still inserts a PHI node. While we could make an alloca for it, it is actually easier to create a PHI node for it, so we still just make the PHI.

Here is the code after the `mem2reg` pass runs:

```
define double @fib(double %x) {
entry:
    %cmptmp = fcmp ult double %x, 3.000000e+00
    %booltmp = uitofp il %cmptmp to double
    %ifcond = fcmp one double %booltmp, 0.000000e+00
    br il %ifcond, label %then, label %else

then:
    br label %ifcont

else:
    %subtmp = fsub double %x, 1.000000e+00
    %calltmp = call double @fib(double %subtmp)
    %subtmp5 = fsub double %x, 2.000000e+00
    %calltmp6 = call double @fib(double %subtmp5)
    %addtmp = fadd double %calltmp, %calltmp6
    br label %ifcont

ifcont:    ; preds = %else, %then
%iftmp = phi double [ 1.000000e+00, %then ], [ %addtmp, %else ]
ret double %iftmp
}
```

This is a trivial case for `mem2reg`, since there are no redefinitions of the variable. The point of showing this is to calm your tension about inserting such blatant inefficiencies :).

After the rest of the optimizers run, we get:

```
define double @fib(double %x) {
entry:
    %cmptmp = fcmp ult double %x, 3.000000e+00
    %booltmp = uitofp il %cmptmp to double
    %ifcond = fcmp ueq double %booltmp, 0.000000e+00
    br il %ifcond, label %else, label %ifcont

else:
    %subtmp = fsub double %x, 1.000000e+00
    %calltmp = call double @fib(double %subtmp)
    %subtmp5 = fsub double %x, 2.000000e+00
    %calltmp6 = call double @fib(double %subtmp5)
    %addtmp = fadd double %calltmp, %calltmp6
```

```

    ret double %addtmp

ifcont:
    ret double 1.000000e+00
}

```

Here we see that the `simplifcfg` pass decided to clone the return instruction into the end of the ‘else’ block. This allowed it to eliminate some branches and the PHI node.

Now that all symbol table references are updated to use stack variables, we’ll add the assignment operator.

New Assignment Operator

With our current framework, adding a new assignment operator is really simple. We will parse it just like any other binary operator, but handle it internally (instead of allowing the user to define it). The first step is to set a precedence:

```

let main () =
  (* Install standard binary operators.
   * 1 is the lowest precedence. *)
  Hashtbl.add Parser.binop_precedence '=' 2;
  Hashtbl.add Parser.binop_precedence '<' 10;
  Hashtbl.add Parser.binop_precedence '+' 20;
  Hashtbl.add Parser.binop_precedence '-' 20;
  ...

```

Now that the parser knows the precedence of the binary operator, it takes care of all the parsing and AST generation. We just need to implement codegen for the assignment operator. This looks like:

```

let rec codegen_expr = function
  begin match op with
  | '=' ->
    (* Special case '=' because we don't want to emit the LHS as an
     * expression. *)
    let name =
      match lhs with
      | Ast.Variable name -> name
      | _ -> raise (Error "destination of '=' must be a variable")
    in

```

Unlike the rest of the binary operators, our assignment operator doesn’t follow the “emit LHS, emit RHS, do computation” model. As such, it is handled as a special case before the other binary operators are handled. The other strange thing is that it requires the LHS to be a variable. It is invalid to have “(x+1) = expr” - only things like “x = expr” are allowed.

```

  (* Codegen the rhs. *)
  let val_ = codegen_expr rhs in

  (* Lookup the name. *)
  let variable = try Hashtbl.find named_values name with
  | Not_found -> raise (Error "unknown variable name")
  in
  ignore(build_store val_ variable builder);
  val_
| _ ->
  ...

```

Once we have the variable, codegen'ing the assignment is straightforward: we emit the RHS of the assignment, create a store, and return the computed value. Returning a value allows for chained assignments like “X = (Y = Z)”.

Now that we have an assignment operator, we can mutate loop variables and arguments. For example, we can now run code like this:

```
# Function to print a double.
extern printd(x);

# Define ':' for sequencing: as a low-precedence operator that ignores operands
# and just returns the RHS.
def binary : l (x y) y;

def test(x)
  printd(x) :
  x = 4 :
  printd(x);

test(123);
```

When run, this example prints “123” and then “4”, showing that we did actually mutate the value! Okay, we have now officially implemented our goal: getting this to work requires SSA construction in the general case. However, to be really useful, we want the ability to define our own local variables, lets add this next!

User-defined Local Variables

Adding var/in is just like any other other extensions we made to Kaleidoscope: we extend the lexer, the parser, the AST and the code generator. The first step for adding our new ‘var/in’ construct is to extend the lexer. As before, this is pretty trivial, the code looks like this:

```
type token =
  ...
  (* var definition *)
  | Var

...

and lex_ident buffer = parser
  ...
  | "in" -> [< 'Token.In; stream >]
  | "binary" -> [< 'Token.Binary; stream >]
  | "unary" -> [< 'Token.Unary; stream >]
  | "var" -> [< 'Token.Var; stream >]
  ...
```

The next step is to define the AST node that we will construct. For var/in, it looks like this:

```
type expr =
  ...
  (* variant for var/in. *)
  | Var of (string * expr option) array * expr
  ...
```

var/in allows a list of names to be defined all at once, and each name can optionally have an initializer value. As such, we capture this information in the VarNames vector. Also, var/in has a body, this body is allowed to access the variables defined by the var/in.

With this in place, we can define the parser pieces. The first thing we do is add it as a primary expression:

```
(* primary
 * ::= identifier
 * ::= numberexpr
 * ::= parenexpr
 * ::= ifexpr
 * ::= forexpr
 * ::= varexpr *)
let rec parse_primary = parser
...
(* varexpr
 * ::= 'var' identifier ('=' expression?
 *      (',' identifier ('=' expression)?)* 'in' expression *)
| [< 'Token.Var;
   (* At least one variable name is required. *)
   'Token.Ident id ?? "expected identifier after var";
   init=parse_var_init;
   var_names=parse_var_names [(id, init)];
   (* At this point, we have to have 'in'. *)
   'Token.In ?? "expected 'in' keyword after 'var'";
   body=parse_expr >] ->
   Ast.Var (Array.of_list (List.rev var_names), body)
...

and parse_var_init = parser
(* read in the optional initializer. *)
| [< 'Token.Kwd '='; e=parse_expr >] -> Some e
| [< >] -> None

and parse_var_names accumulator = parser
| [< 'Token.Kwd ',';
   'Token.Ident id ?? "expected identifier list after var";
   init=parse_var_init;
   e=parse_var_names ((id, init) :: accumulator) >] -> e
| [< >] -> accumulator
```

Now that we can parse and represent the code, we need to support emission of LLVM IR for it. This code starts out with:

```
let rec codegen_expr = function
...
| Ast.Var (var_names, body)
  let old_bindings = ref [] in

  let the_function = block_parent (insertion_block builder) in

  (* Register all variables and emit their initializer. *)
  Array.iter (fun (var_name, init) ->
```

Basically it loops over all the variables, installing them one at a time. For each variable we put into the symbol table, we remember the previous value that we replace in OldBindings.

```
(* Emit the initializer before adding the variable to scope, this
 * prevents the initializer from referencing the variable itself, and
 * permits stuff like this:
 *   var a = 1 in
```

```

    *      var a = a in ...    # refers to outer 'a'. *)
let init_val =
  match init with
  | Some init -> codegen_expr init
  (* If not specified, use 0.0. *)
  | None -> const_float double_type 0.0
in

let alloca = create_entry_block_alloc the_function var_name in
ignore(build_store init_val alloca builder);

(* Remember the old variable binding so that we can restore the binding
 * when we unrecurse. *)

begin
  try
    let old_value = Hashtbl.find named_values var_name in
    old_bindings := (var_name, old_value) :: !old_bindings;
    with Not_found > ()
  end;

  (* Remember this binding. *)
  Hashtbl.add named_values var_name alloca;
) var_names;

```

There are more comments here than code. The basic idea is that we emit the initializer, create the alloca, then update the symbol table to point to it. Once all the variables are installed in the symbol table, we evaluate the body of the var/in expression:

```

(* Codegen the body, now that all vars are in scope. *)
let body_val = codegen_expr body in

```

Finally, before returning, we restore the previous variable bindings:

```

(* Pop all our variables from scope. *)
List.iter (fun (var_name, old_value) ->
  Hashtbl.add named_values var_name old_value
) !old_bindings;

(* Return the body computation. *)
body_val

```

The end result of all of this is that we get properly scoped variable definitions, and we even (trivially) allow mutation of them :).

With this, we completed what we set out to do. Our nice iterative fib example from the intro compiles and runs just fine. The mem2reg pass optimizes all of our stack variables into SSA registers, inserting PHI nodes where needed, and our front-end remains simple: no “iterated dominance frontier” computation anywhere in sight.

Full Code Listing

Here is the complete code listing for our running example, enhanced with mutable variables and var/in support. To build this example, use:

```

# Compile
ocamlbuild toy.byte

```

```
# Run
./toy.byte
```

Here is the code:

_tags:

```
<{lexer,parser}.ml>: use_camlp4, pp(camlp4of)
<*. {byte,native}>: g++, use_llvm, use_llvm_analysis
<*. {byte,native}>: use_llvm_executionengine, use_llvm_target
<*. {byte,native}>: use_llvm_scalar_opts, use_bindings
```

myocamlbuild.ml:

```
open Ocamlbuild_plugin;;

ocaml_lib ~extern:true "llvm";;
ocaml_lib ~extern:true "llvm_analysis";;
ocaml_lib ~extern:true "llvm_executionengine";;
ocaml_lib ~extern:true "llvm_target";;
ocaml_lib ~extern:true "llvm_scalar_opts";;

flag ["link"; "ocaml"; "g++"] (S[A"-cc"; A"g++"; A"-cclib"; A"-rdynamic"]);;
dep ["link"; "ocaml"; "use_bindings"] ["bindings.o"];;
```

token.ml:

```
(=====
 * Lexer Tokens
 *=====)

(* The lexer returns these 'Kwd' if it is an unknown character, otherwise one of
 * these others for known things. *)
type token =
  (* commands *)
  | Def | Extern

  (* primary *)
  | Ident of string | Number of float

  (* unknown *)
  | Kwd of char

  (* control *)
  | If | Then | Else
  | For | In

  (* operators *)
  | Binary | Unary

  (* var definition *)
  | Var
```

lexer.ml:

```
(=====
 * Lexer
 *=====)
```

```

let rec lex = parser
  (* Skip any whitespace. *)
  | [< ' (' | '\n' | '\r' | '\t'); stream >] -> lex stream

  (* identifier: [a-zA-Z][a-zA-Z0-9] *)
  | [< ' ('A' .. 'Z' | 'a' .. 'z' as c); stream >] ->
    let buffer = Buffer.create 1 in
    Buffer.add_char buffer c;
    lex_ident buffer stream

  (* number: [0-9.]+ *)
  | [< ' ('0' .. '9' as c); stream >] ->
    let buffer = Buffer.create 1 in
    Buffer.add_char buffer c;
    lex_number buffer stream

  (* Comment until end of line. *)
  | [< ' ('#'); stream >] ->
    lex_comment stream

  (* Otherwise, just return the character as its ascii value. *)
  | [< 'c; stream >] ->
    [< 'Token.Kwd c; lex stream >]

  (* end of stream. *)
  | [< >] -> [< >]

and lex_number buffer = parser
  | [< ' ('0' .. '9' | '.' as c); stream >] ->
    Buffer.add_char buffer c;
    lex_number buffer stream
  | [< stream=lex >] ->
    [< 'Token.Number (float_of_string (Buffer.contents buffer)); stream >]

and lex_ident buffer = parser
  | [< ' ('A' .. 'Z' | 'a' .. 'z' | '0' .. '9' as c); stream >] ->
    Buffer.add_char buffer c;
    lex_ident buffer stream
  | [< stream=lex >] ->
    match Buffer.contents buffer with
    | "def" -> [< 'Token.Def; stream >]
    | "extern" -> [< 'Token.Extern; stream >]
    | "if" -> [< 'Token.If; stream >]
    | "then" -> [< 'Token.Then; stream >]
    | "else" -> [< 'Token.Else; stream >]
    | "for" -> [< 'Token.For; stream >]
    | "in" -> [< 'Token.In; stream >]
    | "binary" -> [< 'Token.Binary; stream >]
    | "unary" -> [< 'Token.Unary; stream >]
    | "var" -> [< 'Token.Var; stream >]
    | id -> [< 'Token.Ident id; stream >]

and lex_comment = parser
  | [< ' ('\n'); stream=lex >] -> stream
  | [< 'c; e=lex_comment >] -> e
  | [< >] -> [< >]

```


ast.ml:

```

(*=====
 * Abstract Syntax Tree (aka Parse Tree)
 *=====*)

(* expr - Base type for all expression nodes. *)
type expr =
  (* variant for numeric literals like "1.0". *)
  | Number of float

  (* variant for referencing a variable, like "a". *)
  | Variable of string

  (* variant for a unary operator. *)
  | Unary of char * expr

  (* variant for a binary operator. *)
  | Binary of char * expr * expr

  (* variant for function calls. *)
  | Call of string * expr array

  (* variant for if/then/else. *)
  | If of expr * expr * expr

  (* variant for for/in. *)
  | For of string * expr * expr * expr option * expr

  (* variant for var/in. *)
  | Var of (string * expr option) array * expr

(* proto - This type represents the "prototype" for a function, which captures
 * its name, and its argument names (thus implicitly the number of arguments the
 * function takes). *)
type proto =
  | Prototype of string * string array
  | BinOpPrototype of string * string array * int

(* func - This type represents a function definition itself. *)
type func = Function of proto * expr

```

parser.ml:

```

(*=====
 * Parser
 *=====*)

(* binop_precedence - This holds the precedence for each binary operator that is
 * defined *)
let binop_precedence: (char, int) Hashtbl.t = Hashtbl.create 10

(* precedence - Get the precedence of the pending binary operator token. *)
let precedence c = try Hashtbl.find binop_precedence c with Not_found -> -1

(* primary
 * ::= identifier
 * ::= numberexpr
 * ::= parenexpr

```

```

* ::= ifexpr
* ::= forexpr
* ::= varexpr *)
let rec parse_primary = parser
  (* numberexpr ::= number *)
  | [< 'Token.Number n >] -> Ast.Number n

  (* parenexpr ::= '(' expression ')' *)
  | [< 'Token.Kwd '('; e=parse_expr; 'Token.Kwd ')' ?? "expected ')'" >] -> e

  (* identifierexpr
  * ::= identifier
  * ::= identifier '(' argumentexpr ')' *)
  | [< 'Token.Ident id; stream >] ->
    let rec parse_args accumulator = parser
      | [< e=parse_expr; stream >] ->
        begin parser
          | [< 'Token.Kwd ','; e=parse_args (e :: accumulator) >] -> e
          | [< >] -> e :: accumulator
        end stream
      | [< >] -> accumulator
    in
    let rec parse_ident id = parser
      (* Call. *)
      | [< 'Token.Kwd '(';
        args=parse_args [];
        'Token.Kwd ')' ?? "expected ')'">] ->
        Ast.Call (id, Array.of_list (List.rev args))

      (* Simple variable ref. *)
      | [< >] -> Ast.Variable id
    in
    parse_ident id stream

  (* ifexpr ::= 'if' expr 'then' expr 'else' expr *)
  | [< 'Token.If; c=parse_expr;
    'Token.Then ?? "expected 'then'"; t=parse_expr;
    'Token.Else ?? "expected 'else'"; e=parse_expr >] ->
    Ast.If (c, t, e)

  (* forexpr
  ::= 'for' identifier '=' expr ',' expr (',' expr)? 'in' expression *)
  | [< 'Token.For;
    'Token.Ident id ?? "expected identifier after for";
    'Token.Kwd '=' ?? "expected '=' after for";
    stream >] ->
    begin parser
      | [<
        start=parse_expr;
        'Token.Kwd ',' ?? "expected ',' after for";
        end_=parse_expr;
        stream >] ->
        let step =
          begin parser
            | [< 'Token.Kwd ','; step=parse_expr >] -> Some step
            | [< >] -> None
          end stream
        in

```

```

begin parser
| [< 'Token.In; body=parse_expr >] ->
  Ast.For (id, start, end_, step, body)
| [< >] ->
  raise (Stream.Error "expected 'in' after for")
end stream
| [< >] ->
  raise (Stream.Error "expected '=' after for")
end stream

(* varexpr
* ::= 'var' identifier ('=' expression?
*      (' identifier ('=' expression)?)* 'in' expression *)
| [< 'Token.Var;
  (* At least one variable name is required. *)
  'Token.Ident id ?? "expected identifier after var";
  init=parse_var_init;
  var_names=parse_var_names [(id, init)];
  (* At this point, we have to have 'in'. *)
  'Token.In ?? "expected 'in' keyword after 'var'";
  body=parse_expr >] ->
  Ast.Var (Array.of_list (List.rev var_names), body)

| [< >] -> raise (Stream.Error "unknown token when expecting an expression.")

(* unary
* ::= primary
* ::= '!' unary *)
and parse_unary = parser
(* If this is a unary operator, read it. *)
| [< 'Token.Kwd op when op != '(' && op != ')'; operand=parse_expr >] ->
  Ast.Unary (op, operand)

(* If the current token is not an operator, it must be a primary expr. *)
| [< stream >] -> parse_primary stream

(* binoprhs
* ::= ('+' primary)* *)
and parse_bin_rhs expr_prec lhs stream =
  match Stream.peek stream with
  (* If this is a binop, find its precedence. *)
  | Some (Token.Kwd c) when Hashtbl.mem binop_precedence c ->
    let token_prec = precedence c in

    (* If this is a binop that binds at least as tightly as the current binop,
    * consume it, otherwise we are done. *)
    if token_prec < expr_prec then lhs else begin
      (* Eat the binop. *)
      Stream.junk stream;

      (* Parse the primary expression after the binary operator. *)
      let rhs = parse_unary stream in

      (* Okay, we know this is a binop. *)
      let rhs =
        match Stream.peek stream with
        | Some (Token.Kwd c2) ->
          (* If BinOp binds less tightly with rhs than the operator after

```

```

        * rhs, let the pending operator take rhs as its lhs. *)
    let next_prec = precedence c2 in
    if token_prec < next_prec
    then parse_bin_rhs (token_prec + 1) rhs stream
    else rhs
    | _ -> rhs
in

    (* Merge lhs/rhs. *)
    let lhs = Ast.Binary (c, lhs, rhs) in
    parse_bin_rhs expr_prec lhs stream
end
| _ -> lhs

and parse_var_init = parser
    (* read in the optional initializer. *)
    | [< 'Token.Kwd '='; e=parse_expr >] -> Some e
    | [< >] -> None

and parse_var_names accumulator = parser
    | [< 'Token.Kwd ',';
        'Token.Ident id ?? "expected identifier list after var";
        init=parse_var_init;
        e=parse_var_names ((id, init) :: accumulator) >] -> e
    | [< >] -> accumulator

(* expression
 * ::= primary binoprhs *)
and parse_expr = parser
    | [< lhs=parse_unary; stream >] -> parse_bin_rhs 0 lhs stream

(* prototype
 * ::= id '(' id* ')'
 * ::= binary LETTER number? (id, id)
 * ::= unary LETTER number? (id) *)
let parse_prototype =
    let rec parse_args accumulator = parser
        | [< 'Token.Ident id; e=parse_args (id::accumulator) >] -> e
        | [< >] -> accumulator
    in
    let parse_operator = parser
        | [< 'Token.Unary >] -> "unary", 1
        | [< 'Token.Binary >] -> "binary", 2
    in
    let parse_binary_precedence = parser
        | [< 'Token.Number n >] -> int_of_float n
        | [< >] -> 30
    in
    parser
    | [< 'Token.Ident id;
        'Token.Kwd '(' ?? "expected '(' in prototype";
        args=parse_args [];
        'Token.Kwd ')' ?? "expected ')' in prototype" >] ->
        (* success. *)
        Ast.Prototype (id, Array.of_list (List.rev args))
    | [< (prefix, kind)=parse_operator;
        'Token.Kwd op ?? "expected an operator";
        (* Read the precedence if present. *)
```

```

    binary_precedence=parse_binary_precedence;
    'Token.Kwd '(' ?? "expected '(' in prototype";
    args=parse_args [];
    'Token.Kwd ')' ?? "expected ')' in prototype" >] ->
let name = prefix ^ (String.make 1 op) in
let args = Array.of_list (List.rev args) in

(* Verify right number of arguments for operator. *)
if Array.length args != kind
then raise (Stream.Error "invalid number of operands for operator")
else
    if kind == 1 then
        Ast.Prototype (name, args)
    else
        Ast.BinOpPrototype (name, args, binary_precedence)
| [< >] ->
    raise (Stream.Error "expected function name in prototype")

(* definition ::= 'def' prototype expression *)
let parse_definition = parser
| [< 'Token.Def; p=parse_prototype; e=parse_expr >] ->
    Ast.Function (p, e)

(* toplevelexpr ::= expression *)
let parse_toplevel = parser
| [< e=parse_expr >] ->
    (* Make an anonymous proto. *)
    Ast.Function (Ast.Prototype ("", [||]), e)

(* external ::= 'extern' prototype *)
let parse_extern = parser
| [< 'Token.Extern; e=parse_prototype >] -> e

```

codegen.ml:

```

(*=====
 * Code Generation
 *=====*)

open Llvml

exception Error of string

let context = global_context ()
let the_module = create_module context "my cool jit"
let builder = builder context
let named_values:(string, llvalue) Hashtbl.t = Hashtbl.create 10
let double_type = double_type context

(* Create an alloca instruction in the entry block of the function. This
 * is used for mutable variables etc. *)
let create_entry_block_alloca the_function var_name =
    let builder = builder_at context (instr_begin (entry_block the_function)) in
    build_alloca double_type var_name builder

let rec codegen_expr = function
| Ast.Number n -> const_float double_type n
| Ast.Variable name ->

```

```

    let v = try Hashtbl.find named_values name with
    | Not_found -> raise (Error "unknown variable name")
  in
    (* Load the value. *)
    build_load v name builder
| Ast.Unary (op, operand) ->
  let operand = codegen_expr operand in
  let callee = "unary" ^ (String.make 1 op) in
  let callee =
    match lookup_function callee the_module with
    | Some callee -> callee
    | None -> raise (Error "unknown unary operator")
  in
    build_call callee [|operand|] "unop" builder
| Ast.Binary (op, lhs, rhs) ->
  begin match op with
  | '=' ->
    (* Special case '=' because we don't want to emit the LHS as an
     * expression. *)
    let name =
      match lhs with
      | Ast.Variable name -> name
      | _ -> raise (Error "destination of '=' must be a variable")
    in
      (* Codegen the rhs. *)
      let val_ = codegen_expr rhs in

      (* Lookup the name. *)
      let variable = try Hashtbl.find named_values name with
      | Not_found -> raise (Error "unknown variable name")
      in
        ignore(build_store val_ variable builder);
        val_
    | _ ->
      let lhs_val = codegen_expr lhs in
      let rhs_val = codegen_expr rhs in
      begin
        match op with
        | '+' -> build_add lhs_val rhs_val "addtmp" builder
        | '-' -> build_sub lhs_val rhs_val "subtmp" builder
        | '*' -> build_mul lhs_val rhs_val "multmp" builder
        | '<' ->
          (* Convert bool 0/1 to double 0.0 or 1.0 *)
          let i = build_fcmp Fcmp.Ult lhs_val rhs_val "cmptmp" builder in
          build_uitofp i double_type "booltmp" builder
        | _ ->
          (* If it wasn't a builtin binary operator, it must be a user_
      ↪defined
          * one. Emit a call to it. *)
          let callee = "binary" ^ (String.make 1 op) in
          let callee =
            match lookup_function callee the_module with
            | Some callee -> callee
            | None -> raise (Error "binary operator not found!")
          in
            build_call callee [|lhs_val; rhs_val|] "binop" builder
      end
  end

```

```

    end
| Ast.Call (callee, args) ->
    (* Look up the name in the module table. *)
    let callee =
        match lookup_function callee the_module with
        | Some callee -> callee
        | None -> raise (Error "unknown function referenced")
    in
    let params = params callee in

    (* If argument mismatch error. *)
    if Array.length params == Array.length args then () else
        raise (Error "incorrect # arguments passed");
    let args = Array.map codegen_expr args in
    build_call callee args "calltmp" builder
| Ast.If (cond, then_, else_) ->
    let cond = codegen_expr cond in

    (* Convert condition to a bool by comparing equal to 0.0 *)
    let zero = const_float double_type 0.0 in
    let cond_val = build_fcmp Fcmp.One cond zero "ifcond" builder in

    (* Grab the first block so that we might later add the conditional branch
     * to it at the end of the function. *)
    let start_bb = insertion_block builder in
    let the_function = block_parent start_bb in

    let then_bb = append_block context "then" the_function in

    (* Emit 'then' value. *)
    position_at_end then_bb builder;
    let then_val = codegen_expr then_ in

    (* Codegen of 'then' can change the current block, update then_bb for the
     * phi. We create a new name because one is used for the phi node, and the
     * other is used for the conditional branch. *)
    let new_then_bb = insertion_block builder in

    (* Emit 'else' value. *)
    let else_bb = append_block context "else" the_function in
    position_at_end else_bb builder;
    let else_val = codegen_expr else_ in

    (* Codegen of 'else' can change the current block, update else_bb for the
     * phi. *)
    let new_else_bb = insertion_block builder in

    (* Emit merge block. *)
    let merge_bb = append_block context "ifcont" the_function in
    position_at_end merge_bb builder;
    let incoming = [(then_val, new_then_bb); (else_val, new_else_bb)] in
    let phi = build_phi incoming "iftmp" builder in

    (* Return to the start block to add the conditional branch. *)
    position_at_end start_bb builder;
    ignore (build_cond_br cond_val then_bb else_bb builder);

    (* Set a unconditional branch at the end of the 'then' block and the

```

```

    * 'else' block to the 'merge' block. *)
position_at_end new_then_bb builder; ignore (build_br merge_bb builder);
position_at_end new_else_bb builder; ignore (build_br merge_bb builder);

(* Finally, set the builder to the end of the merge block. *)
position_at_end merge_bb builder;

phi
| Ast.For (var_name, start, end_, step, body) ->
  (* Output this as:
   *   var = alloca double
   *   ...
   *   start = startexpr
   *   store start -> var
   *   goto loop
   * loop:
   *   ...
   *   bodyexpr
   *   ...
   * loopend:
   *   step = stepexpr
   *   endcond = endexpr
   *
   *   curvar = load var
   *   nextvar = curvar + step
   *   store nextvar -> var
   *   br endcond, loop, endloop
   * outloop: *)

let the_function = block_parent (insertion_block builder) in

  (* Create an alloca for the variable in the entry block. *)
let alloca = create_entry_block_alloca the_function var_name in

  (* Emit the start code first, without 'variable' in scope. *)
let start_val = codegen_expr start in

  (* Store the value into the alloca. *)
  ignore(build_store start_val alloca builder);

  (* Make the new basic block for the loop header, inserting after current
   * block. *)
let loop_bb = append_block context "loop" the_function in

  (* Insert an explicit fall through from the current block to the
   * loop_bb. *)
  ignore (build_br loop_bb builder);

  (* Start insertion in loop_bb. *)
  position_at_end loop_bb builder;

  (* Within the loop, the variable is defined equal to the PHI node. If it
   * shadows an existing variable, we have to restore it, so save it
   * now. *)
let old_val =
  try Some (Hashtbl.find named_values var_name) with Not_found -> None
in
  Hashtbl.add named_values var_name alloca;

```



```

(* Emit the body of the loop. This, like any other expr, can change the
 * current BB. Note that we ignore the value computed by the body, but
 * don't allow an error *)
ignore (codegen_expr body);

(* Emit the step value. *)
let step_val =
  match step with
  | Some step -> codegen_expr step
  (* If not specified, use 1.0. *)
  | None -> const_float double_type 1.0
in

(* Compute the end condition. *)
let end_cond = codegen_expr end_ in

(* Reload, increment, and restore the alloca. This handles the case where
 * the body of the loop mutates the variable. *)
let cur_var = build_load alloca var_name builder in
let next_var = build_add cur_var step_val "nextvar" builder in
ignore(build_store next_var alloca builder);

(* Convert condition to a bool by comparing equal to 0.0. *)
let zero = const_float double_type 0.0 in
let end_cond = build_fcmp Fcmp.One end_cond zero "loopcond" builder in

(* Create the "after loop" block and insert it. *)
let after_bb = append_block context "afterloop" the_function in

(* Insert the conditional branch into the end of loop_end_bb. *)
ignore (build_cond_br end_cond loop_bb after_bb builder);

(* Any new code will be inserted in after_bb. *)
position_at_end after_bb builder;

(* Restore the unshadowed variable. *)
begin match old_val with
| Some old_val -> Hashtbl.add named_values var_name old_val
| None -> ()
end;

(* for expr always returns 0.0. *)
const_null double_type
| Ast.Var (var_names, body) ->
  let old_bindings = ref [] in

  let the_function = block_parent (insertion_block builder) in

  (* Register all variables and emit their initializer. *)
  Array.iter (fun (var_name, init) ->
    (* Emit the initializer before adding the variable to scope, this
     * prevents the initializer from referencing the variable itself, and
     * permits stuff like this:
     *   var a = 1 in
     *     var a = a in ... # refers to outer 'a'. *)
    let init_val =
      match init with

```

```

    | Some init -> codegen_expr init
    (* If not specified, use 0.0. *)
    | None -> const_float double_type 0.0
  in

  let alloca = create_entry_block_alloca the_function var_name in
  ignore(build_store init_val alloca builder);

  (* Remember the old variable binding so that we can restore the binding
   * when we unrecurse. *)
  begin
    try
      let old_value = Hashtbl.find named_values var_name in
      old_bindings := (var_name, old_value) :: !old_bindings;
      with Not_found -> ()
    end;

    (* Remember this binding. *)
    Hashtbl.add named_values var_name alloca;
  ) var_names;

  (* Codegen the body, now that all vars are in scope. *)
  let body_val = codegen_expr body in

  (* Pop all our variables from scope. *)
  List.iter (fun (var_name, old_value) ->
    Hashtbl.add named_values var_name old_value
  ) !old_bindings;

  (* Return the body computation. *)
  body_val

let codegen_proto = function
| Ast.Prototype (name, args) | Ast.BinOpPrototype (name, args, _) ->
  (* Make the function type: double(double,double) etc. *)
  let doubles = Array.make (Array.length args) double_type in
  let ft = function_type double_type doubles in
  let f =
    match lookup_function name the_module with
    | None -> declare_function name ft the_module

    (* If 'f' conflicted, there was already something named 'name'. If it
     * has a body, don't allow redefinition or reextern. *)
    | Some f ->
      (* If 'f' already has a body, reject this. *)
      if block_begin f <> At_end f then
        raise (Error "redefinition of function");

      (* If 'f' took a different number of arguments, reject. *)
      if element_type (type_of f) <> ft then
        raise (Error "redefinition of function with different # args");
      f
  in

  (* Set names for all arguments. *)
  Array.iteri (fun i a ->
    let n = args.(i) in
    set_value_name n a;

```

```

    Hashtbl.add named_values n a;
  ) (params f);
  f

(* Create an alloca for each argument and register the argument in the symbol
 * table so that references to it will succeed. *)
let create_argument_alloca the_function proto =
  let args = match proto with
  | Ast.Prototype (_, args) | Ast.BinOpPrototype (_, args, _) -> args
  in
  Array.iteri (fun i ai ->
    let var_name = args.(i) in
    (* Create an alloca for this variable. *)
    let alloca = create_entry_block_alloca the_function var_name in

    (* Store the initial value into the alloca. *)
    ignore(build_store ai alloca builder);

    (* Add arguments to variable symbol table. *)
    Hashtbl.add named_values var_name alloca;
  ) (params the_function)

let codegen_func the_fpm = function
| Ast.Function (proto, body) ->
  Hashtbl.clear named_values;
  let the_function = codegen_proto proto in

  (* If this is an operator, install it. *)
  begin match proto with
  | Ast.BinOpPrototype (name, args, prec) ->
    let op = name.[String.length name - 1] in
    Hashtbl.add Parser.binop_precedence op prec;
  | _ -> ()
  end;

  (* Create a new basic block to start insertion into. *)
  let bb = append_block context "entry" the_function in
  position_at_end bb builder;

  try
    (* Add all arguments to the symbol table and create their allocas. *)
    create_argument_alloca the_function proto;

    let ret_val = codegen_expr body in

    (* Finish off the function. *)
    let _ = build_ret ret_val builder in

    (* Validate the generated code, checking for consistency. *)
    Llvm_analysis.assert_valid_function the_function;

    (* Optimize the function. *)
    let _ = PassManager.run_function the_function the_fpm in

    the_function
  with e ->
    delete_function the_function;
    raise e

```

toplevel.ml:

```
(*=====*)
* Top-Level parsing and JIT Driver
*=====*)

open Llvml
open Llvml_executionengine

(* top ::= definition | external | expression | ';' *)
let rec main_loop the_fpm the_execution_engine stream =
  match Stream.peek stream with
  | None -> ()

  (* ignore top-level semicolons. *)
  | Some (Token.Kwd ';') ->
    Stream.junk stream;
    main_loop the_fpm the_execution_engine stream

  | Some token ->
    begin
      try match token with
      | Token.Def ->
        let e = Parser.parse_definition stream in
        print_endline "parsed a function definition.";
        dump_value (Codegen.codegen_func the_fpm e);
      | Token.Extern ->
        let e = Parser.parse_extern stream in
        print_endline "parsed an extern.";
        dump_value (Codegen.codegen_proto e);
      | _ ->
        (* Evaluate a top-level expression into an anonymous function. *)
        let e = Parser.parse_toplevel stream in
        print_endline "parsed a top-level expr";
        let the_function = Codegen.codegen_func the_fpm e in
        dump_value the_function;

        (* JIT the function, returning a function pointer. *)
        let result = ExecutionEngine.run_function the_function [||]
          the_execution_engine in

        print_string "Evaluated to ";
        print_float (GenericValue.as_float Codegen.double_type result);
        print_newline ();
        with Stream.Error s | Codegen.Error s ->
          (* Skip token for error recovery. *)
          Stream.junk stream;
          print_endline s;
      end;
      print_string "ready> "; flush stdout;
      main_loop the_fpm the_execution_engine stream
    end
```

toy.ml:

```
(*=====*)
* Main driver code.
*=====*)

open Llvml
```

```

open Llvm_executionengine
open Llvm_target
open Llvm_scalar_opts

let main () =
  ignore (initialize_native_target ());

  (* Install standard binary operators.
   * 1 is the lowest precedence. *)
  Hashtbl.add Parser.binop_precedence '=' 2;
  Hashtbl.add Parser.binop_precedence '<' 10;
  Hashtbl.add Parser.binop_precedence '+' 20;
  Hashtbl.add Parser.binop_precedence '-' 20;
  Hashtbl.add Parser.binop_precedence '*' 40;      (* highest. *)

  (* Prime the first token. *)
  print_string "ready> "; flush stdout;
  let stream = Lexer.lex (Stream.of_channel stdin) in

  (* Create the JIT. *)
  let the_execution_engine = ExecutionEngine.create Codegen.the_module in
  let the_fpm = PassManager.create_function Codegen.the_module in

  (* Set up the optimizer pipeline. Start with registering info about how the
   * target lays out data structures. *)
  DataLayout.add (ExecutionEngine.target_data the_execution_engine) the_fpm;

  (* Promote allocas to registers. *)
  add_memory_to_register_promotion the_fpm;

  (* Do simple "peephole" optimizations and bit-twiddling optzn. *)
  add_instruction_combination the_fpm;

  (* reassociate expressions. *)
  add_reassociation the_fpm;

  (* Eliminate Common SubExpressions. *)
  add_gvn the_fpm;

  (* Simplify the control flow graph (deleting unreachable blocks, etc). *)
  add_cfg_simplification the_fpm;

  ignore (PassManager.initialize the_fpm);

  (* Run the main "interpreter loop" now. *)
  Toplevel.main_loop the_fpm the_execution_engine stream;

  (* Print out all the generated code. *)
  dump_module Codegen.the_module
;;

main ()

```

bindings.c

```

#include <stdio.h>

/* putchar - putchar that takes a double and returns 0. */

```

```
extern double putchar(double X) {
    putchar((char)X);
    return 0;
}

/* printf - printf that takes a double prints it as "%f\n", returning 0. */
extern double printf(double X) {
    printf("%f\n", X);
    return 0;
}
```

Next: Conclusion and other useful LLVM tidbits

Kaleidoscope: Conclusion and other useful LLVM tidbits

- *Tutorial Conclusion*
- *Properties of the LLVM IR*
 - *Target Independence*
 - *Safety Guarantees*
 - *Language-Specific Optimizations*
- *Tips and Tricks*
 - *Implementing portable offsetof/sizeof*
 - *Garbage Collected Stack Frames*

Tutorial Conclusion

Welcome to the final chapter of the “Implementing a language with LLVM” tutorial. In the course of this tutorial, we have grown our little Kaleidoscope language from being a useless toy, to being a semi-interesting (but probably still useless) toy. :)

It is interesting to see how far we’ve come, and how little code it has taken. We built the entire lexer, parser, AST, code generator, and an interactive run-loop (with a JIT!) by-hand in under 700 lines of (non-comment/non-blank) code.

Our little language supports a couple of interesting features: it supports user defined binary and unary operators, it uses JIT compilation for immediate evaluation, and it supports a few control flow constructs with SSA construction.

Part of the idea of this tutorial was to show you how easy and fun it can be to define, build, and play with languages. Building a compiler need not be a scary or mystical process! Now that you’ve seen some of the basics, I strongly encourage you to take the code and hack on it. For example, try adding:

- **global variables** - While global variables have questional value in modern software engineering, they are often useful when putting together quick little hacks like the Kaleidoscope compiler itself. Fortunately, our current setup makes it very easy to add global variables: just have value lookup check to see if an unresolved variable is in the global variable symbol table before rejecting it. To create a new global variable, make an instance of the LLVM `GlobalVariable` class.
- **typed variables** - Kaleidoscope currently only supports variables of type double. This gives the language a very nice elegance, because only supporting one type means that you never have to specify types. Different languages

have different ways of handling this. The easiest way is to require the user to specify types for every variable definition, and record the type of the variable in the symbol table along with its Value*.

- **arrays, structs, vectors, etc** - Once you add types, you can start extending the type system in all sorts of interesting ways. Simple arrays are very easy and are quite useful for many different applications. Adding them is mostly an exercise in learning how the LLVM [getelementptr](#) instruction works: it is so nifty/unconventional, it has its own FAQ! If you add support for recursive types (e.g. linked lists), make sure to read the [section in the LLVM Programmer's Manual](#) that describes how to construct them.
- **standard runtime** - Our current language allows the user to access arbitrary external functions, and we use it for things like “printf” and “putchar”. As you extend the language to add higher-level constructs, often these constructs make the most sense if they are lowered to calls into a language-supplied runtime. For example, if you add hash tables to the language, it would probably make sense to add the routines to a runtime, instead of inlining them all the way.
- **memory management** - Currently we can only access the stack in Kaleidoscope. It would also be useful to be able to allocate heap memory, either with calls to the standard libc malloc/free interface or with a garbage collector. If you would like to use garbage collection, note that LLVM fully supports [Accurate Garbage Collection](#) including algorithms that move objects and need to scan/update the stack.
- **debugger support** - LLVM supports generation of [DWARF Debug info](#) which is understood by common debuggers like GDB. Adding support for debug info is fairly straightforward. The best way to understand it is to compile some C/C++ code with “clang -g -O0” and taking a look at what it produces.
- **exception handling support** - LLVM supports generation of [zero cost exceptions](#) which interoperate with code compiled in other languages. You could also generate code by implicitly making every function return an error value and checking it. You could also make explicit use of setjmp/longjmp. There are many different ways to go here.
- **object orientation, generics, database access, complex numbers, geometric programming, ...** - Really, there is no end of crazy features that you can add to the language.
- **unusual domains** - We've been talking about applying LLVM to a domain that many people are interested in: building a compiler for a specific language. However, there are many other domains that can use compiler technology that are not typically considered. For example, LLVM has been used to implement OpenGL graphics acceleration, translate C++ code to ActionScript, and many other cute and clever things. Maybe you will be the first to JIT compile a regular expression interpreter into native code with LLVM?

Have fun - try doing something crazy and unusual. Building a language like everyone else always has, is much less fun than trying something a little crazy or off the wall and seeing how it turns out. If you get stuck or want to talk about it, feel free to email the [llvm-dev mailing list](#): it has lots of people who are interested in languages and are often willing to help out.

Before we end this tutorial, I want to talk about some “tips and tricks” for generating LLVM IR. These are some of the more subtle things that may not be obvious, but are very useful if you want to take advantage of LLVM's capabilities.

Properties of the LLVM IR

We have a couple common questions about code in the LLVM IR form - lets just get these out of the way right now, shall we?

Target Independence

Kaleidoscope is an example of a “portable language”: any program written in Kaleidoscope will work the same way on any target that it runs on. Many other languages have this property, e.g. lisp, java, haskell, javascript, python, etc (note that while these languages are portable, not all their libraries are).

One nice aspect of LLVM is that it is often capable of preserving target independence in the IR: you can take the LLVM IR for a Kaleidoscope-compiled program and run it on any target that LLVM supports, even emitting C code and compiling that on targets that LLVM doesn't support natively. You can trivially tell that the Kaleidoscope compiler generates target-independent code because it never queries for any target-specific information when generating code.

The fact that LLVM provides a compact, target-independent, representation for code gets a lot of people excited. Unfortunately, these people are usually thinking about C or a language from the C family when they are asking questions about language portability. I say "unfortunately", because there is really no way to make (fully general) C code portable, other than shipping the source code around (and of course, C source code is not actually portable in general either - ever port a really old application from 32- to 64-bits?).

The problem with C (again, in its full generality) is that it is heavily laden with target specific assumptions. As one simple example, the preprocessor often destructively removes target-independence from the code when it processes the input text:

```
#ifdef __i386__
    int X = 1;
#else
    int X = 42;
#endif
```

While it is possible to engineer more and more complex solutions to problems like this, it cannot be solved in full generality in a way that is better than shipping the actual source code.

That said, there are interesting subsets of C that can be made portable. If you are willing to fix primitive types to a fixed size (say `int` = 32-bits, and `long` = 64-bits), don't care about ABI compatibility with existing binaries, and are willing to give up some other minor features, you can have portable code. This can make sense for specialized domains such as an in-kernel language.

Safety Guarantees

Many of the languages above are also "safe" languages: it is impossible for a program written in Java to corrupt its address space and crash the process (assuming the JVM has no bugs). Safety is an interesting property that requires a combination of language design, runtime support, and often operating system support.

It is certainly possible to implement a safe language in LLVM, but LLVM IR does not itself guarantee safety. The LLVM IR allows unsafe pointer casts, use after free bugs, buffer over-runs, and a variety of other problems. Safety needs to be implemented as a layer on top of LLVM and, conveniently, several groups have investigated this. Ask on the [llvm-dev mailing list](#) if you are interested in more details.

Language-Specific Optimizations

One thing about LLVM that turns off many people is that it does not solve all the world's problems in one system (sorry 'world hunger', someone else will have to solve you some other day). One specific complaint is that people perceive LLVM as being incapable of performing high-level language-specific optimization: LLVM "loses too much information".

Unfortunately, this is really not the place to give you a full and unified version of "Chris Lattner's theory of compiler design". Instead, I'll make a few observations:

First, you're right that LLVM does lose information. For example, as of this writing, there is no way to distinguish in the LLVM IR whether an SSA-value came from a C "int" or a C "long" on an ILP32 machine (other than debug info). Both get compiled down to an 'i32' value and the information about what it came from is lost. The more general issue here, is that the LLVM type system uses "structural equivalence" instead of "name equivalence". Another place this surprises people is if you have two types in a high-level language that have the same structure (e.g. two different

structs that have a single int field): these types will compile down into a single LLVM type and it will be impossible to tell what it came from.

Second, while LLVM does lose information, LLVM is not a fixed target: we continue to enhance and improve it in many different ways. In addition to adding new features (LLVM did not always support exceptions or debug info), we also extend the IR to capture important information for optimization (e.g. whether an argument is sign or zero extended, information about pointers aliasing, etc). Many of the enhancements are user-driven: people want LLVM to include some specific feature, so they go ahead and extend it.

Third, it is *possible and easy* to add language-specific optimizations, and you have a number of choices in how to do it. As one trivial example, it is easy to add language-specific optimization passes that “know” things about code compiled for a language. In the case of the C family, there is an optimization pass that “knows” about the standard C library functions. If you call “exit(0)” in main(), it knows that it is safe to optimize that into “return 0;” because C specifies what the ‘exit’ function does.

In addition to simple library knowledge, it is possible to embed a variety of other language-specific information into the LLVM IR. If you have a specific need and run into a wall, please bring the topic up on the llvm-dev list. At the very worst, you can always treat LLVM as if it were a “dumb code generator” and implement the high-level optimizations you desire in your front-end, on the language-specific AST.

Tips and Tricks

There is a variety of useful tips and tricks that you come to know after working on/with LLVM that aren’t obvious at first glance. Instead of letting everyone rediscover them, this section talks about some of these issues.

Implementing portable offsetof/sizeof

One interesting thing that comes up, if you are trying to keep the code generated by your compiler “target independent”, is that you often need to know the size of some LLVM type or the offset of some field in an llvm structure. For example, you might need to pass the size of a type into a function that allocates memory.

Unfortunately, this can vary widely across targets: for example the width of a pointer is trivially target-specific. However, there is a [clever way to use the getelementptr instruction](#) that allows you to compute this in a portable way.

Garbage Collected Stack Frames

Some languages want to explicitly manage their stack frames, often so that they are garbage collected or to allow easy implementation of closures. There are often better ways to implement these features than explicit stack frames, but [LLVM does support them](#), if you want. It requires your front-end to convert the code into [Continuation Passing Style](#) and the use of tail calls (which LLVM also supports).

2.17.3 Building a JIT in LLVM

Building a JIT: Starting out with KaleidoscopeJIT

- [Chapter 1 Introduction](#)
- [JIT API Basics](#)
- [KaleidoscopeJIT](#)

• *Full Code Listing*

Chapter 1 Introduction

Welcome to Chapter 1 of the “Building an ORC-based JIT in LLVM” tutorial. This tutorial runs through the implementation of a JIT compiler using LLVM’s On-Request-Compilation (ORC) APIs. It begins with a simplified version of the KaleidoscopeJIT class used in the [Implementing a language with LLVM](#) tutorials and then introduces new features like optimization, lazy compilation and remote execution.

The goal of this tutorial is to introduce you to LLVM’s ORC JIT APIs, show how these APIs interact with other parts of LLVM, and to teach you how to recombine them to build a custom JIT that is suited to your use-case.

The structure of the tutorial is:

- **Chapter #1:** Investigate the simple KaleidoscopeJIT class. This will introduce some of the basic concepts of the ORC JIT APIs, including the idea of an ORC *Layer*.
- **Chapter #2:** Extend the basic KaleidoscopeJIT by adding a new layer that will optimize IR and generated code.
- **Chapter #3:** Further extend the JIT by adding a Compile-On-Demand layer to lazily compile IR.
- **Chapter #4:** Improve the laziness of our JIT by replacing the Compile-On-Demand layer with a custom layer that uses the ORC Compile Callbacks API directly to defer IR-generation until functions are called.
- **Chapter #5:** Add process isolation by JITing code into a remote process with reduced privileges using the JIT Remote APIs.

To provide input for our JIT we will use the Kaleidoscope REPL from [Chapter 7](#) of the “Implementing a language in LLVM tutorial”, with one minor modification: We will remove the FunctionPassManager from the code for that chapter and replace it with optimization support in our JIT class in Chapter #2.

Finally, a word on API generations: ORC is the 3rd generation of LLVM JIT API. It was preceded by MCJIT, and before that by the (now deleted) legacy JIT. These tutorials don’t assume any experience with these earlier APIs, but readers acquainted with them will see many familiar elements. Where appropriate we will make this connection with the earlier APIs explicit to help people who are transitioning from them to ORC.

JIT API Basics

The purpose of a JIT compiler is to compile code “on-the-fly” as it is needed, rather than compiling whole programs to disk ahead of time as a traditional compiler does. To support that aim our initial, bare-bones JIT API will be:

1. `Handle addModule(Module &M)` – Make the given IR module available for execution.
2. `JITSymbol findSymbol(const std::string &Name)` – Search for pointers to symbols (functions or variables) that have been added to the JIT.
3. `void removeModule(Handle H)` – Remove a module from the JIT, releasing any memory that had been used for the compiled code.

A basic use-case for this API, executing the ‘main’ function from a module, will look like:

```
std::unique_ptr<Module> M = buildModule();
JIT J;
Handle H = J.addModule(*M);
int (*Main)(int, char*[]) =
    (int (*)(int, char*[]))J.findSymbol("main").getAddress();
int Result = Main();
J.removeModule(H);
```

The APIs that we build in these tutorials will all be variations on this simple theme. Behind the API we will refine the implementation of the JIT to add support for optimization and lazy compilation. Eventually we will extend the API itself to allow higher-level program representations (e.g. ASTs) to be added to the JIT.

KaleidoscopeJIT

In the previous section we described our API, now we examine a simple implementation of it: The KaleidoscopeJIT class¹ that was used in the [Implementing a language with LLVM](#) tutorials. We will use the REPL code from [Chapter 7](#) of that tutorial to supply the input for our JIT: Each time the user enters an expression the REPL will add a new IR module containing the code for that expression to the JIT. If the expression is a top-level expression like '1+1' or 'sin(x)', the REPL will also use the findSymbol method of our JIT class find and execute the code for the expression, and then use the removeModule method to remove the code again (since there's no way to re-invoke an anonymous expression). In later chapters of this tutorial we'll modify the REPL to enable new interactions with our JIT class, but for now we will take this setup for granted and focus our attention on the implementation of our JIT itself.

Our KaleidoscopeJIT class is defined in the KaleidoscopeJIT.h header. After the usual include guards and #includes², we get to the definition of our class:

```
#ifndef LLVM_EXECUTIONENGINE_ORC_KALEIDOSCOPEJIT_H
#define LLVM_EXECUTIONENGINE_ORC_KALEIDOSCOPEJIT_H

#include "llvm/ExecutionEngine/ExecutionEngine.h"
#include "llvm/ExecutionEngine/RTDyldMemoryManager.h"
#include "llvm/ExecutionEngine/Orc/CompileUtils.h"
#include "llvm/ExecutionEngine/Orc/IRCompileLayer.h"
#include "llvm/ExecutionEngine/Orc/LambdaResolver.h"
#include "llvm/ExecutionEngine/Orc/ObjectLinkingLayer.h"
#include "llvm/IR/Mangler.h"
#include "llvm/Support/DynamicLibrary.h"

namespace llvm {
namespace orc {

class KaleidoscopeJIT {
private:
    std::unique_ptr<TargetMachine> TM;
    const DataLayout DL;
    ObjectLinkingLayer<> ObjectLayer;
    IRCompileLayer<decltype(ObjectLayer)> CompileLayer;

public:
    typedef decltype(CompileLayer)::ModuleSetHandleT ModuleHandleT;
```

¹ Actually we use a cut-down version of KaleidoscopeJIT that makes a simplifying assumption: symbols cannot be re-defined. This will make it impossible to re-define symbols in the REPL, but will make our symbol lookup logic simpler. Re-introducing support for symbol redefinition is left as an exercise for the reader. (The KaleidoscopeJIT.h used in the original tutorials will be a helpful reference).

²

File	Reason for inclusion
ExecutionEngine.h	Access to the EngineBuilder::selectTarget method.
RTDyldMemoryManager.h	Access to the RTDyldMemoryManager::getSymbolAddressInProcess method.
CompileUtils.h	Provides the SimpleCompiler class.
IRCompileLayer.h	Provides the IRCompileLayer class.
LambdaResolver.h	Access the createLambdaResolver function, which provides easy construction of symbol resolvers.
ObjectLinkingLayer.h	Provides the ObjectLinkingLayer class.
Mangler.h	Provides the Mangler class for platform specific name-mangling.
DynamicLibrary.h	Provides the DynamicLibrary class, which makes symbols in the host process searchable.

Our class begins with four members: A `TargetMachine`, `TM`, which will be used to build our LLVM compiler instance; A `DataLayout`, `DL`, which will be used for symbol mangling (more on that later), and two ORC *layers*: an `ObjectLinkingLayer` and an `IRCompileLayer`. We'll be talking more about layers in the next chapter, but for now you can think of them as analogous to LLVM Passes: they wrap up useful JIT utilities behind an easy to compose interface. The first layer, `ObjectLinkingLayer`, is the foundation of our JIT: it takes in-memory object files produced by a compiler and links them on the fly to make them executable. This JIT-on-top-of-a-linker design was introduced in MCJIT, however the linker was hidden inside the MCJIT class. In ORC we expose the linker so that clients can access and configure it directly if they need to. In this tutorial our `ObjectLinkingLayer` will just be used to support the next layer in our stack: the `IRCompileLayer`, which will be responsible for taking LLVM IR, compiling it, and passing the resulting in-memory object files down to the object linking layer below.

That's it for member variables, after that we have a single typedef: `ModuleHandleT`. This is the handle type that will be returned from our JIT's `addModule` method, and can be passed to the `removeModule` method to remove a module. The `IRCompileLayer` class already provides a convenient handle type (`IRCompileLayer::ModuleSetHandleT`), so we just alias our `ModuleHandleT` to this.

```
KaleidoscopeJIT()
    : TM(EngineBuilder().selectTarget()), DL(TM->createDataLayout()),
      CompileLayer(ObjectLayer, SimpleCompiler(*TM)) {
    llvm::sys::DynamicLibrary::LoadLibraryPermanently(nullptr);
}

TargetMachine &getTargetMachine() { return *TM; }
```

Next up we have our class constructor. We begin by initializing `TM` using the `EngineBuilder::selectTarget` helper method, which constructs a `TargetMachine` for the current process. Next we use our newly created `TargetMachine` to initialize `DL`, our `DataLayout`. Then we initialize our `IRCompileLayer`. Our `IRCompile` layer needs two things: (1) A reference to our object linking layer, and (2) a compiler instance to use to perform the actual compilation from IR to object files. We use the off-the-shelf `SimpleCompiler` instance for now. Finally, in the body of the constructor, we call the `DynamicLibrary::LoadLibraryPermanently` method with a `nullptr` argument. Normally the `LoadLibraryPermanently` method is called with the path of a dynamic library to load, but when passed a null pointer it will 'load' the host process itself, making its exported symbols available for execution.

```
ModuleHandle addModule(std::unique_ptr<Module> M) {
    // Build our symbol resolver:
    // Lambda 1: Look back into the JIT itself to find symbols that are part of
    //           the same "logical dylib".
    // Lambda 2: Search for external symbols in the host process.
    auto Resolver = createLambdaResolver(
        [&](const std::string &Name) {
            if (auto Sym = CompileLayer.findSymbol(Name, false))
                return Sym;
            return JITSymbol(nullptr);
        },
        [] (const std::string &S) {
            if (auto SymAddr =
                RTDyldMemoryManager::getSymbolAddressInProcess(Name))
                return JITSymbol(SymAddr, JITSymbolFlags::Exported);
            return JITSymbol(nullptr);
        });

    // Build a singleton module set to hold our module.
    std::vector<std::unique_ptr<Module>> Ms;
    Ms.push_back(std::move(M));

    // Add the set to the JIT with the resolver we created above and a newly
    // created SectionMemoryManager.
```

```

return CompileLayer.addModuleSet(std::move(Ms),
                                make_unique<SectionMemoryManager>(),
                                std::move(Resolver));
}

```

Now we come to the first of our JIT API methods: `addModule`. This method is responsible for adding IR to the JIT and making it available for execution. In this initial implementation of our JIT we will make our modules “available for execution” by adding them straight to the `IRCompileLayer`, which will immediately compile them. In later chapters we will teach our JIT to be lazier and instead add the Modules to a “pending” list to be compiled if and when they are first executed.

To add our module to the `IRCompileLayer` we need to supply two auxiliary objects (as well as the module itself): a memory manager and a symbol resolver. The memory manager will be responsible for managing the memory allocated to JIT’d machine code, setting memory permissions, and registering exception handling tables (if the JIT’d code uses exceptions). For our memory manager we will use the `SectionMemoryManager` class: another off-the-shelf utility that provides all the basic functionality we need. The second auxiliary class, the symbol resolver, is more interesting for us. It exists to tell the JIT where to look when it encounters an *external symbol* in the module we are adding. External symbols are any symbol not defined within the module itself, including calls to functions outside the JIT and calls to functions defined in other modules that have already been added to the JIT. It may seem as though modules added to the JIT should “know about one another” by default, but since we would still have to supply a symbol resolver for references to code outside the JIT it turns out to be easier to just re-use this one mechanism for all symbol resolution. This has the added benefit that the user has full control over the symbol resolution process. Should we search for definitions within the JIT first, then fall back on external definitions? Or should we prefer external definitions where available and only JIT code if we don’t already have an available implementation? By using a single symbol resolution scheme we are free to choose whatever makes the most sense for any given use case.

Building a symbol resolver is made especially easy by the `createLambdaResolver` function. This function takes two lambdas³ and returns a `JITSymbolResolver` instance. The first lambda is used as the implementation of the resolver’s `findSymbolInLogicalDylib` method, which searches for symbol definitions that should be thought of as being part of the same “logical” dynamic library as this Module. If you are familiar with static linking: this means that `findSymbolInLogicalDylib` should expose symbols with common linkage and hidden visibility. If all this sounds foreign you can ignore the details and just remember that this is the first method that the linker will use to try to find a symbol definition. If the `findSymbolInLogicalDylib` method returns a null result then the linker will call the second symbol resolver method, called `findSymbol`, which searches for symbols that should be thought of as external to (but visible from) the module and its logical dylib. In this tutorial we will adopt the following simple scheme: All modules added to the JIT will behave as if they were linked into a single, ever-growing logical dylib. To implement this our first lambda (the one defining `findSymbolInLogicalDylib`) will just search for JIT’d code by calling the `CompileLayer`’s `findSymbol` method. If we don’t find a symbol in the JIT itself we’ll fall back to our second lambda, which implements `findSymbol`. This will use the `RTDyldMemoryManager::getSymbolAddressInProcess` method to search for the symbol within the program itself. If we can’t find a symbol definition via either of these paths, the JIT will refuse to accept our module, returning a “symbol not found” error.

Now that we’ve built our symbol resolver, we’re ready to add our module to the JIT. We do this by calling the `CompileLayer`’s `addModuleSet` method⁴. Since we only have a single Module and `addModuleSet` expects a collection, we will create a vector of modules and add our module as the only member. Since we have already typedef’d our `ModuleHandleT` type to be the same as the `CompileLayer`’s handle type, we can return the handle from `addModuleSet` directly from our `addModule` method.

```

JITSymbol findSymbol(const std::string Name) {
    std::string MangledName;
    raw_string_ostream MangledNameStream(MangledName);
    Mangler::getNameWithPrefix(MangledNameStream, Name, DL);
}

```

³ Actually they don’t have to be lambdas, any object with a call operator will do, including plain old functions or `std::functions`.

⁴ ORC layers accept sets of Modules, rather than individual ones, so that all Modules in the set could be co-located by the memory manager, though this feature is not yet implemented.

```

    return CompileLayer.findSymbol(MangledNameStream.str(), true);
}

void removeModule(ModuleHandle H) {
    CompileLayer.removeModuleSet(H);
}

```

Now that we can add code to our JIT, we need a way to find the symbols we've added to it. To do that we call the `findSymbol` method on our `IRCompileLayer`, but with a twist: We have to *mangle* the name of the symbol we're searching for first. The reason for this is that the ORC JIT components use mangled symbols internally the same way a static compiler and linker would, rather than using plain IR symbol names. The kind of mangling will depend on the `DataLayout`, which in turn depends on the target platform. To allow us to remain portable and search based on the un-mangled name, we just re-produce this mangling ourselves.

We now come to the last method in our JIT API: `removeModule`. This method is responsible for destructing the `MemoryManager` and `SymbolResolver` that were added with a given module, freeing any resources they were using in the process. In our Kaleidoscope demo we rely on this method to remove the module representing the most recent top-level expression, preventing it from being treated as a duplicate definition when the next top-level expression is entered. It is generally good to free any module that you know you won't need to call further, just to free up the resources dedicated to it. However, you don't strictly need to do this: All resources will be cleaned up when your JIT class is destructed, if they haven't been freed before then.

This brings us to the end of Chapter 1 of Building a JIT. You now have a basic but fully functioning JIT stack that you can use to take LLVM IR and make it executable within the context of your JIT process. In the next chapter we'll look at how to extend this JIT to produce better quality code, and in the process take a deeper look at the ORC layer concept.

Next: [Extending the KaleidoscopeJIT](#)

Full Code Listing

Here is the complete code listing for our running example. To build this example, use:

```

# Compile
clang++ -g toy.cpp `llvm-config --cxxflags --ldflags --system-libs --libs core orc_
↪native` -O3 -o toy
# Run
./toy

```

Here is the code:

```

//===- KaleidoscopeJIT.h - A simple JIT for Kaleidoscope -----*- C++ -*-===//
//
//                               The LLVM Compiler Infrastructure
//
// This file is distributed under the University of Illinois Open Source
// License. See LICENSE.TXT for details.
//
//===-----//
//
// Contains a simple JIT definition for use in the kaleidoscope tutorials.
//
//===-----//

#ifndef LLVM_EXECUTIONENGINE_ORC_KALEIDOSCOPEJIT_H
#define LLVM_EXECUTIONENGINE_ORC_KALEIDOSCOPEJIT_H

```

```

#include "llvm/ADT/STLExtras.h"
#include "llvm/ExecutionEngine/ExecutionEngine.h"
#include "llvm/ExecutionEngine/JITSymbol.h"
#include "llvm/ExecutionEngine/RTDyldMemoryManager.h"
#include "llvm/ExecutionEngine/SectionMemoryManager.h"
#include "llvm/ExecutionEngine/Orc/CompileUtils.h"
#include "llvm/ExecutionEngine/Orc/IRCompileLayer.h"
#include "llvm/ExecutionEngine/Orc/LambdaResolver.h"
#include "llvm/ExecutionEngine/Orc/RTDyldObjectLinkingLayer.h"
#include "llvm/IR/DataLayout.h"
#include "llvm/IR/Mangler.h"
#include "llvm/Support/DynamicLibrary.h"
#include "llvm/Support/raw_ostream.h"
#include "llvm/Target/TargetMachine.h"
#include <algorithm>
#include <memory>
#include <string>
#include <vector>

namespace llvm {
namespace orc {

class KaleidoscopeJIT {
private:
    std::unique_ptr<TargetMachine> TM;
    const DataLayout DL;
    RTDyldObjectLinkingLayer ObjectLayer;
    IRCompileLayer<decltype(ObjectLayer), SimpleCompiler> CompileLayer;

public:
    using ModuleHandle = decltype(CompileLayer)::ModuleHandleT;

    KaleidoscopeJIT()
        : TM(EngineBuilder().selectTarget()), DL(TM->createDataLayout()),
          ObjectLayer([]() { return std::make_shared<SectionMemoryManager>(); }),
          CompileLayer(ObjectLayer, SimpleCompiler(*TM)) {
        llvm::sys::DynamicLibrary::LoadLibraryPermanently(nullptr);
    }

    TargetMachine &getTargetMachine() { return *TM; }

    ModuleHandle addModule(std::unique_ptr<Module> M) {
        // Build our symbol resolver:
        // Lambda 1: Look back into the JIT itself to find symbols that are part of
        //           the same "logical dylib".
        // Lambda 2: Search for external symbols in the host process.
        auto Resolver = createLambdaResolver(
            [&](const std::string &Name) {
                if (auto Sym = CompileLayer.findSymbol(Name, false))
                    return Sym;
                return JITSymbol(nullptr);
            },
            [](const std::string &Name) {
                if (auto SymAddr =
                    RTDyldMemoryManager::getSymbolAddressInProcess(Name))
                    return JITSymbol(SymAddr, JITSymbolFlags::Exported);
                return JITSymbol(nullptr);
            });
    }
};

```



```
    });

    // Add the set to the JIT with the resolver we created above and a newly
    // created SectionMemoryManager.
    return cantFail(CompileLayer.addModule(std::move(M),
                                           std::move(Resolver)));
}

JITSymbol findSymbol(const std::string Name) {
    std::string MangledName;
    raw_string_ostream MangledNameStream(MangledName);
    Mangler::getNameWithPrefix(MangledNameStream, Name, DL);
    return CompileLayer.findSymbol(MangledNameStream.str(), true);
}

void removeModule(ModuleHandle H) {
    cantFail(CompileLayer.removeModule(H));
}
};

} // end namespace orc
} // end namespace llvm

#endif // LLVM_EXECUTIONENGINE_ORC_KALEIDOSCOPEJIT_H
```

Building a JIT: Adding Optimizations – An introduction to ORC Layers

- *Chapter 2 Introduction*
- *Optimizing Modules using the IRTransformLayer*
- *Full Code Listing*

This tutorial is under active development. It is incomplete and details may change frequently. Nonetheless we invite you to try it out as it stands, and we welcome any feedback.

Chapter 2 Introduction

Welcome to Chapter 2 of the “Building an ORC-based JIT in LLVM” tutorial. In [Chapter 1](#) of this series we examined a basic JIT class, KaleidoscopeJIT, that could take LLVM IR modules as input and produce executable code in memory. KaleidoscopeJIT was able to do this with relatively little code by composing two off-the-shelf *ORC layers*: IRCompileLayer and ObjectLinkingLayer, to do much of the heavy lifting.

In this layer we’ll learn more about the ORC layer concept by using a new layer, IRTransformLayer, to add IR optimization support to KaleidoscopeJIT.

Optimizing Modules using the IRTransformLayer

In [Chapter 4](#) of the “Implementing a language with LLVM” tutorial series the `llvm::FunctionPassManager` is introduced as a means for optimizing LLVM IR. Interested readers may read that chapter for details, but in short: to optimize a Module we create an `llvm::FunctionPassManager` instance, configure it with a set of optimizations, then run the PassManager on a Module to mutate it into a (hopefully) more optimized but semantically equivalent form. In the

original tutorial series the FunctionPassManager was created outside the KaleidoscopeJIT and modules were optimized before being added to it. In this Chapter we will make optimization a phase of our JIT instead. For now this will provide us a motivation to learn more about ORC layers, but in the long term making optimization part of our JIT will yield an important benefit: When we begin lazily compiling code (i.e. deferring compilation of each function until the first time it's run), having optimization managed by our JIT will allow us to optimize lazily too, rather than having to do all our optimization up-front.

To add optimization support to our JIT we will take the KaleidoscopeJIT from Chapter 1 and compose an ORC *IRTransformLayer* on top. We will look at how the IRTransformLayer works in more detail below, but the interface is simple: the constructor for this layer takes a reference to the layer below (as all layers do) plus an *IR optimization function* that it will apply to each Module that is added via addModuleSet:

```
class KaleidoscopeJIT {
private:
    std::unique_ptr<TargetMachine> TM;
    const DataLayout DL;
    ObjectLinkingLayer<> ObjectLayer;
    IRCompileLayer<decltype(ObjectLayer)> CompileLayer;

    typedef std::function<std::unique_ptr<Module>(std::unique_ptr<Module>)>
        OptimizeFunction;

    IRTransformLayer<decltype(CompileLayer), OptimizeFunction> OptimizeLayer;

public:
    typedef decltype(OptimizeLayer)::ModuleSetHandleT ModuleHandle;

    KaleidoscopeJIT()
        : TM(EngineBuilder().selectTarget()), DL(TM->createDataLayout()),
          CompileLayer(ObjectLayer, SimpleCompiler(*TM)),
          OptimizeLayer(CompileLayer,
                        [this](std::unique_ptr<Module> M) {
                            return optimizeModule(std::move(M));
                        }) {
        llvm::sys::DynamicLibrary::LoadLibraryPermanently(nullptr);
    }
}
```

Our extended KaleidoscopeJIT class starts out the same as it did in Chapter 1, but after the CompileLayer we introduce a typedef for our optimization function. In this case we use a std::function (a handy wrapper for “function-like” things) from a single unique_ptr<Module> input to a std::unique_ptr<Module> output. With our optimization function typedef in place we can declare our OptimizeLayer, which sits on top of our CompileLayer.

To initialize our OptimizeLayer we pass it a reference to the CompileLayer below (standard practice for layers), and we initialize the OptimizeFunction using a lambda that calls out to an “optimizeModule” function that we will define below.

```
// ...
auto Resolver = createLambdaResolver(
    [&](const std::string &Name) {
        if (auto Sym = OptimizeLayer.findSymbol(Name, false))
            return Sym;
        return JITSymbol(nullptr);
    },
    // ...
```

```
// ...
return OptimizeLayer.addModuleSet(std::move(Ms),
                                   make_unique<SectionMemoryManager>(),
```

```

std::move(Resolver));
// ...

// ...
return OptimizeLayer.findSymbol(MangledNameStream.str(), true);
// ...

// ...
OptimizeLayer.removeModuleSet(H);
// ...

```

Next we need to replace references to ‘CompileLayer’ with references to OptimizeLayer in our key methods: addModule, findSymbol, and removeModule. In addModule we need to be careful to replace both references: the findSymbol call inside our resolver, and the call through to addModuleSet.

```

std::unique_ptr<Module> optimizeModule(std::unique_ptr<Module> M) {
    // Create a function pass manager.
    auto FPM = llvm::make_unique<legacy::FunctionPassManager>(M.get());

    // Add some optimizations.
    FPM->add(createInstructionCombiningPass());
    FPM->add(createReassociatePass());
    FPM->add(createGVNPass());
    FPM->add(createCFGSimplificationPass());
    FPM->doInitialization();

    // Run the optimizations over all functions in the module being added to
    // the JIT.
    for (auto &F : *M)
        FPM->run(F);

    return M;
}

```

At the bottom of our JIT we add a private method to do the actual optimization: *optimizeModule*. This function sets up a FunctionPassManager, adds some passes to it, runs it over every function in the module, and then returns the mutated module. The specific optimizations are the same ones used in [Chapter 4](#) of the “Implementing a language with LLVM” tutorial series. Readers may visit that chapter for a more in-depth discussion of these, and of IR optimization in general.

And that’s it in terms of changes to KaleidoscopeJIT: When a module is added via addModule the OptimizeLayer will call our optimizeModule function before passing the transformed module on to the CompileLayer below. Of course, we could have called optimizeModule directly in our addModule function and not gone to the bother of using the IRTransformLayer, but doing so gives us another opportunity to see how layers compose. It also provides a neat entry point to the *layer* concept itself, because IRTransformLayer turns out to be one of the simplest implementations of the layer concept that can be devised:

```

template <typename BaseLayerT, typename TransformFtor>
class IRTransformLayer {
public:
    typedef typename BaseLayerT::ModuleSetHandleT ModuleSetHandleT;

    IRTransformLayer(BaseLayerT &BaseLayer,
                    TransformFtor Transform = TransformFtor())
        : BaseLayer(BaseLayer), Transform(std::move(Transform)) {}

    template <typename ModuleSetT, typename MemoryManagerPtrT,

```

```

    typename SymbolResolverPtrT>
ModuleSetHandleT addModuleSet (ModuleSetT Ms,
                               MemoryManagerPtrT MemMgr,
                               SymbolResolverPtrT Resolver) {

    for (auto I = Ms.begin(), E = Ms.end(); I != E; ++I)
        *I = Transform(std::move(*I));

    return BaseLayer.addModuleSet(std::move(Ms), std::move(MemMgr),
                                   std::move(Resolver));
}

void removeModuleSet (ModuleSetHandleT H) { BaseLayer.removeModuleSet (H); }

JITSymbol findSymbol(const std::string &Name, bool ExportedSymbolsOnly) {
    return BaseLayer.findSymbol (Name, ExportedSymbolsOnly);
}

JITSymbol findSymbolIn (ModuleSetHandleT H, const std::string &Name,
                        bool ExportedSymbolsOnly) {
    return BaseLayer.findSymbolIn (H, Name, ExportedSymbolsOnly);
}

void emitAndFinalize (ModuleSetHandleT H) {
    BaseLayer.emitAndFinalize (H);
}

TransformFtor& getTransform() { return Transform; }

const TransformFtor& getTransform() const { return Transform; }

private:
    BaseLayerT &BaseLayer;
    TransformFtor Transform;
};

```

This is the whole definition of `IRTransformLayer`, from `llvm/include/llvm/ExecutionEngine/Orc/IRTransformLayer.h`, stripped of its comments. It is a template class with two template arguments: `BaseLayerT` and `TransformFtor` that provide the type of the base layer and the type of the “transform functor” (in our case a `std::function`) respectively. This class is concerned with two very simple jobs: (1) Running every IR Module that is added with `addModuleSet` through the transform functor, and (2) conforming to the ORC layer interface. The interface consists of one typedef and five methods:

Interface	Description
ModuleSetHandleT	Provides a handle that can be used to identify a module set when calling <code>findSymbolIn</code> , <code>removeModuleSet</code> , or <code>emitAndFinalize</code> .
addModuleSet	Takes a given set of Modules and makes them “available for execution. This means that symbols in those modules should be searchable via <code>findSymbol</code> and <code>findSymbolIn</code> , and the address of the symbols should be read/writable (for data symbols), or executable (for function symbols) after <code>JITSymbol::getAddress()</code> is called. Note: This means that <code>addModuleSet</code> doesn’t have to compile (or do any other work) up-front. It <i>can</i> , like <code>IRCompileLayer</code> , act eagerly, but it can also simply record the module and take no further action until somebody calls <code>JITSymbol::getAddress()</code> . In <code>IRTransformLayer</code> ’s case <code>addModuleSet</code> eagerly applies the transform functor to each module in the set, then passes the resulting set of mutated modules down to the layer below.
removeModuleSet	Removes a set of modules from the JIT. Code or data defined in these modules will no longer be available, and the memory holding the JIT’d definitions will be freed.
findSymbol	Searches for the named symbol in all modules that have previously been added via <code>addModuleSet</code> (and not yet removed by a call to <code>removeModuleSet</code>). In <code>IRTransformLayer</code> we just pass the query on to the layer below. In our REPL this is our default way to search for function definitions.
findSymbolIn	Searches for the named symbol in the module set indicated by the given <code>ModuleSetHandleT</code> . This is just an optimized search, better for lookup-speed when you know exactly a symbol definition should be found. In <code>IRTransformLayer</code> we just pass this query on to the layer below. In our REPL we use this method to search for functions representing top-level expressions, since we know exactly where we’ll find them: in the top-level expression module we just added.
emitAndFinalize	Forces all of the actions required to make the code and data in a module set (represented by a <code>ModuleSetHandleT</code>) accessible. Behaves as if some symbol in the set had been searched for and <code>JITSymbol::getSymbolAddress</code> called. This is rarely needed, but can be useful when dealing with layers that usually behave lazily if the user wants to trigger early compilation (for example, to use idle CPU time to eagerly compile code in the background).

This interface attempts to capture the natural operations of a JIT (with some wrinkles like `emitAndFinalize` for performance), similar to the basic JIT API operations we identified in Chapter 1. Conforming to the layer concept allows classes to compose neatly by implementing their behaviors in terms of the these same operations, carried out on the layer below. For example, an eager layer (like `IRTransformLayer`) can implement `addModuleSet` by running each module in the set through its transform up-front and immediately passing the result to the layer below. A lazy layer, by contrast, could implement `addModuleSet` by squirreling away the modules doing no other up-front work, but applying the transform (and calling `addModuleSet` on the layer below) when the client calls `findSymbol` instead. The JIT’d program behavior will be the same either way, but these choices will have different performance characteristics: Doing work eagerly means the JIT takes longer up-front, but proceeds smoothly once this is done. Deferring work allows the JIT to get up-and-running quickly, but will force the JIT to pause and wait whenever some code or data is needed that hasn’t already been processed.

Our current REPL is eager: Each function definition is optimized and compiled as soon as it’s typed in. If we were to make the transform layer lazy (but not change things otherwise) we could defer optimization until the first time we reference a function in a top-level expression (see if you can figure out why, then check out the answer below¹). In the next chapter, however we’ll introduce fully lazy compilation, in which function’s aren’t compiled until they’re first called at run-time. At this point the trade-offs get much more interesting: the lazier we are, the quicker we can start

¹ When we add our top-level expression to the JIT, any calls to functions that we defined earlier will appear to the `ObjectLinkingLayer` as external symbols. The `ObjectLinkingLayer` will call the `SymbolResolver` that we defined in `addModuleSet`, which in turn calls `findSymbol` on the `OptimizeLayer`, at which point even a lazy transform layer will have to do its work.

executing the first function, but the more often we'll have to pause to compile newly encountered functions. If we only code-gen lazily, but optimize eagerly, we'll have a slow startup (which everything is optimized) but relatively short pauses as each function just passes through code-gen. If we both optimize and code-gen lazily we can start executing the first function more quickly, but we'll have longer pauses as each function has to be both optimized and code-gen'd when it's first executed. Things become even more interesting if we consider interprocedural optimizations like inlining, which must be performed eagerly. These are complex trade-offs, and there is no one-size-fits all solution to them, but by providing composable layers we leave the decisions to the person implementing the JIT, and make it easy for them to experiment with different configurations.

Next: Adding Per-function Lazy Compilation

Full Code Listing

Here is the complete code listing for our running example with an IRTransformLayer added to enable optimization. To build this example, use:

```
# Compile
clang++ -g toy.cpp `llvm-config --cxxflags --ldflags --system-libs --libs core orc_
↪native` -O3 -o toy
# Run
./toy
```

Here is the code:

```
//====- KaleidoscopeJIT.h - A simple JIT for Kaleidoscope -----*- C++ -*-====//
//
//                               The LLVM Compiler Infrastructure
//
// This file is distributed under the University of Illinois Open Source
// License. See LICENSE.TXT for details.
//
//=====//
//
// Contains a simple JIT definition for use in the kaleidoscope tutorials.
//
//=====//

#ifndef LLVM_EXECUTIONENGINE_ORC_KALEIDOSCOPEJIT_H
#define LLVM_EXECUTIONENGINE_ORC_KALEIDOSCOPEJIT_H

#include "llvm/ADT/STLExtras.h"
#include "llvm/ExecutionEngine/ExecutionEngine.h"
#include "llvm/ExecutionEngine/JITSymbol.h"
#include "llvm/ExecutionEngine/RTDyldMemoryManager.h"
#include "llvm/ExecutionEngine/SectionMemoryManager.h"
#include "llvm/ExecutionEngine/Orc/CompileUtils.h"
#include "llvm/ExecutionEngine/Orc/IRCompileLayer.h"
#include "llvm/ExecutionEngine/Orc/IRTransformLayer.h"
#include "llvm/ExecutionEngine/Orc/LambdaResolver.h"
#include "llvm/ExecutionEngine/Orc/RTDyldObjectLinkingLayer.h"
#include "llvm/IR/DataLayout.h"
#include "llvm/IR/LegacyPassManager.h"
#include "llvm/IR/Mangler.h"
#include "llvm/Support/DynamicLibrary.h"
#include "llvm/Target/TargetMachine.h"
#include "llvm/Support/raw_ostream.h"
#include "llvm/Transforms/Scalar.h"
```

```

#include "llvm/Transforms/Scalar/GVN.h"
#include <algorithm>
#include <memory>
#include <string>
#include <vector>

namespace llvm {
namespace orc {

class KaleidoscopeJIT {
private:
    std::unique_ptr<TargetMachine> TM;
    const DataLayout DL;
    RTDyldObjectLinkingLayer ObjectLayer;
    IRCompileLayer<decltype(ObjectLayer), SimpleCompiler> CompileLayer;

    using OptimizeFunction =
        std::function<std::shared_ptr<Module>(std::shared_ptr<Module>)>>;

    IRTransformLayer<decltype(CompileLayer), OptimizeFunction> OptimizeLayer;

public:
    using ModuleHandle = decltype(OptimizeLayer)::ModuleHandleT;

    KaleidoscopeJIT()
        : TM(EngineBuilder().selectTarget(), DL(TM->createDataLayout()),
            ObjectLayer([]() { return std::make_shared<SectionMemoryManager>(); })),
          CompileLayer(ObjectLayer, SimpleCompiler(*TM)),
          OptimizeLayer(CompileLayer,
                        [this](std::shared_ptr<Module> M) {
                            return optimizeModule(std::move(M));
                        }) {
        llvm::sys::DynamicLibrary::LoadLibraryPermanently(nullptr);
    }

    TargetMachine &getTargetMachine() { return *TM; }

    ModuleHandle addModule(std::unique_ptr<Module> M) {
        // Build our symbol resolver:
        // Lambda 1: Look back into the JIT itself to find symbols that are part of
        //           the same "logical dylib".
        // Lambda 2: Search for external symbols in the host process.
        auto Resolver = createLambdaResolver(
            [&](const std::string &Name) {
                if (auto Sym = OptimizeLayer.findSymbol(Name, false))
                    return Sym;
                return JITSymbol(nullptr);
            },
            [&](const std::string &Name) {
                if (auto SymAddr =
                    RTDyldMemoryManager::getSymbolAddressInProcess(Name))
                    return JITSymbol(SymAddr, JITSymbolFlags::Exported);
                return JITSymbol(nullptr);
            });

        // Add the set to the JIT with the resolver we created above and a newly
        // created SectionMemoryManager.
        return cantFail(OptimizeLayer.addModule(std::move(M),

```

```

        std::move(Resolver)));
    }

JITSymbol findSymbol(const std::string Name) {
    std::string MangledName;
    raw_string_ostream MangledNameStream(MangledName);
    Mangler::getNameWithPrefix(MangledNameStream, Name, DL);
    return OptimizeLayer.findSymbol(MangledNameStream.str(), true);
}

void removeModule(ModuleHandle H) {
    cantFail(OptimizeLayer.removeModule(H));
}

private:
std::shared_ptr<Module> optimizeModule(std::shared_ptr<Module> M) {
    // Create a function pass manager.
    auto FPM = llvm::make_unique<legacy::FunctionPassManager>(M.get());

    // Add some optimizations.
    FPM->add(createInstructionCombiningPass());
    FPM->add(createReassociatePass());
    FPM->add(createGVNPass());
    FPM->add(createCFGSimplificationPass());
    FPM->doInitialization();

    // Run the optimizations over all functions in the module being added to
    // the JIT.
    for (auto &F : *M)
        FPM->run(F);

    return M;
}
};

} // end namespace orc
} // end namespace llvm

#endif // LLVM_EXECUTIONENGINE_ORC_KALEIDOSCOPEJIT_H

```

Building a JIT: Per-function Lazy Compilation

- *Chapter 3 Introduction*
- *Lazy Compilation*
- *Full Code Listing*

This tutorial is under active development. It is incomplete and details may change frequently. Nonetheless we invite you to try it out as it stands, and we welcome any feedback.

Chapter 3 Introduction

Welcome to Chapter 3 of the “Building an ORC-based JIT in LLVM” tutorial. This chapter discusses lazy JITing and shows you how to enable it by adding an ORC CompileOnDemand layer the JIT from [Chapter 2](#).

Lazy Compilation

When we add a module to the KaleidoscopeJIT class from Chapter 2 it is immediately optimized, compiled and linked for us by the IRTransformLayer, IRCompileLayer and ObjectLinkingLayer respectively. This scheme, where all the work to make a Module executable is done up front, is simple to understand and its performance characteristics are easy to reason about. However, it will lead to very high startup times if the amount of code to be compiled is large, and may also do a lot of unnecessary compilation if only a few compiled functions are ever called at runtime. A truly “just-in-time” compiler should allow us to defer the compilation of any given function until the moment that function is first called, improving launch times and eliminating redundant work. In fact, the ORC APIs provide us with a layer to lazily compile LLVM IR: *CompileOnDemandLayer*.

The CompileOnDemandLayer class conforms to the layer interface described in Chapter 2, but its addModuleSet method behaves quite differently from the layers we have seen so far: rather than doing any work up front, it just scans the Modules being added and arranges for each function in them to be compiled the first time it is called. To do this, the CompileOnDemandLayer creates two small utilities for each function that it scans: a *stub* and a *compile callback*. The stub is a pair of a function pointer (which will be pointed at the function’s implementation once the function has been compiled) and an indirect jump through the pointer. By fixing the address of the indirect jump for the lifetime of the program we can give the function a permanent “effective address”, one that can be safely used for indirection and function pointer comparison even if the function’s implementation is never compiled, or if it is compiled more than once (due to, for example, recompiling the function at a higher optimization level) and changes address. The second utility, the compile callback, represents a re-entry point from the program into the compiler that will trigger compilation and then execution of a function. By initializing the function’s stub to point at the function’s compile callback, we enable lazy compilation: The first attempted call to the function will follow the function pointer and trigger the compile callback instead. The compile callback will compile the function, update the function pointer for the stub, then execute the function. On all subsequent calls to the function, the function pointer will point at the already-compiled function, so there is no further overhead from the compiler. We will look at this process in more detail in the next chapter of this tutorial, but for now we’ll trust the CompileOnDemandLayer to set all the stubs and callbacks up for us. All we need to do is to add the CompileOnDemandLayer to the top of our stack and we’ll get the benefits of lazy compilation. We just need a few changes to the source:

```
...
#include "llvm/ExecutionEngine/SectionMemoryManager.h"
#include "llvm/ExecutionEngine/Orc/CompileOnDemandLayer.h"
#include "llvm/ExecutionEngine/Orc/CompileUtils.h"
...

...
class KaleidoscopeJIT {
private:
    std::unique_ptr<TargetMachine> TM;
    const DataLayout DL;
    std::unique_ptr<JITCompileCallbackManager> CompileCallbackManager;
    ObjectLinkingLayer<> ObjectLayer;
    IRCompileLayer<decltype(ObjectLayer)> CompileLayer;

    typedef std::function<std::unique_ptr<Module>(std::unique_ptr<Module>)>
        OptimizeFunction;

    IRTransformLayer<decltype(CompileLayer), OptimizeFunction> OptimizeLayer;
    CompileOnDemandLayer<decltype(OptimizeLayer)> CODLayer;
```



```
public:
    typedef decltype(CODLayer)::ModuleSetHandleT ModuleHandle;
```

First we need to include the `CompileOnDemandLayer.h` header, then add two new members: a `std::unique_ptr<CompileCallbackManager>` and a `CompileOnDemandLayer`, to our class. The `CompileCallbackManager` member is used by the `CompileOnDemandLayer` to create the compile callback needed for each function.

```
KaleidoscopeJIT()
: TM(EngineBuilder().selectTarget()), DL(TM->createDataLayout()),
  CompileLayer(ObjectLayer, SimpleCompiler(*TM)),
  OptimizeLayer(CompileLayer,
    [this](std::unique_ptr<Module> M) {
        return optimizeModule(std::move(M));
    }),
  CompileCallbackManager(
    orc::createLocalCompileCallbackManager(TM->getTargetTriple(), 0)),
  CODLayer(OptimizeLayer,
    [this](Function &F) { return std::set<Function*>({&F}); },
    *CompileCallbackManager,
    orc::createLocalIndirectStubsManagerBuilder(
      TM->getTargetTriple())) {
    llvm::sys::DynamicLibrary::LoadLibraryPermanently(nullptr);
}
```

Next we have to update our constructor to initialize the new members. To create an appropriate compile callback manager we use the `createLocalCompileCallbackManager` function, which takes a `TargetMachine` and a `JITTargetAddress` to call if it receives a request to compile an unknown function. In our simple JIT this situation is unlikely to come up, so we'll cheat and just pass '0' here. In a production quality JIT you could give the address of a function that throws an exception in order to unwind the JIT'd code's stack.

Now we can construct our `CompileOnDemandLayer`. Following the pattern from previous layers we start by passing a reference to the next layer down in our stack – the `OptimizeLayer`. Next we need to supply a 'partitioning function': when a not-yet-compiled function is called, the `CompileOnDemandLayer` will call this function to ask us what we would like to compile. At a minimum we need to compile the function being called (given by the argument to the partitioning function), but we could also request that the `CompileOnDemandLayer` compile other functions that are unconditionally called (or highly likely to be called) from the function being called. For `KaleidoscopeJIT` we'll keep it simple and just request compilation of the function that was called. Next we pass a reference to our `CompileCallbackManager`. Finally, we need to supply an "indirect stubs manager builder": a utility function that constructs `IndirectStubManagers`, which are in turn used to build the stubs for the functions in each module. The `CompileOnDemandLayer` will call the indirect stub manager builder once for each call to `addModuleSet`, and use the resulting indirect stubs manager to create stubs for all functions in all modules in the set. If/when the module set is removed from the JIT the indirect stubs manager will be deleted, freeing any memory allocated to the stubs. We supply this function by using the `createLocalIndirectStubsManagerBuilder` utility.

```
// ...
    if (auto Sym = CODLayer.findSymbol(Name, false))
// ...
return CODLayer.addModuleSet(std::move(Ms),
                             make_unique<SectionMemoryManager>(),
                             std::move(Resolver));
// ...

// ...
return CODLayer.findSymbol(MangledNameStream.str(), true);
// ...
```

```
// ...
CODLayer.removeModuleSet(H);
// ...
```

Finally, we need to replace the references to `OptimizeLayer` in our `addModule`, `findSymbol`, and `removeModule` methods. With that, we're up and running.

To be done:

**** Chapter conclusion.****

Full Code Listing

Here is the complete code listing for our running example with a `CompileOnDemand` layer added to enable lazy function-at-a-time compilation. To build this example, use:

```
# Compile
clang++ -g toy.cpp `llvm-config --cxxflags --ldflags --system-libs --libs core orc_
↪native` -O3 -o toy
# Run
./toy
```

Here is the code:

```
//===- KaleidoscopeJIT.h - A simple JIT for Kaleidoscope -----*- C++ -*-===//
//
//                               The LLVM Compiler Infrastructure
//
// This file is distributed under the University of Illinois Open Source
// License. See LICENSE.TXT for details.
//
//===-----//
//
// Contains a simple JIT definition for use in the kaleidoscope tutorials.
//
//===-----//

#ifndef LLVM_EXECUTIONENGINE_ORC_KALEIDOSCOPEJIT_H
#define LLVM_EXECUTIONENGINE_ORC_KALEIDOSCOPEJIT_H

#include "llvm/ADT/STLExtras.h"
#include "llvm/ExecutionEngine/ExecutionEngine.h"
#include "llvm/ExecutionEngine/JITSymbol.h"
#include "llvm/ExecutionEngine/RTDyldMemoryManager.h"
#include "llvm/ExecutionEngine/RuntimeDyld.h"
#include "llvm/ExecutionEngine/SectionMemoryManager.h"
#include "llvm/ExecutionEngine/Orc/CompileOnDemandLayer.h"
#include "llvm/ExecutionEngine/Orc/CompileUtils.h"
#include "llvm/ExecutionEngine/Orc/IRCompileLayer.h"
#include "llvm/ExecutionEngine/Orc/IRTransformLayer.h"
#include "llvm/ExecutionEngine/Orc/LambdaResolver.h"
#include "llvm/ExecutionEngine/Orc/RTDyldObjectLinkingLayer.h"
#include "llvm/IR/DataLayout.h"
#include "llvm/IR/LegacyPassManager.h"
#include "llvm/IR/Mangler.h"
#include "llvm/Support/DynamicLibrary.h"
#include "llvm/Support/raw_ostream.h"
```

```

#include "llvm/Target/TargetMachine.h"
#include "llvm/Transforms/Scalar.h"
#include "llvm/Transforms/Scalar/GVN.h"
#include <algorithm>
#include <memory>
#include <set>
#include <string>
#include <vector>

namespace llvm {
namespace orc {

class KaleidoscopeJIT {
private:
    std::unique_ptr<TargetMachine> TM;
    const DataLayout DL;
    RTDyldObjectLinkingLayer ObjectLayer;
    IRCompileLayer<decltype(ObjectLayer), SimpleCompiler> CompileLayer;

    using OptimizeFunction =
        std::function<std::shared_ptr<Module>(std::shared_ptr<Module>)>;

    IRTransformLayer<decltype(CompileLayer), OptimizeFunction> OptimizeLayer;

    std::unique_ptr<JITCompileCallbackManager> CompileCallbackManager;
    CompileOnDemandLayer<decltype(OptimizeLayer)> CODLayer;

public:
    using ModuleHandle = decltype(CODLayer)::ModuleHandleT;

    KaleidoscopeJIT()
        : TM(EngineBuilder().selectTarget()), DL(TM->createDataLayout()),
          ObjectLayer([]() { return std::make_shared<SectionMemoryManager>(); }),
          CompileLayer(ObjectLayer, SimpleCompiler(*TM)),
          OptimizeLayer(CompileLayer,
                        [this](std::shared_ptr<Module> M) {
                            return optimizeModule(std::move(M));
                        }),
          CompileCallbackManager(
              orc::createLocalCompileCallbackManager(TM->getTargetTriple(), 0)),
          CODLayer(OptimizeLayer,
                  [](Function &F) { return std::set<Function*>({&F}); },
                  *CompileCallbackManager,
                  orc::createLocalIndirectStubsManagerBuilder(
                      TM->getTargetTriple())) {
        llvm::sys::DynamicLibrary::LoadLibraryPermanently(nullptr);
    }

    TargetMachine &getTargetMachine() { return *TM; }

    ModuleHandle addModule(std::unique_ptr<Module> M) {
        // Build our symbol resolver:
        // Lambda 1: Look back into the JIT itself to find symbols that are part of
        //           the same "logical dylib".
        // Lambda 2: Search for external symbols in the host process.
        auto Resolver = createLambdaResolver(
            [&](const std::string &Name) {
                if (auto Sym = CODLayer.findSymbol(Name, false))

```

```
        return Sym;
        return JITSymbol(nullptr);
    },
    [] (const std::string &Name) {
        if (auto SymAddr =
            RTDyldMemoryManager::getSymbolAddressInProcess(Name))
            return JITSymbol(SymAddr, JITSymbolFlags::Exported);
        return JITSymbol(nullptr);
    });

    // Add the set to the JIT with the resolver we created above and a newly
    // created SectionMemoryManager.
    return cantFail(CODLayer.addModule(std::move(M), std::move(Resolver)));
}

JITSymbol findSymbol(const std::string Name) {
    std::string MangledName;
    raw_string_ostream MangledNameStream(MangledName);
    Mangler::getNameWithPrefix(MangledNameStream, Name, DL);
    return CODLayer.findSymbol(MangledNameStream.str(), true);
}

void removeModule(ModuleHandle H) {
    cantFail(CODLayer.removeModule(H));
}

private:
    std::shared_ptr<Module> optimizeModule(std::shared_ptr<Module> M) {
        // Create a function pass manager.
        auto FPM = llvm::make_unique<legacy::FunctionPassManager>(M.get());

        // Add some optimizations.
        FPM->add(createInstructionCombiningPass());
        FPM->add(createReassociatePass());
        FPM->add(createGVNPass());
        FPM->add(createCFGSimplificationPass());
        FPM->doInitialization();

        // Run the optimizations over all functions in the module being added to
        // the JIT.
        for (auto &F : *M)
            FPM->run(F);

        return M;
    }
};

} // end namespace orc
} // end namespace llvm

#endif // LLVM_EXECUTIONENGINE_ORC_KALEIDOSCOPEJIT_H
```

Next: Extreme Laziness – Using Compile Callbacks to JIT directly from ASTs

Building a JIT: Extreme Laziness - Using Compile Callbacks to JIT from ASTs

- [Chapter 4 Introduction](#)
- [Full Code Listing](#)

This tutorial is under active development. It is incomplete and details may change frequently. Nonetheless we invite you to try it out as it stands, and we welcome any feedback.

Chapter 4 Introduction

Welcome to Chapter 4 of the “Building an ORC-based JIT in LLVM” tutorial. This chapter introduces the Compile Callbacks and Indirect Stubs APIs and shows how they can be used to replace the CompileOnDemand layer from [Chapter 3](#) with a custom lazy-JITing scheme that JITs directly from Kaleidoscope ASTs.

To be done:

- (1) **Describe the drawbacks of JITing from IR (have to compile to IR first, which reduces the benefits of laziness).**
- (2) **Describe CompileCallbackManagers and IndirectStubManagers in detail.**
- (3) **Run through the implementation of addFunctionAST.**

Full Code Listing

Here is the complete code listing for our running example that JITs lazily from Kaleidoscope ASTs. To build this example, use:

```
# Compile
clang++ -g toy.cpp `llvm-config --cxxflags --ldflags --system-libs --libs core orc_
↪native` -O3 -o toy
# Run
./toy
```

Here is the code:

```
//====- KaleidoscopeJIT.h - A simple JIT for Kaleidoscope -----*- C++ -*-====//
//
//                               The LLVM Compiler Infrastructure
//
// This file is distributed under the University of Illinois Open Source
// License. See LICENSE.TXT for details.
//
//=====//
//
// Contains a simple JIT definition for use in the kaleidoscope tutorials.
//
//=====//

#ifndef LLVM_EXECUTIONENGINE_ORC_KALEIDOSCOPEJIT_H
#define LLVM_EXECUTIONENGINE_ORC_KALEIDOSCOPEJIT_H

#include "llvm/ADT/STLExtras.h"
#include "llvm/ExecutionEngine/ExecutionEngine.h"
#include "llvm/ExecutionEngine/JITSymbol.h"
#include "llvm/ExecutionEngine/RTDyldMemoryManager.h"
#include "llvm/ExecutionEngine/SectionMemoryManager.h"
#include "llvm/ExecutionEngine/Orc/CompileUtils.h"
```

```
#include "llvm/ExecutionEngine/Orc/IndirectionUtils.h"
#include "llvm/ExecutionEngine/Orc/IRCompileLayer.h"
#include "llvm/ExecutionEngine/Orc/IRTransformLayer.h"
#include "llvm/ExecutionEngine/Orc/LambdaResolver.h"
#include "llvm/ExecutionEngine/Orc/RTDyldObjectLinkingLayer.h"
#include "llvm/IR/DataLayout.h"
#include "llvm/IR/LegacyPassManager.h"
#include "llvm/IR/Mangler.h"
#include "llvm/Support/DynamicLibrary.h"
#include "llvm/Support/Error.h"
#include "llvm/Support/raw_ostream.h"
#include "llvm/Target/TargetMachine.h"
#include "llvm/Transforms/Scalar.h"
#include "llvm/Transforms/Scalar/GVN.h"
#include <algorithm>
#include <cassert>
#include <cstdlib>
#include <memory>
#include <string>
#include <vector>

class PrototypeAST;
class ExprAST;

/// FunctionAST - This class represents a function definition itself.
class FunctionAST {
    std::unique_ptr<PrototypeAST> Proto;
    std::unique_ptr<ExprAST> Body;

public:
    FunctionAST(std::unique_ptr<PrototypeAST> Proto,
               std::unique_ptr<ExprAST> Body)
        : Proto(std::move(Proto)), Body(std::move(Body)) {}

    const PrototypeAST& getProto() const;
    const std::string& getName() const;
    llvm::Function *codegen();
};

/// This will compile FnAST to IR, rename the function to add the given
/// suffix (needed to prevent a name-clash with the function's stub),
/// and then take ownership of the module that the function was compiled
/// into.
std::unique_ptr<llvm::Module>
irgenAndTakeOwnership(FunctionAST &FnAST, const std::string &Suffix);

namespace llvm {
namespace orc {

class KaleidoscopeJIT {
private:
    std::unique_ptr<TargetMachine> TM;
    const DataLayout DL;
    RTDyldObjectLinkingLayer ObjectLayer;
    IRCompileLayer<decltype(ObjectLayer), SimpleCompiler> CompileLayer;

    using OptimizeFunction =
        std::function<std::shared_ptr<Module>(std::shared_ptr<Module>)>;
```

```

IRTransformLayer<decltype(CompileLayer), OptimizeFunction> OptimizeLayer;

std::unique_ptr<JITCompileCallbackManager> CompileCallbackMgr;
std::unique_ptr<IndirectStubsManager> IndirectStubsMgr;

public:
    using ModuleHandle = decltype(OptimizeLayer)::ModuleHandleT;

    KaleidoscopeJIT()
        : TM(EngineBuilder().selectTarget()),
          DL(TM->createDataLayout()),
          ObjectLayer([]() { return std::make_shared<SectionMemoryManager>(); }),
          CompileLayer(ObjectLayer, SimpleCompiler(*TM)),
          OptimizeLayer(CompileLayer,
                        [this](std::shared_ptr<Module> M) {
                            return optimizeModule(std::move(M));
                        }),
          CompileCallbackMgr(
              orc::createLocalCompileCallbackManager(TM->getTargetTriple(), 0)) {
        auto IndirectStubsMgrBuilder =
            orc::createLocalIndirectStubsManagerBuilder(TM->getTargetTriple());
        IndirectStubsMgr = IndirectStubsMgrBuilder();
        llvm::sys::DynamicLibrary::LoadLibraryPermanently(nullptr);
    }

    TargetMachine &getTargetMachine() { return *TM; }

    ModuleHandle addModule(std::unique_ptr<Module> M) {
        // Build our symbol resolver:
        // Lambda 1: Look back into the JIT itself to find symbols that are part of
        //           the same "logical dylib".
        // Lambda 2: Search for external symbols in the host process.
        auto Resolver = createLambdaResolver(
            [&](const std::string &Name) {
                if (auto Sym = IndirectStubsMgr->findStub(Name, false))
                    return Sym;
                if (auto Sym = OptimizeLayer.findSymbol(Name, false))
                    return Sym;
                return JITSymbol(nullptr);
            },
            [&](const std::string &Name) {
                if (auto SymAddr =
                    RTDyldMemoryManager::getSymbolAddressInProcess(Name))
                    return JITSymbol(SymAddr, JITSymbolFlags::Exported);
                return JITSymbol(nullptr);
            });

        // Add the set to the JIT with the resolver we created above and a newly
        // created SectionMemoryManager.
        return cantFail(OptimizeLayer.addModule(std::move(M),
                                                  std::move(Resolver)));
    }

    Error addFunctionAST(std::unique_ptr<FunctionAST> FnAST) {
        // Create a CompileCallback - this is the re-entry point into the compiler
        // for functions that haven't been compiled yet.
        auto CCInfo = CompileCallbackMgr->getCompileCallback();
    }

```

```

// Create an indirect stub. This serves as the functions "canonical
// definition" - an unchanging (constant address) entry point to the
// function implementation.
// Initially we point the stub's function-pointer at the compile callback
// that we just created. In the compile action for the callback (see below)
// we will update the stub's function pointer to point at the function
// implementation that we just implemented.
if (auto Err = IndirectStubsMgr->createStub(mangle(FnAST->getName()),
                                           CCInfo.getAddress(),
                                           JITSymbolFlags::Exported))

    return Err;

// Move ownership of FnAST to a shared pointer - C++11 lambdas don't support
// capture-by-move, which is required for unique_ptr.
auto SharedFnAST = std::shared_ptr<FunctionAST>(std::move(FnAST));

// Set the action to compile our AST. This lambda will be run if/when
// execution hits the compile callback (via the stub).
//
// The steps to compile are:
// (1) IRGen the function.
// (2) Add the IR module to the JIT to make it executable like any other
//     module.
// (3) Use findSymbol to get the address of the compiled function.
// (4) Update the stub pointer to point at the implementation so that
//     subsequent calls go directly to it and bypass the compiler.
// (5) Return the address of the implementation: this lambda will actually
//     be run inside an attempted call to the function, and we need to
//     continue on to the implementation to complete the attempted call.
//     The JIT runtime (the resolver block) will use the return address of
//     this function as the address to continue at once it has reset the
//     CPU state to what it was immediately before the call.
CCInfo.setCompileAction(
    [this, SharedFnAST]() {
        auto M = irgenAndTakeOwnership(*SharedFnAST, "$impl");
        addModule(std::move(M));
        auto Sym = findSymbol(SharedFnAST->getName() + "$impl");
        assert(Sym && "Couldn't find compiled function?");
        JITTargetAddress SymAddr = cantFail(Sym.getAddress());
        if (auto Err =
            IndirectStubsMgr->updatePointer(mangle(SharedFnAST->getName()),
                                           SymAddr)) {
            logAllUnhandledErrors(std::move(Err), errs(),
                                "Error updating function pointer: ");
            exit(1);
        }

        return SymAddr;
    });

return Error::success();
}

JITSymbol findSymbol(const std::string Name) {
    return OptimizeLayer.findSymbol(mangle(Name), true);
}

```



```

void removeModule(ModuleHandle H) {
    cantFail(OptimizeLayer.removeModule(H));
}

private:
std::string mangle(const std::string &Name) {
    std::string MangledName;
    raw_string_ostream MangledNameStream(MangledName);
    Mangler::getNameWithPrefix(MangledNameStream, Name, DL);
    return MangledNameStream.str();
}

std::shared_ptr<Module> optimizeModule(std::shared_ptr<Module> M) {
    // Create a function pass manager.
    auto FPM = llvm::make_unique<legacy::FunctionPassManager>(M.get());

    // Add some optimizations.
    FPM->add(createInstructionCombiningPass());
    FPM->add(createReassociatePass());
    FPM->add(createGVNPass());
    FPM->add(createCFGSimplificationPass());
    FPM->doInitialization();

    // Run the optimizations over all functions in the module being added to
    // the JIT.
    for (auto &F : *M)
        FPM->run(F);

    return M;
}
};

} // end namespace orc
} // end namespace llvm

#endif // LLVM_EXECUTIONENGINE_ORC_KALEIDOSCOPEJIT_H

```

Next: Remote-JITing – Process-isolation and laziness-at-a-distance

Building a JIT: Remote-JITing – Process Isolation and Laziness at a Distance

- [Chapter 5 Introduction](#)
- [Full Code Listing](#)

This tutorial is under active development. It is incomplete and details may change frequently. Nonetheless we invite you to try it out as it stands, and we welcome any feedback.

Chapter 5 Introduction

Welcome to Chapter 5 of the “Building an ORC-based JIT in LLVM” tutorial. This chapter introduces the ORC RemoteJIT Client/Server APIs and shows how to use them to build a JIT stack that will execute its code via a communications channel with a different process. This can be a separate process on the same machine, a process on a different

machine, or even a process on a different platform/architecture. The code builds on top of the lazy-AST-compiling JIT stack from [Chapter 4](#).

To be done – this is going to be a long one:

(1) Introduce channels, RPC, RemoteJIT Client and Server APIs

(2) Describe the client code in greater detail. Discuss modifications of the KaleidoscopeJIT class, and the REPL itself.

(3) Describe the server code.

(4) Describe how to run the demo.

Full Code Listing

Here is the complete code listing for our running example that JITs lazily from Kaleidoscope ASTS. To build this example, use:

```
# Compile
clang++ -g toy.cpp `llvm-config --cxxflags --ldflags --system-libs --libs core orc_
↪native` -O3 -o toy
# Run
./toy
```

Here is the code for the modified KaleidoscopeJIT:

```
//===-- KaleidoscopeJIT.h - A simple JIT for Kaleidoscope -----*- C++ -*-===//
//
//                               The LLVM Compiler Infrastructure
//
// This file is distributed under the University of Illinois Open Source
// License. See LICENSE.TXT for details.
//
//===-----//
//
// Contains a simple JIT definition for use in the kaleidoscope tutorials.
//
//===-----//

#ifndef LLVM_EXECUTIONENGINE_ORC_KALEIDOSCOPEJIT_H
#define LLVM_EXECUTIONENGINE_ORC_KALEIDOSCOPEJIT_H

#include "RemoteJITUtils.h"
#include "llvm/ADT/SmallVector.h"
#include "llvm/ADT/STLExtras.h"
#include "llvm/ADT/Triple.h"
#include "llvm/ExecutionEngine/ExecutionEngine.h"
#include "llvm/ExecutionEngine/JITSymbol.h"
#include "llvm/ExecutionEngine/Orc/CompileUtils.h"
#include "llvm/ExecutionEngine/Orc/IndirectionUtils.h"
#include "llvm/ExecutionEngine/Orc/IRCompileLayer.h"
#include "llvm/ExecutionEngine/Orc/IRTransformLayer.h"
#include "llvm/ExecutionEngine/Orc/LambdaResolver.h"
#include "llvm/ExecutionEngine/Orc/RTDyldObjectLinkingLayer.h"
#include "llvm/ExecutionEngine/Orc/OrcRemoteTargetClient.h"
#include "llvm/IR/DataLayout.h"
#include "llvm/IR/LegacyPassManager.h"
#include "llvm/IR/Mangler.h"
```

```

#include "llvm/Support/DynamicLibrary.h"
#include "llvm/Support/Error.h"
#include "llvm/Support/raw_ostream.h"
#include "llvm/Target/TargetMachine.h"
#include "llvm/Transforms/Scalar.h"
#include "llvm/Transforms/Scalar/GVN.h"
#include <algorithm>
#include <cassert>
#include <cstdlib>
#include <memory>
#include <string>
#include <vector>

class PrototypeAST;
class ExprAST;

/// FunctionAST - This class represents a function definition itself.
class FunctionAST {
    std::unique_ptr<PrototypeAST> Proto;
    std::unique_ptr<ExprAST> Body;

public:
    FunctionAST(std::unique_ptr<PrototypeAST> Proto,
                std::unique_ptr<ExprAST> Body)
        : Proto(std::move(Proto)), Body(std::move(Body)) {}

    const PrototypeAST& getProto() const;
    const std::string& getName() const;
    llvm::Function *codegen();
};

/// This will compile FnAST to IR, rename the function to add the given
/// suffix (needed to prevent a name-clash with the function's stub),
/// and then take ownership of the module that the function was compiled
/// into.
std::unique_ptr<llvm::Module>
irgenAndTakeOwnership(FunctionAST &FnAST, const std::string &Suffix);

namespace llvm {
namespace orc {

/// Typedef the remote-client API.
using MyRemote = remote::OrcRemoteTargetClient<FDRPCChannel>;

class KaleidoscopeJIT {
private:
    std::unique_ptr<TargetMachine> TM;
    const DataLayout DL;
    RTDyldObjectLinkingLayer ObjectLayer;
    IRCompileLayer<decltype(ObjectLayer), SimpleCompiler> CompileLayer;

    using OptimizeFunction =
        std::function<std::shared_ptr<Module>(std::shared_ptr<Module>)>;

    IRTransformLayer<decltype(CompileLayer), OptimizeFunction> OptimizeLayer;

    JITCompileCallbackManager *CompileCallbackMgr;
    std::unique_ptr<IndirectStubsManager> IndirectStubsMgr;

```

```

MyRemote &Remote;

public:
    using ModuleHandle = decltype(OptimizeLayer)::ModuleHandleT;

    KaleidoscopeJIT(MyRemote &Remote)
        : TM(EngineBuilder().selectTarget(Triple(Remote.getTargetTriple()), "",
                                              "", SmallVector<std::string, 0>())),
          DL(TM->createDataLayout()),
          ObjectLayer([&Remote]() {
              std::unique_ptr<MyRemote::RCMemoryManager> MemMgr;
              if (auto Err = Remote.createRemoteMemoryManager(MemMgr)) {
                  logAllUnhandledErrors(std::move(Err), errs(),
                                         "Error creating remote memory manager:");
                  exit(1);
              }
              return MemMgr;
          }),
          CompileLayer(ObjectLayer, SimpleCompiler(*TM)),
          OptimizeLayer(CompileLayer,
                        [this](std::shared_ptr<Module> M) {
                            return optimizeModule(std::move(M));
                        }),
          Remote(Remote) {
        auto CCMgrOrErr = Remote.enableCompileCallbacks(0);
        if (!CCMgrOrErr) {
            logAllUnhandledErrors(CCMgrOrErr.takeError(), errs(),
                                   "Error enabling remote compile callbacks:");
            exit(1);
        }
        CompileCallbackMgr = &*CCMgrOrErr;
        std::unique_ptr<MyRemote::RCIndirectStubsManager> ISM;
        if (auto Err = Remote.createIndirectStubsManager(ISM)) {
            logAllUnhandledErrors(std::move(Err), errs(),
                                   "Error creating indirect stubs manager:");
            exit(1);
        }
        IndirectStubsMgr = std::move(ISM);
        llvm::sys::DynamicLibrary::LoadLibraryPermanently(nullptr);
    }

    TargetMachine &getTargetMachine() { return *TM; }

    ModuleHandle addModule(std::unique_ptr<Module> M) {
        // Build our symbol resolver:
        // Lambda 1: Look back into the JIT itself to find symbols that are part of
        //           the same "logical dylib".
        // Lambda 2: Search for external symbols in the host process.
        auto Resolver = createLambdaResolver(
            [&](const std::string &Name) {
                if (auto Sym = IndirectStubsMgr->findStub(Name, false))
                    return Sym;
                if (auto Sym = OptimizeLayer.findSymbol(Name, false))
                    return Sym;
                return JITSymbol(nullptr);
            },
            [&](const std::string &Name) {
                if (auto AddrOrErr = Remote.getSymbolAddress(Name))

```

```

        return JITSymbol(*AddrOrErr, JITSymbolFlags::Exported);
    else {
        logAllUnhandledErrors(AddrOrErr.takeError(), errs(),
                               "Error resolving remote symbol:");
        exit(1);
    }
    return JITSymbol(nullptr);
});

// Add the set to the JIT with the resolver we created above and a newly
// created SectionMemoryManager.
return cantFail(OptimizeLayer.addModule(std::move(M),
                                         std::move(Resolver)));
}

Error addFunctionAST(std::unique_ptr<FunctionAST> FnAST) {
    // Create a CompileCallback - this is the re-entry point into the compiler
    // for functions that haven't been compiled yet.
    auto CCInfo = CompileCallbackMgr->getCompileCallback();

    // Create an indirect stub. This serves as the functions "canonical
    // definition" - an unchanging (constant address) entry point to the
    // function implementation.
    // Initially we point the stub's function-pointer at the compile callback
    // that we just created. In the compile action for the callback (see below)
    // we will update the stub's function pointer to point at the function
    // implementation that we just implemented.
    if (auto Err = IndirectStubsMgr->createStub(mangle(FnAST->getName()),
                                                CCInfo.getAddress(),
                                                JITSymbolFlags::Exported))

        return Err;

    // Move ownership of FnAST to a shared pointer - C++11 lambdas don't support
    // capture-by-move, which is required for unique_ptr.
    auto SharedFnAST = std::shared_ptr<FunctionAST>(std::move(FnAST));

    // Set the action to compile our AST. This lambda will be run if/when
    // execution hits the compile callback (via the stub).
    //
    // The steps to compile are:
    // (1) IRGen the function.
    // (2) Add the IR module to the JIT to make it executable like any other
    //     module.
    // (3) Use findSymbol to get the address of the compiled function.
    // (4) Update the stub pointer to point at the implementation so that
    //     subsequent calls go directly to it and bypass the compiler.
    // (5) Return the address of the implementation: this lambda will actually
    //     be run inside an attempted call to the function, and we need to
    //     continue on to the implementation to complete the attempted call.
    //     The JIT runtime (the resolver block) will use the return address of
    //     this function as the address to continue at once it has reset the
    //     CPU state to what it was immediately before the call.
    CCInfo.setCompileAction(
        [this, SharedFnAST]() {
            auto M = irgenAndTakeOwnership(*SharedFnAST, "$impl");
            addModule(std::move(M));
            auto Sym = findSymbol(SharedFnAST->getName() + "$impl");
            assert(Sym && "Couldn't find compiled function?");

```

```

    JITTargetAddress SymAddr = cantFail(Sym.getAddress());
    if (auto Err =
        IndirectStubsMgr->updatePointer(mangle(SharedFnAST->getName()),
                                        SymAddr)) {
        logAllUnhandledErrors(std::move(Err), errs(),
                              "Error updating function pointer: ");
        exit(1);
    }

    return SymAddr;
});

return Error::success();
}

Error executeRemoteExpr(JITTargetAddress ExprAddr) {
    return Remote.callVoidVoid(ExprAddr);
}

JITSymbol findSymbol(const std::string Name) {
    return OptimizeLayer.findSymbol(mangle(Name), true);
}

void removeModule(ModuleHandle H) {
    cantFail(OptimizeLayer.removeModule(H));
}

private:
std::string mangle(const std::string &Name) {
    std::string MangledName;
    raw_string_ostream MangledNameStream(MangledName);
    Mangler::getNameWithPrefix(MangledNameStream, Name, DL);
    return MangledNameStream.str();
}

std::shared_ptr<Module> optimizeModule(std::shared_ptr<Module> M) {
    // Create a function pass manager.
    auto FPM = llvm::make_unique<legacy::FunctionPassManager>(M.get());

    // Add some optimizations.
    FPM->add(createInstructionCombiningPass());
    FPM->add(createReassociatePass());
    FPM->add(createGVNPass());
    FPM->add(createCFGSimplificationPass());
    FPM->doInitialization();

    // Run the optimizations over all functions in the module being added to
    // the JIT.
    for (auto &F : *M)
        FPM->run(F);

    return M;
}
};

} // end namespace orc
} // end namespace llvm

```

```
#endif // LLVM_EXECUTIONENGINE_ORC_KALEIDOSCOPEJIT_H
```

And the code for the JIT server:

```
#include "../RemoteJITUtils.h"
#include "llvm/ExecutionEngine/RTDyldMemoryManager.h"
#include "llvm/ExecutionEngine/Orc/OrcRemoteTargetServer.h"
#include "llvm/ExecutionEngine/Orc/OrcABISupport.h"
#include "llvm/Support/CommandLine.h"
#include "llvm/Support/DynamicLibrary.h"
#include "llvm/Support/Error.h"
#include "llvm/Support/raw_ostream.h"
#include "llvm/Support/TargetSelect.h"
#include <stdint>
#include <stdio>
#include <cstring>
#include <string>
#include <netinet/in.h>
#include <sys/socket.h>

using namespace llvm;
using namespace llvm::orc;

// Command line argument for TCP port.
cl::opt<uint32_t> Port("port",
                      cl::desc("TCP port to listen on"),
                      cl::init(20000));

ExitOnError ExitOnErr;

using MainFun = int (*)(int, const char*[]);

template <typename NativePtrT>
NativePtrT MakeNative(uint64_t P) {
    return reinterpret_cast<NativePtrT>(static_cast<uintptr_t>(P));
}

extern "C"
void printExprResult(double Val) {
    printf("Expression evaluated to: %f\n", Val);
}

// --- LAZY COMPILE TEST ---
int main(int argc, char* argv[]) {
    if (argc == 0)
        ExitOnErr.setBanner("jit_server: ");
    else
        ExitOnErr.setBanner(std::string(argv[0]) + ": ");

    // --- Initialize LLVM ---
    cl::ParseCommandLineOptions(argc, argv, "LLVM lazy JIT example.\n");

    InitializeNativeTarget();
    InitializeNativeTargetAsmPrinter();
    InitializeNativeTargetAsmParser();

    if (sys::DynamicLibrary::LoadLibraryPermanently(nullptr)) {
        errs() << "Error loading program symbols.\n";
    }
}
```

```

    return 1;
}

// --- Initialize remote connection ---

int sockfd = socket(PF_INET, SOCK_STREAM, 0);
sockaddr_in servAddr, clientAddr;
socklen_t clientAddrLen = sizeof(clientAddr);
memset(&servAddr, 0, sizeof(servAddr));
servAddr.sin_family = PF_INET;
servAddr.sin_family = INADDR_ANY;
servAddr.sin_port = htons(Port);

{
    // avoid "Address already in use" error.
    int yes = 1;
    if (setsockopt(sockfd, SOL_SOCKET, SO_REUSEADDR, &yes, sizeof(int)) == -1) {
        errs() << "Error calling setsockopt.\n";
        return 1;
    }
}

if (bind(sockfd, reinterpret_cast<sockaddr*>(&servAddr),
        sizeof(servAddr)) < 0) {
    errs() << "Error on binding.\n";
    return 1;
}
listen(sockfd, 1);
int newsockfd = accept(sockfd, reinterpret_cast<sockaddr*>(&clientAddr),
                      &clientAddrLen);

auto SymbolLookup =
    [] (const std::string &Name) {
        return RTDyldMemoryManager::getSymbolAddressInProcess(Name);
    };

auto RegisterEHFrames =
    [] (uint8_t *Addr, uint32_t Size) {
        RTDyldMemoryManager::registerEHFramesInProcess(Addr, Size);
    };

auto DeregisterEHFrames =
    [] (uint8_t *Addr, uint32_t Size) {
        RTDyldMemoryManager::deregisterEHFramesInProcess(Addr, Size);
    };

FDRPCChannel TCPChannel(newsockfd, newsockfd);

using MyServerT = remote::OrcRemoteTargetServer<FDRPCChannel, OrcX86_64_SysV>;

MyServerT Server(TCPChannel, SymbolLookup, RegisterEHFrames, DeregisterEHFrames);

while (!Server.receivedTerminate())
    ExitOnErr(Server.handleOne());

return 0;
}

```


2.17.4 External Tutorials

Tutorial: Creating an LLVM Backend for the Cpu0 Architecture A step-by-step tutorial for developing an LLVM backend. Under active development at <https://github.com/Jonathan2251/lbd> (please contribute!).

Howto: Implementing LLVM Integrated Assembler A simple guide for how to implement an LLVM integrated assembler for an architecture.

2.17.5 Advanced Topics

1. Writing an Optimization for LLVM

2.18 LLVM 5.0.0 Release Notes

- *Introduction*
- *Non-comprehensive list of changes in this release*
 - *Changes to the LLVM IR*
 - *Changes to the Arm Targets*
 - *Changes to the MIPS Target*
 - *Changes to the PowerPC Target*
 - *Changes to the X86 Target*
 - *Changes to the AMDGPU Target*
 - *Changes to the AVR Target*
 - *Changes to the C API*
- *External Open Source Projects Using LLVM 5*
 - *Zig Programming Language*
 - *LDC - the LLVM-based D compiler*
- *Additional Information*

2.18.1 Introduction

This document contains the release notes for the LLVM Compiler Infrastructure, release 5.0.0. Here we describe the status of LLVM, including major improvements from the previous release, improvements in various subprojects of LLVM, and some of the current users of the code. All LLVM releases may be downloaded from the [LLVM releases web site](#).

For more information about LLVM, including information about the latest release, please check out the [main LLVM web site](#). If you have questions or comments, the [LLVM Developer's Mailing List](#) is a good place to send them.

2.18.2 Non-comprehensive list of changes in this release

- LLVM's `WeakVH` has been renamed to `WeakTrackingVH` and a new `WeakVH` has been introduced. The new `WeakVH` nulls itself out on deletion, but does not track values across RAUW.
- A new library named `BinaryFormat` has been created which holds a collection of code which previously lived in `Support`. This includes the `file_magic` structure and `identify_magic` functions, as well as all the structure and type definitions for DWARF, ELF, COFF, WASM, and MachO file formats.
- The tool `llvm-pdbdump` has been renamed `llvm-pdbutil` to better reflect its nature as a general purpose PDB manipulation / diagnostics tool that does more than just dumping contents.
- The `BBVectorize` pass has been removed. It was fully replaced and no longer used back in 2014 but we didn't get around to removing it. Now it is gone. The SLP vectorizer is the suggested non-loop vectorization pass.
- A new tool `opt-viewer.py` has been added to visualize optimization remarks in HTML. The tool processes the YAML files produced by clang with the `-fsave-optimization-record` option.
- A new CMake macro `LLVM_REVERSE_ITERATION` has been added. If enabled, all supported unordered LLVM containers would be iterated in reverse order. This is useful for uncovering non-determinism caused by iteration of unordered containers. Currently, it supports reverse iteration of `SmallPtrSet` and `DenseMap`.
- A new tool `llvm-dlltool` has been added to create short import libraries from GNU style definition files. The tool utilizes the PE COFF SPEC Import Library Format and PE COFF Auxiliary Weak Externals Format to achieve compatibility with LLD and MSVC LINK.

Changes to the LLVM IR

- The `datalayout` string may now indicate an address space to use for the pointer type of `alloca` rather than the default of 0.
- Added `speculatable` attribute indicating a function which has no side-effects which could inhibit hoisting of calls.

Changes to the Arm Targets

During this release the AArch64 target has:

- A much improved Global ISel at O0.
- Support for ARMv8.1 8.2 and 8.3 instructions.
- New scheduler information for ThunderX2.
- Some SVE type changes but not much more than that.
- Made instruction fusion more aggressive, resulting in speedups for code making use of AArch64 AES instructions. AES fusion has been enabled for most Cortex-A cores and the `AArch64MacroFusion` pass was moved to the generic `MacroFusion` pass.
- Added preferred function alignments for most Cortex-A cores.
- OpenMP "offload-to-self" base support.

During this release the ARM target has:

- Improved, but still mostly broken, Global ISel.
- Scheduling models update, new schedule for Cortex-A57.

- Hardware breakpoint support in LLDB.
- New assembler error handling, with spelling corrections and multiple suggestions on how to fix problems.
- Improved mixed ARM/Thumb code generation. Some cases in which wrong relocations were emitted have been fixed.
- Added initial support for mixed ARM/Thumb link-time optimization, using the thumb-mode target feature.

Changes to the MIPS Target

- The microMIPS64R6 backend is deprecated and will be removed in the next release.
- The MIPS backend now directly supports vector types for arguments and return values (previously this required ABI specific LLVM IR).
- Added documentation for how the MIPS backend handles address lowering.
- Added a GCC compatible option `-m(no-)madd4` to control the generation of four operand multiply addition/subtraction instructions.
- Added basic support for the XRay instrumentation system.
- Added support for more assembly aliases and macros.
- Added support for the `micromips` and `nomicromips` function attributes which control micromips code generation on a per function basis.
- Added the `long-calls` feature for non-pic environments. This feature is used where the callee is out of range of the caller using a standard call sequence. It must be enabled specifically.
- Added support for performing microMIPS code generation via function attributes.
- Added experimental support for the static relocation model for the N64 ABI.
- Added partial support for the MT ASE.
- Added basic support for code size reduction for microMIPS.
- Fixed numerous bugs including: multi-precision arithmetic support, various vectorization bugs, debug information for thread local variables, debug sections lacking the correct flags, crashing when disassembling sections whose size is not a multiple of two or four.

Changes to the PowerPC Target

- Additional support and exploitation of POWER ISA 3.0: `vabsdub`, `vabsduh`, `vabsduw`, `modsw`, `moduw`, `modsd`, `modud`, `lxv`, `stxv`, `vextublx`, `vextubrx`, `vextuhlx`, `vextuhrx`, `vextuwlx`, `vextuwrx`, `vextsb2w`, `vextsb2d`, `vextsh2w`, `vextsh2d`, and `vextsw2d`
- Implemented Optimal Code Sequences from The PowerPC Compiler Writer's Guide.
- Enable `-fomit-frame-pointer` by default.
- Improved handling of bit reverse intrinsic.
- Improved handling of `memcpy` and `memcmp` functions.
- Improved handling of branches with static branch hints.
- Improved codegen for atomic `load_acquire`.
- Improved block placement during code layout
- Many improvements to instruction selection and code generation

Changes to the X86 Target

- Added initial AMD Ryzen (znver1) scheduler support.
- Added support for Intel Goldmont CPUs.
- Add support for `avx512vpopcntdq` instructions.
- Added heuristics to convert `CMOV` into branches when it may be profitable.
- More aggressive inlining of `memcmp` calls.
- Improve `vXi64` shuffles on 32-bit targets.
- Improved use of `PMOVMASKB` for `any_of/all_of` comparison reductions.
- Improved Silvermont, Sandybridge, and Jaguar (btver2) schedulers.
- Improved support for AVX512 vector rotations.
- Added support for AMD Lightweight Profiling (LWP) instructions.
- Avoid using slow `LEA` instructions.
- Use alternative sequences for multiply by constant.
- Improved lowering of strided shuffles.
- Improved the AVX512 cost model used by the vectorizer.
- Fix scalar code performance when AVX512 is enabled by making `il`'s illegal.
- Fixed many inline assembly bugs.
- Preliminary support for tracing NetBSD processes and core files with a single thread in LLDB.

Changes to the AMDGPU Target

- Initial gfx9 support

Changes to the AVR Target

This release consists mainly of bugfixes and implementations of features required for compiling basic Rust programs.

- Enable the branch relaxation pass so that we don't crash on large stack load/stores
- Add support for lowering bit-rotations to the native `ror` and `rol` instructions
- Fix bug where function pointers were treated as pointers to RAM and not pointers to program memory
- Fix broken code generation for shift-by-variable expressions
- Support zero-sized types in argument lists; this is impossible in C, but possible in Rust

Changes to the C API

- Deprecated the `LLVMAddBBVectorizePass` interface since the `BBVectorize` pass has been removed. It is now a no-op and will be removed in the next release. Use `LLVMAddSLPVectorizePass` instead to get the supported SLP vectorizer.

2.18.3 External Open Source Projects Using LLVM 5

Zig Programming Language

Zig is an open-source programming language designed for robustness, optimality, and clarity. It integrates closely with C and is intended to eventually take the place of C. It uses LLVM to produce highly optimized native code and to cross-compile for any target out of the box. Zig is in alpha; with a beta release expected in September.

LDC - the LLVM-based D compiler

D is a language with C-like syntax and static typing. It pragmatically combines efficiency, control, and modeling power, with safety and programmer productivity. D supports powerful concepts like Compile-Time Function Execution (CTFE) and Template Meta-Programming, provides an innovative approach to concurrency and offers many classical paradigms.

LDC uses the frontend from the reference compiler combined with LLVM as backend to produce efficient native code. LDC targets x86/x86_64 systems like Linux, OS X, FreeBSD and Windows and also Linux on ARM and PowerPC (32/64 bit). Ports to other architectures like AArch64 and MIPS64 are underway.

2.18.4 Additional Information

A wide variety of additional information is available on the [LLVM web page](#), in particular in the [documentation](#) section. The web page also contains versions of the API documentation which is up-to-date with the Subversion version of the source code. You can access versions of these documents specific to this release by going into the `llvm/docs/` directory in the LLVM tree.

If you have any questions or comments about LLVM, please feel free to contact us via the [mailing lists](#).

2.19 LLVM's Analysis and Transform Passes

- *Introduction*
- *Analysis Passes*
 - *-aa-eval: Exhaustive Alias Analysis Precision Evaluator*
 - *-basicaa: Basic Alias Analysis (stateless AA impl)*
 - *-basiccg: Basic CallGraph Construction*
 - *-count-aa: Count Alias Analysis Query Responses*
 - *-da: Dependence Analysis*
 - *-debug-aa: AA use debugger*
 - *-domfrontier: Dominance Frontier Construction*
 - *-domtree: Dominator Tree Construction*
 - *-dot-callgraph: Print Call Graph to “dot” file*
 - *-dot-cfg: Print CFG of function to “dot” file*
 - *-dot-cfg-only: Print CFG of function to “dot” file (with no function bodies)*

- `-dot-dom`: *Print dominance tree of function to “dot” file*
- `-dot-dom-only`: *Print dominance tree of function to “dot” file (with no function bodies)*
- `-dot-postdom`: *Print postdominance tree of function to “dot” file*
- `-dot-postdom-only`: *Print postdominance tree of function to “dot” file (with no function bodies)*
- `-globalsmodref-aa`: *Simple mod/ref analysis for globals*
- `-instcount`: *Counts the various types of Instructions*
- `-intervals`: *Interval Partition Construction*
- `-iv-users`: *Induction Variable Users*
- `-lazy-value-info`: *Lazy Value Information Analysis*
- `-libcall-aa`: *LibCall Alias Analysis*
- `-lint`: *Statically lint-checks LLVM IR*
- `-loops`: *Natural Loop Information*
- `-memdep`: *Memory Dependence Analysis*
- `-module-debuginfo`: *Decodes module-level debug info*
- `-postdomfrontier`: *Post-Dominance Frontier Construction*
- `-postdomtree`: *Post-Dominator Tree Construction*
- `-print-alias-sets`: *Alias Set Printer*
- `-print-callgraph`: *Print a call graph*
- `-print-callgraph-sccs`: *Print SCCs of the Call Graph*
- `-print-cfg-sccs`: *Print SCCs of each function CFG*
- `-print-dom-info`: *Dominator Info Printer*
- `-print-externalfnconstants`: *Print external fn callsites passed constants*
- `-print-function`: *Print function to stderr*
- `-print-module`: *Print module to stderr*
- `-print-used-types`: *Find Used Types*
- `-regions`: *Detect single entry single exit regions*
- `-scalar-evolution`: *Scalar Evolution Analysis*
- `-scev-aa`: *ScalarEvolution-based Alias Analysis*
- `-targetdata`: *Target Data Layout*
- *Transform Passes*
 - `-adce`: *Aggressive Dead Code Elimination*
 - `-always-inline`: *Inliner for always_inline functions*
 - `-argpromotion`: *Promote ‘by reference’ arguments to scalars*
 - `-bb-vectorize`: *Basic-Block Vectorization*
 - `-block-placement`: *Profile Guided Basic Block Placement*

- `-break-crit-edges`: *Break critical edges in CFG*
- `-codegenprepare`: *Optimize for code generation*
- `-constmerge`: *Merge Duplicate Global Constants*
- `-constprop`: *Simple constant propagation*
- `-dce`: *Dead Code Elimination*
- `-deadargelim`: *Dead Argument Elimination*
- `-deadtypeelim`: *Dead Type Elimination*
- `-die`: *Dead Instruction Elimination*
- `-dse`: *Dead Store Elimination*
- `-functionattrs`: *Deduce function attributes*
- `-globaldce`: *Dead Global Elimination*
- `-globalopt`: *Global Variable Optimizer*
- `-gvn`: *Global Value Numbering*
- `-indvars`: *Canonicalize Induction Variables*
- `-inline`: *Function Integration/Inlining*
- `-instcombine`: *Combine redundant instructions*
- `-internalize`: *Internalize Global Symbols*
- `-ipconstprop`: *Interprocedural constant propagation*
- `-ipsccp`: *Interprocedural Sparse Conditional Constant Propagation*
- `-jump-threading`: *Jump Threading*
- `-lcssa`: *Loop-Closed SSA Form Pass*
- `-licm`: *Loop Invariant Code Motion*
- `-loop-deletion`: *Delete dead loops*
- `-loop-extract`: *Extract loops into new functions*
- `-loop-extract-single`: *Extract at most one loop into a new function*
- `-loop-reduce`: *Loop Strength Reduction*
- `-loop-rotate`: *Rotate Loops*
- `-loop-simplify`: *Canonicalize natural loops*
- `-loop-unroll`: *Unroll loops*
- `-loop-unswitch`: *Unswitch loops*
- `-loweratomic`: *Lower atomic intrinsics to non-atomic form*
- `-lowerinvoke`: *Lower invokes to calls, for unwindless code generators*
- `-lowerswitch`: *Lower SwitchInsts to branches*
- `-mem2reg`: *Promote Memory to Register*
- `-memcpyopt`: *MemCpy Optimization*

- *-mergefunc: Merge Functions*
- *-mergereturn: Unify function exit nodes*
- *-partial-inliner: Partial Inliner*
- *-prune-eh: Remove unused exception handling info*
- *-reassociate: Reassociate expressions*
- *-reg2mem: Demote all values to stack slots*
- *-sroa: Scalar Replacement of Aggregates*
- *-sccp: Sparse Conditional Constant Propagation*
- *-simplifycfg: Simplify the CFG*
- *-sink: Code sinking*
- *-strip: Strip all symbols from a module*
- *-strip-dead-debug-info: Strip debug info for unused symbols*
- *-strip-dead-prototypes: Strip Unused Function Prototypes*
- *-strip-debug-declare: Strip all `llvm.dbg.declare` intrinsics*
- *-strip-nondebug: Strip all symbols, except `dbg` symbols, from a module*
- *-tailcallelim: Tail Call Elimination*
- *Utility Passes*
 - *-deadarghaX0r: Dead Argument Hacking (BUGPOINT USE ONLY; DO NOT USE)*
 - *-extract-blocks: Extract Basic Blocks From Module (for bugpoint use)*
 - *-instnamer: Assign names to anonymous instructions*
 - *-verify: Module Verifier*
 - *-view-cfg: View CFG of function*
 - *-view-cfg-only: View CFG of function (with no function bodies)*
 - *-view-dom: View dominance tree of function*
 - *-view-dom-only: View dominance tree of function (with no function bodies)*
 - *-view-postdom: View postdominance tree of function*
 - *-view-postdom-only: View postdominance tree of function (with no function bodies)*

2.19.1 Introduction

This document serves as a high level summary of the optimization features that LLVM provides. Optimizations are implemented as Passes that traverse some portion of a program to either collect information or transform the program. The table below divides the passes that LLVM provides into three categories. Analysis passes compute information that other passes can use or for debugging or program visualization purposes. Transform passes can use (or invalidate) the analysis passes. Transform passes all mutate the program in some way. Utility passes provides some utility but don't otherwise fit categorization. For example passes to extract functions to bitcode or write a module to bitcode are neither analysis nor transform passes. The table of contents above provides a quick summary of each pass and links to the more complete pass description later in the document.

2.19.2 Analysis Passes

This section describes the LLVM Analysis Passes.

-aa-eval1: Exhaustive Alias Analysis Precision Evaluator

This is a simple N^2 alias analysis accuracy evaluator. Basically, for each function in the program, it simply queries to see how the alias analysis implementation answers alias queries between each pair of pointers in the function.

This is inspired and adapted from code by: Naveen Neelakantam, Francesco Spadini, and Wojciech Stryjewski.

-basicaa: Basic Alias Analysis (stateless AA impl)

A basic alias analysis pass that implements identities (two different globals cannot alias, etc), but does no stateful analysis.

-basiccg: Basic CallGraph Construction

Yet to be written.

-count-aa: Count Alias Analysis Query Responses

A pass which can be used to count how many alias queries are being made and how the alias analysis implementation being used responds.

-da: Dependence Analysis

Dependence analysis framework, which is used to detect dependences in memory accesses.

-debug-aa: AA use debugger

This simple pass checks alias analysis users to ensure that if they create a new value, they do not query AA without informing it of the value. It acts as a shim over any other AA pass you want.

Yes keeping track of every value in the program is expensive, but this is a debugging pass.

-domfrontier: Dominance Frontier Construction

This pass is a simple dominator construction algorithm for finding forward dominator frontiers.

-domtree: Dominator Tree Construction

This pass is a simple dominator construction algorithm for finding forward dominators.

-dot-callgraph: Print Call Graph to “dot” file

This pass, only available in `opt`, prints the call graph into a `.dot` graph. This graph can then be processed with the “dot” tool to convert it to postscript or some other suitable format.

-dot-cfg: Print CFG of function to “dot” file

This pass, only available in `opt`, prints the control flow graph into a `.dot` graph. This graph can then be processed with the `dot` tool to convert it to postscript or some other suitable format.

-dot-cfg-only: Print CFG of function to “dot” file (with no function bodies)

This pass, only available in `opt`, prints the control flow graph into a `.dot` graph, omitting the function bodies. This graph can then be processed with the `dot` tool to convert it to postscript or some other suitable format.

-dot-dom: Print dominance tree of function to “dot” file

This pass, only available in `opt`, prints the dominator tree into a `.dot` graph. This graph can then be processed with the `dot` tool to convert it to postscript or some other suitable format.

-dot-dom-only: Print dominance tree of function to “dot” file (with no function bodies)

This pass, only available in `opt`, prints the dominator tree into a `.dot` graph, omitting the function bodies. This graph can then be processed with the `dot` tool to convert it to postscript or some other suitable format.

-dot-postdom: Print postdominance tree of function to “dot” file

This pass, only available in `opt`, prints the post dominator tree into a `.dot` graph. This graph can then be processed with the `dot` tool to convert it to postscript or some other suitable format.

-dot-postdom-only: Print postdominance tree of function to “dot” file (with no function bodies)

This pass, only available in `opt`, prints the post dominator tree into a `.dot` graph, omitting the function bodies. This graph can then be processed with the `dot` tool to convert it to postscript or some other suitable format.

-globalsmodref-aa: Simple mod/ref analysis for globals

This simple pass provides alias and mod/ref information for global values that do not have their address taken, and keeps track of whether functions read or write memory (are “pure”). For this simple (but very common) case, we can provide pretty accurate and useful information.

-instcount: Counts the various types of Instructions

This pass collects the count of all instructions and reports them.

-intervals: Interval Partition Construction

This analysis calculates and represents the interval partition of a function, or a preexisting interval partition.

In this way, the interval partition may be used to reduce a flow graph down to its degenerate single node interval partition (unless it is irreducible).

-iv-users: Induction Variable Users

Bookkeeping for “interesting” users of expressions computed from induction variables.

-lazy-value-info: Lazy Value Information Analysis

Interface for lazy computation of value constraint information.

-libcall-aa: LibCall Alias Analysis

LibCall Alias Analysis.

-lint: Statically lint-checks LLVM IR

This pass statically checks for common and easily-identified constructs which produce undefined or likely unintended behavior in LLVM IR.

It is not a guarantee of correctness, in two ways. First, it isn’t comprehensive. There are checks which could be done statically which are not yet implemented. Some of these are indicated by TODO comments, but those aren’t comprehensive either. Second, many conditions cannot be checked statically. This pass does no dynamic instrumentation, so it can’t check for all possible problems.

Another limitation is that it assumes all code will be executed. A store through a null pointer in a basic block which is never reached is harmless, but this pass will warn about it anyway.

Optimization passes may make conditions that this pass checks for more or less obvious. If an optimization pass appears to be introducing a warning, it may be that the optimization pass is merely exposing an existing condition in the code.

This code may be run before *instcombine*. In many cases, *instcombine* checks for the same kinds of things and turns instructions with undefined behavior into unreachable (or equivalent). Because of this, this pass makes some effort to look through bitcasts and so on.

-loops: Natural Loop Information

This analysis is used to identify natural loops and determine the loop depth of various nodes of the CFG. Note that the loops identified may actually be several natural loops that share the same header node... not just a single natural loop.

-memdep: Memory Dependence Analysis

An analysis that determines, for a given memory operation, what preceding memory operations it depends on. It builds on alias analysis information, and tries to provide a lazy, caching interface to a common kind of alias information query.

-module-debuginfo: Decodes module-level debug info

This pass decodes the debug info metadata in a module and prints in a (sufficiently-prepared-) human-readable form.

For example, run this pass from `opt` along with the `-analyze` option, and it’ll print to standard output.

-postdomfrontier: Post-Dominance Frontier Construction

This pass is a simple post-dominator construction algorithm for finding post-dominator frontiers.

-postdomtree: Post-Dominator Tree Construction

This pass is a simple post-dominator construction algorithm for finding post-dominators.

-print-alias-sets: Alias Set Printer

Yet to be written.

-print-callgraph: Print a call graph

This pass, only available in `opt`, prints the call graph to standard error in a human-readable form.

-print-callgraph-sccs: Print SCCs of the Call Graph

This pass, only available in `opt`, prints the SCCs of the call graph to standard error in a human-readable form.

-print-cfg-sccs: Print SCCs of each function CFG

This pass, only available in `opt`, prints the SCCs of each function CFG to standard error in a human-readable form.

-print-dom-info: Dominator Info Printer

Dominator Info Printer.

-print-externalfnconstants: Print external fn callsites passed constants

This pass, only available in `opt`, prints out call sites to external functions that are called with constant arguments. This can be useful when looking for standard library functions we should constant fold or handle in alias analyses.

-print-function: Print function to stderr

The `PrintFunctionPass` class is designed to be pipelined with other `FunctionPasses`, and prints out the functions of the module as they are processed.

-print-module: Print module to stderr

This pass simply prints out the entire module when it is executed.

-print-used-types: Find Used Types

This pass is used to seek out all of the types in use by the program. Note that this analysis explicitly does not include types only used by the symbol table.

-regions: Detect single entry single exit regions

The `RegionInfo` pass detects single entry single exit regions in a function, where a region is defined as any subgraph that is connected to the remaining graph at only two spots. Furthermore, an hierarchical region tree is built.

-scalar-evolution: Scalar Evolution Analysis

The `ScalarEvolution` analysis can be used to analyze and categorize scalar expressions in loops. It specializes in recognizing general induction variables, representing them with the abstract and opaque `SCEV` class. Given this analysis, trip counts of loops and other important properties can be obtained.

This analysis is primarily useful for induction variable substitution and strength reduction.

-scev-aa: ScalarEvolution-based Alias Analysis

Simple alias analysis implemented in terms of `ScalarEvolution` queries.

This differs from traditional loop dependence analysis in that it tests for dependencies within a single iteration of a loop, rather than dependencies between different iterations.

`ScalarEvolution` has a more complete understanding of pointer arithmetic than `BasicAliasAnalysis`' collection of ad-hoc analyses.

-targetdata: Target Data Layout

Provides other passes access to information on how the size and alignment required by the target ABI for various data types.

2.19.3 Transform Passes

This section describes the LLVM Transform Passes.

-adce: Aggressive Dead Code Elimination

ADCE aggressively tries to eliminate code. This pass is similar to *DCE* but it assumes that values are dead until proven otherwise. This is similar to *SCCP*, except applied to the liveness of values.

-always-inline: Inliner for `always_inline` functions

A custom inliner that handles only functions that are marked as “always inline”.

-argpromotion: Promote ‘by reference’ arguments to scalars

This pass promotes “by reference” arguments to be “by value” arguments. In practice, this means looking for internal functions that have pointer arguments. If it can prove, through the use of alias analysis, that an argument is *only* loaded, then it can pass the value into the function instead of the address of the value. This can cause recursive simplification of code and lead to the elimination of allocas (especially in C++ template code like the STL).

This pass also handles aggregate arguments that are passed into a function, scalarizing them if the elements of the aggregate are only loaded. Note that it refuses to scalarize aggregates which would require passing in more than three operands to the function, because passing thousands of operands for a large array or structure is unprofitable!

Note that this transformation could also be done for arguments that are only stored to (returning the value instead), but does not currently. This case would be best handled when and if LLVM starts supporting multiple return values from functions.

-bb-vectorize: Basic-Block Vectorization

This pass combines instructions inside basic blocks to form vector instructions. It iterates over each basic block, attempting to pair compatible instructions, repeating this process until no additional pairs are selected for vectorization. When the outputs of some pair of compatible instructions are used as inputs by some other pair of compatible instructions, those pairs are part of a potential vectorization chain. Instruction pairs are only fused into vector instructions when they are part of a chain longer than some threshold length. Moreover, the pass attempts to find the best possible chain for each pair of compatible instructions. These heuristics are intended to prevent vectorization in cases where it would not yield a performance increase of the resulting code.

-block-placement: Profile Guided Basic Block Placement

This pass is a very simple profile guided basic block placement algorithm. The idea is to put frequently executed blocks together at the start of the function and hopefully increase the number of fall-through conditional branches. If there is no profile information for a particular function, this pass basically orders blocks in depth-first order.

-break-crit-edges: Break critical edges in CFG

Break all of the critical edges in the CFG by inserting a dummy basic block. It may be “required” by passes that cannot deal with critical edges. This transformation obviously invalidates the CFG, but can update forward dominator (set, immediate dominators, tree, and frontier) information.

-codegenprepare: Optimize for code generation

This pass munges the code in the input function to better prepare it for SelectionDAG-based code generation. This works around limitations in its basic-block-at-a-time approach. It should eventually be removed.

-constmerge: Merge Duplicate Global Constants

Merges duplicate global constants together into a single constant that is shared. This is useful because some passes (i.e., TraceValues) insert a lot of string constants into the program, regardless of whether or not an existing string is available.

-constprop: Simple constant propagation

This pass implements constant propagation and merging. It looks for instructions involving only constant operands and replaces them with a constant value instead of an instruction. For example:

```
add i32 1, 2
```

becomes

```
i32 3
```

NOTE: this pass has a habit of making definitions be dead. It is a good idea to run a *Dead Instruction Elimination* pass sometime after running this pass.

-dce: Dead Code Elimination

Dead code elimination is similar to *dead instruction elimination*, but it rechecks instructions that were used by removed instructions to see if they are newly dead.

-deadargelim: Dead Argument Elimination

This pass deletes dead arguments from internal functions. Dead argument elimination removes arguments which are directly dead, as well as arguments only passed into function calls as dead arguments of other functions. This pass also deletes dead arguments in a similar way.

This pass is often useful as a cleanup pass to run after aggressive interprocedural passes, which add possibly-dead arguments.

-deadtypeelim: Dead Type Elimination

This pass is used to cleanup the output of GCC. It eliminate names for types that are unused in the entire translation unit, using the *find used types* pass.

-die: Dead Instruction Elimination

Dead instruction elimination performs a single pass over the function, removing instructions that are obviously dead.

-dse: Dead Store Elimination

A trivial dead store elimination that only considers basic-block local redundant stores.

-functionattrs: Deduce function attributes

A simple interprocedural pass which walks the call-graph, looking for functions which do not access or only read non-local memory, and marking them `readnone/readonly`. In addition, it marks function arguments (of pointer type) “`nocapture`” if a call to the function does not create any copies of the pointer value that outlive the call. This more or less means that the pointer is only dereferenced, and not returned from the function or stored in a global. This pass is implemented as a bottom-up traversal of the call-graph.

-globaldce: Dead Global Elimination

This transform is designed to eliminate unreachable internal globals from the program. It uses an aggressive algorithm, searching out globals that are known to be alive. After it finds all of the globals which are needed, it deletes whatever is left over. This allows it to delete recursive chunks of the program which are unreachable.

-globalopt: Global Variable Optimizer

This pass transforms simple global variables that never have their address taken. If obviously true, it marks read/write globals as constant, deletes variables only stored to, etc.

-gvn: Global Value Numbering

This pass performs global value numbering to eliminate fully and partially redundant instructions. It also performs redundant load elimination.

-indvars: Canonicalize Induction Variables

This transformation analyzes and transforms the induction variables (and computations derived from them) into simpler forms suitable for subsequent analysis and transformation.

This transformation makes the following changes to each loop with an identifiable induction variable:

- All loops are transformed to have a *single* canonical induction variable which starts at zero and steps by one.
- The canonical induction variable is guaranteed to be the first PHI node in the loop header block.
- Any pointer arithmetic recurrences are raised to use array subscripts.

If the trip count of a loop is computable, this pass also makes the following changes:

- The exit condition for the loop is canonicalized to compare the induction value against the exit value. This turns loops like:

```
for (i = 7; i*i < 1000; ++i)
into
```

```
for (i = 0; i != 25; ++i)
```

- Any use outside of the loop of an expression derived from the indvar is changed to compute the derived value outside of the loop, eliminating the dependence on the exit value of the induction variable. If the only purpose of the loop is to compute the exit value of some derived expression, this transformation will make the loop dead.

This transformation should be followed by strength reduction after all of the desired loop transformations have been performed. Additionally, on targets where it is profitable, the loop could be transformed to count down to zero (the “do loop” optimization).

-inline: Function Integration/Inlining

Bottom-up inlining of functions into callees.

-instcombine: Combine redundant instructions

Combine instructions to form fewer, simple instructions. This pass does not modify the CFG. This pass is where algebraic simplification happens.

This pass combines things like:

```
%Y = add i32 %X, 1
%Z = add i32 %Y, 1
```

into:

```
%Z = add i32 %X, 2
```

This is a simple worklist driven algorithm.

This pass guarantees that the following canonicalizations are performed on the program:

1. If a binary operator has a constant operand, it is moved to the right-hand side.
2. Bitwise operators with constant operands are always grouped so that shifts are performed first, then `ors`, then `ands`, then `xors`.

3. Compare instructions are converted from `<`, `>`, `,` or `to` to `=` or if possible.
4. All `cmp` instructions on boolean values are replaced with logical operations.
5. `add X, X` is represented as `mul X, 2 shl X, 1`
6. Multiplies with a constant power-of-two argument are transformed into shifts.
7. ... etc.

This pass can also simplify calls to specific well-known function calls (e.g. runtime library functions). For example, a call `exit(3)` that occurs within the `main()` function can be transformed into simply `return 3`. Whether or not library calls are simplified is controlled by the `-functionattrs` pass and LLVM's knowledge of library calls on different targets.

`-internalize`: Internalize Global Symbols

This pass loops over all of the functions in the input module, looking for a main function. If a main function is found, all other functions and all global variables with initializers are marked as internal.

`-ipconstprop`: Interprocedural constant propagation

This pass implements an *extremely* simple interprocedural constant propagation pass. It could certainly be improved in many different ways, like using a worklist. This pass makes arguments dead, but does not remove them. The existing dead argument elimination pass should be run after this to clean up the mess.

`-ipsccp`: Interprocedural Sparse Conditional Constant Propagation

An interprocedural variant of *Sparse Conditional Constant Propagation*.

`-jump-threading`: Jump Threading

Jump threading tries to find distinct threads of control flow running through a basic block. This pass looks at blocks that have multiple predecessors and multiple successors. If one or more of the predecessors of the block can be proven to always cause a jump to one of the successors, we forward the edge from the predecessor to the successor by duplicating the contents of this block.

An example of when this can occur is code like this:

```
if () { ...
  X = 4;
}
if (X < 3) {
```

In this case, the unconditional branch at the end of the first if can be revectorized to the false side of the second if.

`-lcssa`: Loop-Closed SSA Form Pass

This pass transforms loops by placing phi nodes at the end of the loops for all values that are live across the loop boundary. For example, it turns the left into the right code:

```
for (...)                for (...)
  if (c)                  if (c)
    X1 = ...              X1 = ...
  else                    else
    X2 = ...              X2 = ...
    X3 = phi(X1, X2)      X3 = phi(X1, X2)
... = X3 + 4              X4 = phi(X3)
                          ... = X4 + 4
```

This is still valid LLVM; the extra phi nodes are purely redundant, and will be trivially eliminated by `InstCombine`. The major benefit of this transformation is that it makes many other loop optimizations, such as `LoopUnswitching`, simpler.

-licm: Loop Invariant Code Motion

This pass performs loop invariant code motion, attempting to remove as much code from the body of a loop as possible. It does this by either hoisting code into the preheader block, or by sinking code to the exit blocks if it is safe. This pass also promotes must-aliased memory locations in the loop to live in registers, thus hoisting and sinking “invariant” loads and stores.

This pass uses alias analysis for two purposes:

1. Moving loop invariant loads and calls out of loops. If we can determine that a load or call inside of a loop never aliases anything stored to, we can hoist it or sink it like any other instruction.
2. Scalar Promotion of Memory. If there is a store instruction inside of the loop, we try to move the store to happen AFTER the loop instead of inside of the loop. This can only happen if a few conditions are true:
 - (a) The pointer stored through is loop invariant.
 - (b) There are no stores or loads in the loop which *may* alias the pointer. There are no calls in the loop which mod/ref the pointer.

If these conditions are true, we can promote the loads and stores in the loop of the pointer to use a temporary alloca'd variable. We then use the `mem2reg` functionality to construct the appropriate SSA form for the variable.

-loop-deletion: Delete dead loops

This file implements the Dead Loop Deletion Pass. This pass is responsible for eliminating loops with non-infinite computable trip counts that have no side effects or volatile instructions, and do not contribute to the computation of the function's return value.

-loop-extract: Extract loops into new functions

A pass wrapper around the `ExtractLoop()` scalar transformation to extract each top-level loop into its own new function. If the loop is the *only* loop in a given function, it is not touched. This is a pass most useful for debugging via `bugpoint`.

-loop-extract-single: Extract at most one loop into a new function

Similar to *Extract loops into new functions*, this pass extracts one natural loop from the program into a function if it can. This is used by `bugpoint`.

-loop-reduce: Loop Strength Reduction

This pass performs a strength reduction on array references inside loops that have as one or more of their components the loop induction variable. This is accomplished by creating a new value to hold the initial value of the array access for the first iteration, and then creating a new GEP instruction in the loop to increment the value by the appropriate amount.

-loop-rotate: Rotate Loops

A simple loop rotation transformation.

-loop-simplify: Canonicalize natural loops

This pass performs several transformations to transform natural loops into a simpler form, which makes subsequent analyses and transformations simpler and more effective.

Loop pre-header insertion guarantees that there is a single, non-critical entry edge from outside of the loop to the loop header. This simplifies a number of analyses and transformations, such as *LICM*.

Loop exit-block insertion guarantees that all exit blocks from the loop (blocks which are outside of the loop that have predecessors inside of the loop) only have predecessors from inside of the loop (and are thus dominated by the loop header). This simplifies transformations such as store-sinking that are built into *LICM*.

This pass also guarantees that loops will have exactly one backedge.

Note that the *simplifycfg* pass will clean up blocks which are split out but end up being unnecessary, so usage of this pass should not pessimize generated code.

This pass obviously modifies the CFG, but updates loop information and dominator information.

-loop-unroll: Unroll loops

This pass implements a simple loop unroller. It works best when loops have been canonicalized by the *indvars* pass, allowing it to determine the trip counts of loops easily.

-loop-unswitch: Unswitch loops

This pass transforms loops that contain branches on loop-invariant conditions to have multiple loops. For example, it turns the left into the right code:

<pre> for (...) A if (lic) B C </pre>	<pre> if (lic) for (...) A; B; C else for (...) A; C </pre>
---	---

This can increase the size of the code exponentially (doubling it every time a loop is unswitched) so we only unswitch if the resultant code will be smaller than a threshold.

This pass expects *LICM* to be run before it to hoist invariant conditions out of the loop, to make the unswitching opportunity obvious.

-loweratomic: Lower atomic intrinsics to non-atomic form

This pass lowers atomic intrinsics to non-atomic form for use in a known non-preemptible environment.

The pass does not verify that the environment is non-preemptible (in general this would require knowledge of the entire call graph of the program including any libraries which may not be available in bitcode form); it simply lowers every atomic intrinsic.

-lowerinvoke: Lower invokes to calls, for unwindless code generators

This transformation is designed for use by code generators which do not yet support stack unwinding. This pass converts `invoke` instructions to `call` instructions, so that any exception-handling `landingpad` blocks become dead code (which can be removed by running the `-simplifycfg` pass afterwards).

-lowerswitch: Lower SwitchInsts to branches

Rewrites switch instructions with a sequence of branches, which allows targets to get away with not implementing the switch instruction until it is convenient.

-mem2reg: Promote Memory to Register

This file promotes memory references to be register references. It promotes `alloca` instructions which only have loads and stores as uses. An `alloca` is transformed by using dominator frontiers to place phi nodes, then traversing the function in depth-first order to rewrite loads and stores as appropriate. This is just the standard SSA construction algorithm to construct “pruned” SSA form.

-memcpyopt: MemCpy Optimization

This pass performs various transformations related to eliminating `memcpy` calls, or transforming sets of stores into `memset`s.

-mergefunc: Merge Functions

This pass looks for equivalent functions that are mergable and folds them.

Total-ordering is introduced among the functions set: we define comparison that answers for every two functions which of them is greater. It allows to arrange functions into the binary tree.

For every new function we check for equivalent in tree.

If equivalent exists we fold such functions. If both functions are overridable, we move the functionality into a new internal function and leave two overridable thanks to it.

If there is no equivalent, then we add this function to tree.

Lookup routine has $O(\log(n))$ complexity, while whole merging process has complexity of $O(n*\log(n))$.

Read [this](#) article for more details.

-mergereturn: Unify function exit nodes

Ensure that functions have at most one `ret` instruction in them. Additionally, it keeps track of which node is the new exit node of the CFG.

-partial-inliner: Partial Inliner

This pass performs partial inlining, typically by inlining an `if` statement that surrounds the body of the function.

-prune-eh: Remove unused exception handling info

This file implements a simple interprocedural pass which walks the call-graph, turning `invoke` instructions into `call` instructions if and only if the callee cannot throw an exception. It implements this as a bottom-up traversal of the call-graph.

-reassociate: Reassociate expressions

This pass reassociates commutative expressions in an order that is designed to promote better constant propagation, GCSE, *LICM*, PRE, etc.

For example: $4 + (x + 5) \rightarrow x + (4 + 5)$

In the implementation of this algorithm, constants are assigned rank = 0, function arguments are rank = 1, and other values are assigned ranks corresponding to the reverse post order traversal of current function (starting at 2), which effectively gives values in deep loops higher rank than values not in loops.

-reg2mem: Demote all values to stack slots

This file demotes all registers to memory references. It is intended to be the inverse of *mem2reg*. By converting to `load` instructions, the only values live across basic blocks are `alloca` instructions and `load` instructions before `phi` nodes. It is intended that this should make CFG hacking much easier. To make later hacking easier, the entry block is split into two, such that all introduced `alloca` instructions (and nothing else) are in the entry block.

-sroa: Scalar Replacement of Aggregates

The well-known scalar replacement of aggregates transformation. This transform breaks up `alloca` instructions of aggregate type (structure or array) into individual `alloca` instructions for each member if possible. Then, if possible, it transforms the individual `alloca` instructions into nice clean scalar SSA form.

-sccp: Sparse Conditional Constant Propagation

Sparse conditional constant propagation and merging, which can be summarized as:

- Assumes values are constant unless proven otherwise
- Assumes BasicBlocks are dead unless proven otherwise
- Proves values to be constant, and replaces them with constants
- Proves conditional branches to be unconditional

Note that this pass has a habit of making definitions be dead. It is a good idea to run a *DCE* pass sometime after running this pass.

-simplifycfg: Simplify the CFG

Performs dead code elimination and basic block merging. Specifically:

- Removes basic blocks with no predecessors.
- Merges a basic block into its predecessor if there is only one and the predecessor only has one successor.
- Eliminates PHI nodes for basic blocks with a single predecessor.
- Eliminates a basic block that only contains an unconditional branch.

-sink: Code sinking

This pass moves instructions into successor blocks, when possible, so that they aren't executed on paths where their results aren't needed.

-strip: Strip all symbols from a module

Performs code stripping. This transformation can delete:

- names for virtual registers
- symbols for internal globals and functions
- debug information

Note that this transformation makes code much less readable, so it should only be used in situations where the strip utility would be used, such as reducing code size or making it harder to reverse engineer code.

-strip-dead-debug-info: Strip debug info for unused symbols

performs code stripping. this transformation can delete:

- names for virtual registers
- symbols for internal globals and functions
- debug information

note that this transformation makes code much less readable, so it should only be used in situations where the strip utility would be used, such as reducing code size or making it harder to reverse engineer code.

-strip-dead-prototypes: Strip Unused Function Prototypes

This pass loops over all of the functions in the input module, looking for dead declarations and removes them. Dead declarations are declarations of functions for which no implementation is available (i.e., declarations for unused library functions).

-strip-debug-declare: Strip all `llvm.dbg.declare` intrinsics

This pass implements code stripping. Specifically, it can delete:

1. names for virtual registers
2. symbols for internal globals and functions
3. debug information

Note that this transformation makes code much less readable, so it should only be used in situations where the ‘strip’ utility would be used, such as reducing code size or making it harder to reverse engineer code.

-strip-nondebug: Strip all symbols, except dbg symbols, from a module

This pass implements code stripping. Specifically, it can delete:

1. names for virtual registers
2. symbols for internal globals and functions
3. debug information

Note that this transformation makes code much less readable, so it should only be used in situations where the ‘strip’ utility would be used, such as reducing code size or making it harder to reverse engineer code.

-tailcallelim: Tail Call Elimination

This file transforms calls of the current function (self recursion) followed by a return instruction with a branch to the entry of the function, creating a loop. This pass also implements the following extensions to the basic algorithm:

1. Trivial instructions between the call and return do not prevent the transformation from taking place, though currently the analysis cannot support moving any really useful instructions (only dead ones).
2. This pass transforms functions that are prevented from being tail recursive by an associative expression to use an accumulator variable, thus compiling the typical naive factorial or fib implementation into efficient code.
3. TRE is performed if the function returns void, if the return returns the result returned by the call, or if the function returns a run-time constant on all exits from the function. It is possible, though unlikely, that the return returns something else (like constant 0), and can still be TRE’d. It can be TRE’d if *all other* return instructions in the function return the exact same value.
4. If it can prove that callees do not access their caller stack frame, they are marked as eligible for tail call elimination (by the code generator).

2.19.4 Utility Passes

This section describes the LLVM Utility Passes.

-deadarghaX0r: Dead Argument Hacking (BUGPOINT USE ONLY; DO NOT USE)

Same as dead argument elimination, but deletes arguments to functions which are external. This is only for use by *bugpoint*.

-extract-blocks: Extract Basic Blocks From Module (for bugpoint use)

This pass is used by bugpoint to extract all blocks from the module into their own functions.

-instnamer: Assign names to anonymous instructions

This is a little utility pass that gives instructions names, this is mostly useful when diffing the effect of an optimization because deleting an unnamed instruction can change all other instruction numbering, making the diff very noisy.

-verify: Module Verifier

Verifies an LLVM IR code. This is useful to run after an optimization which is undergoing testing. Note that `llvm-as` verifies its input before emitting bitcode, and also that malformed bitcode is likely to make LLVM crash. All language front-ends are therefore encouraged to verify their output before performing optimizing transformations.

1. Both of a binary operator's parameters are of the same type.
2. Verify that the indices of mem access instructions match other operands.
3. Verify that arithmetic and other things are only performed on first-class types. Verify that shifts and logicals only happen on integrals f.e.
4. All of the constants in a switch statement are of the correct type.
5. The code is in valid SSA form.
6. It is illegal to put a label into any other type (like a structure) or to return one.
7. Only phi nodes can be self referential: `%x = add i32 %x, %x` is invalid.
8. PHI nodes must have an entry for each predecessor, with no extras.
9. PHI nodes must be the first thing in a basic block, all grouped together.
10. PHI nodes must have at least one entry.
11. All basic blocks should only end with terminator insts, not contain them.
12. The entry node to a function must not have predecessors.
13. All Instructions must be embedded into a basic block.
14. Functions cannot take a void-typed parameter.
15. Verify that a function's argument list agrees with its declared type.
16. It is illegal to specify a name for a void value.
17. It is illegal to have an internal global value with no initializer.
18. It is illegal to have a `ret` instruction that returns a value that does not agree with the function return value type.
19. Function call argument types match the function prototype.
20. All other things that are tested by asserts spread about the code.

Note that this does not provide full security verification (like Java), but instead just tries to ensure that code is well-formed.

-view-cfg: View CFG of function

Displays the control flow graph using the GraphViz tool.

-view-cfg-only: View CFG of function (with no function bodies)

Displays the control flow graph using the GraphViz tool, but omitting function bodies.

-view-dom: View dominance tree of function

Displays the dominator tree using the GraphViz tool.

-view-dom-only: View dominance tree of function (with no function bodies)

Displays the dominator tree using the GraphViz tool, but omitting function bodies.

-view-postdom: View postdominance tree of function

Displays the post dominator tree using the GraphViz tool.

-view-postdom-only: View postdominance tree of function (with no function bodies)

Displays the post dominator tree using the GraphViz tool, but omitting function bodies.

2.20 YAML I/O

- *Introduction to YAML*
- *Introduction to YAML I/O*
- *Error Handling*
- *Scalars*
 - *Built-in types*
 - *Unique types*
 - *Hex types*
 - *ScalarEnumerationTraits*
 - *BitValue*
 - *Custom Scalar*
 - *Block Scalars*
- *Mappings*
 - *No Normalization*
 - *Normalization*
 - *Default values*
 - *Order of Keys*
 - *Tags*
 - *Validation*
 - *Flow Mapping*
- *Sequence*
 - *Flow Sequence*
 - *Utility Macros*
- *Document List*

- *User Context Data*
- *Output*
- *Input*

2.20.1 Introduction to YAML

YAML is a human readable data serialization language. The full YAML language spec can be read at yaml.org. The simplest form of yaml is just “scalars”, “mappings”, and “sequences”. A scalar is any number or string. The pound/hash symbol (#) begins a comment line. A mapping is a set of key-value pairs where the key ends with a colon. For example:

```
# a mapping
name:      Tom
hat-size:  7
```

A sequence is a list of items where each item starts with a leading dash (‘-’). For example:

```
# a sequence
- x86
- x86_64
- PowerPC
```

You can combine mappings and sequences by indenting. For example a sequence of mappings in which one of the mapping values is itself a sequence:

```
# a sequence of mappings with one key's value being a sequence
- name:      Tom
  cpus:
    - x86
    - x86_64
- name:      Bob
  cpus:
    - x86
- name:      Dan
  cpus:
    - PowerPC
    - x86
```

Sometime sequences are known to be short and the one entry per line is too verbose, so YAML offers an alternate syntax for sequences called a “Flow Sequence” in which you put comma separated sequence elements into square brackets. The above example could then be simplified to :

```
# a sequence of mappings with one key's value being a flow sequence
- name:      Tom
  cpus:      [ x86, x86_64 ]
- name:      Bob
  cpus:      [ x86 ]
- name:      Dan
  cpus:      [ PowerPC, x86 ]
```

2.20.2 Introduction to YAML I/O

The use of indenting makes the YAML easy for a human to read and understand, but having a program read and write YAML involves a lot of tedious details. The YAML I/O library structures and simplifies reading and writing YAML documents.

YAML I/O assumes you have some “native” data structures which you want to be able to dump as YAML and recreate from YAML. The first step is to try writing example YAML for your data structures. You may find after looking at possible YAML representations that a direct mapping of your data structures to YAML is not very readable. Often the fields are not in the order that a human would find readable. Or the same information is replicated in multiple locations, making it hard for a human to write such YAML correctly.

In relational database theory there is a design step called normalization in which you reorganize fields and tables. The same considerations need to go into the design of your YAML encoding. But, you may not want to change your existing native data structures. Therefore, when writing out YAML there may be a normalization step, and when reading YAML there would be a corresponding denormalization step.

YAML I/O uses a non-invasive, traits based design. YAML I/O defines some abstract base templates. You specialize those templates on your data types. For instance, if you have an enumerated type `FooBar` you could specialize `ScalarEnumerationTraits` on that type and define the `enumeration()` method:

```
using llvm::yaml::ScalarEnumerationTraits;
using llvm::yaml::IO;

template <>
struct ScalarEnumerationTraits<FooBar> {
    static void enumeration(IO &io, FooBar &value) {
        ...
    }
};
```

As with all YAML I/O template specializations, the `ScalarEnumerationTraits` is used for both reading and writing YAML. That is, the mapping between in-memory enum values and the YAML string representation is only in one place. This assures that the code for writing and parsing of YAML stays in sync.

To specify a YAML mappings, you define a specialization on `llvm::yaml::MappingTraits`. If your native data structure happens to be a struct that is already normalized, then the specialization is simple. For example:

```
using llvm::yaml::MappingTraits;
using llvm::yaml::IO;

template <>
struct MappingTraits<Person> {
    static void mapping(IO &io, Person &info) {
        io.mapRequired("name", info.name);
        io.mapOptional("hat-size", info.hatSize);
    }
};
```

A YAML sequence is automatically inferred if you data type has `begin()/end()` iterators and a `push_back()` method. Therefore any of the STL containers (such as `std::vector<>`) will automatically translate to YAML sequences.

Once you have defined specializations for your data types, you can programmatically use YAML I/O to write a YAML document:

```
using llvm::yaml::Output;

Person tom;
tom.name = "Tom";
```

```
tom.hatSize = 8;
Person dan;
dan.name = "Dan";
dan.hatSize = 7;
std::vector<Person> persons;
persons.push_back(tom);
persons.push_back(dan);

Output yout(llvm::outs());
yout << persons;
```

This would write the following:

```
- name:      Tom
  hat-size:  8
- name:      Dan
  hat-size:  7
```

And you can also read such YAML documents with the following code:

```
using llvm::yaml::Input;

typedef std::vector<Person> PersonList;
std::vector<PersonList> docs;

Input yin(document.getBuffer());
yin >> docs;

if ( yin.error() )
    return;

// Process read document
for ( PersonList &pl : docs ) {
    for ( Person &person : pl ) {
        cout << "name=" << person.name;
    }
}
```

One other feature of YAML is the ability to define multiple documents in a single file. That is why reading YAML produces a vector of your document type.

2.20.3 Error Handling

When parsing a YAML document, if the input does not match your schema (as expressed in your XxxTraits<> specializations). YAML I/O will print out an error message and your Input object's error() method will return true. For instance the following document:

```
- name:      Tom
  shoe-size: 12
- name:      Dan
  hat-size:  7
```

Has a key (shoe-size) that is not defined in the schema. YAML I/O will automatically generate this error:

```
YAML:2:2: error: unknown key 'shoe-size'
shoe-size:      12
^~~~~~
```

Similar errors are produced for other input not conforming to the schema.

2.20.4 Scalars

YAML scalars are just strings (i.e. not a sequence or mapping). The YAML I/O library provides support for translating between YAML scalars and specific C++ types.

Built-in types

The following types have built-in support in YAML I/O:

- bool
- float
- double
- StringRef
- std::string
- int64_t
- int32_t
- int16_t
- int8_t
- uint64_t
- uint32_t
- uint16_t
- uint8_t

That is, you can use those types in fields of MappingTraits or as element type in sequence. When reading, YAML I/O will validate that the string found is convertible to that type and error out if not.

Unique types

Given that YAML I/O is trait based, the selection of how to convert your data to YAML is based on the type of your data. But in C++ type matching, typedefs do not generate unique type names. That means if you have two typedefs of unsigned int, to YAML I/O both types look exactly like unsigned int. To facilitate make unique type names, YAML I/O provides a macro which is used like a typedef on built-in types, but expands to create a class with conversion operators to and from the base type. For example:

```
LLVM_YAML_STRONG_TYPEDEF(uint32_t, MyFooFlags)
LLVM_YAML_STRONG_TYPEDEF(uint32_t, MyBarFlags)
```

This generates two classes MyFooFlags and MyBarFlags which you can use in your native data structures instead of uint32_t. They are implicitly converted to and from uint32_t. The point of creating these unique types is that you can now specify traits on them to get different YAML conversions.

Hex types

An example use of a unique type is that YAML I/O provides fixed sized unsigned integers that are written with YAML I/O as hexadecimal instead of the decimal format used by the built-in integer types:

- Hex64
- Hex32
- Hex16
- Hex8

You can use `llvm::yaml::Hex32` instead of `uint32_t` and the only different will be that when YAML I/O writes out that type it will be formatted in hexadecimal.

ScalarEnumerationTraits

YAML I/O supports translating between in-memory enumerations and a set of string values in YAML documents. This is done by specializing `ScalarEnumerationTraits<>` on your enumeration type and define a `enumeration()` method. For instance, suppose you had an enumeration of CPUs and a struct with it as a field:

```
enum CPUs {
    cpu_x86_64 = 5,
    cpu_x86    = 7,
    cpu_PowerPC = 8
};

struct Info {
    CPUs    cpu;
    uint32_t flags;
};
```

To support reading and writing of this enumeration, you can define a `ScalarEnumerationTraits` specialization on `CPUs`, which can then be used as a field type:

```
using llvm::yaml::ScalarEnumerationTraits;
using llvm::yaml::MappingTraits;
using llvm::yaml::IO;

template <>
struct ScalarEnumerationTraits<CPUs> {
    static void enumeration(IO &io, CPUs &value) {
        io.enumCase(value, "x86_64", cpu_x86_64);
        io.enumCase(value, "x86",    cpu_x86);
        io.enumCase(value, "PowerPC", cpu_PowerPC);
    }
};

template <>
struct MappingTraits<Info> {
    static void mapping(IO &io, Info &info) {
        io.mapRequired("cpu",    info.cpu);
        io.mapOptional("flags",  info.flags, 0);
    }
};
```

When reading YAML, if the string found does not match any of the strings specified by `enumCase()` methods, an error is automatically generated. When writing YAML, if the value being written does not match any of the values specified by the `enumCase()` methods, a runtime assertion is triggered.

BitValue

Another common data structure in C++ is a field where each bit has a unique meaning. This is often used in a “flags” field. YAML I/O has support for converting such fields to a flow sequence. For instance suppose you had the following bit flags defined:

```
enum {
    flagsPointy = 1
    flagsHollow = 2
    flagsFlat   = 4
    flagsRound  = 8
};

LLVM_YAML_STRONG_TYPEDEF(uint32_t, MyFlags)
```

To support reading and writing of `MyFlags`, you specialize `ScalarBitSetTraits<>` on `MyFlags` and provide the bit values and their names.

```
using llvm::yaml::ScalarBitSetTraits;
using llvm::yaml::MappingTraits;
using llvm::yaml::IO;

template <>
struct ScalarBitSetTraits<MyFlags> {
    static void bitset(IO &io, MyFlags &value) {
        io.bitSetCase(value, "hollow", flagHollow);
        io.bitSetCase(value, "flat",   flagFlat);
        io.bitSetCase(value, "round",  flagRound);
        io.bitSetCase(value, "pointy", flagPointy);
    }
};

struct Info {
   StringRef  name;
    MyFlags   flags;
};

template <>
struct MappingTraits<Info> {
    static void mapping(IO &io, Info& info) {
        io.mapRequired("name", info.name);
        io.mapRequired("flags", info.flags);
    }
};
```

With the above, YAML I/O (when writing) will test mask each value in the bitset trait against the flags field, and each that matches will cause the corresponding string to be added to the flow sequence. The opposite is done when reading and any unknown string values will result in a error. With the above schema, a same valid YAML document is:

```
name:    Tom
flags:   [ pointy, flat ]
```

Sometimes a “flags” field might contains an enumeration part defined by a bit-mask.

```
enum {
    flagsFeatureA = 1,
    flagsFeatureB = 2,
    flagsFeatureC = 4,

    flagsCPUMask = 24,

    flagsCPU1 = 8,
    flagsCPU2 = 16
};
```

To support reading and writing such fields, you need to use the `maskedBitSet()` method and provide the bit values, their names and the enumeration mask.

```
template <>
struct ScalarBitSetTraits<MyFlags> {
    static void bitset(IO &io, MyFlags &value) {
        io.bitSetCase(value, "featureA", flagsFeatureA);
        io.bitSetCase(value, "featureB", flagsFeatureB);
        io.bitSetCase(value, "featureC", flagsFeatureC);
        io.maskedBitSetCase(value, "CPU1", flagsCPU1, flagsCPUMask);
        io.maskedBitSetCase(value, "CPU2", flagsCPU2, flagsCPUMask);
    }
};
```

YAML I/O (when writing) will apply the enumeration mask to the flags field, and compare the result and values from the bitset. As in case of a regular bitset, each that matches will cause the corresponding string to be added to the flow sequence.

Custom Scalar

Sometimes for readability a scalar needs to be formatted in a custom way. For instance your internal data structure may use an integer for time (seconds since some epoch), but in YAML it would be much nicer to express that integer in some time format (e.g. 4-May-2012 10:30pm). YAML I/O has a way to support custom formatting and parsing of scalar types by specializing `ScalarTraits<>` on your data type. When writing, YAML I/O will provide the native type and your specialization must create a temporary `llvm::StringRef`. When reading, YAML I/O will provide an `llvm::StringRef` of scalar and your specialization must convert that to your native data type. An outline of a custom scalar type looks like:

```
using llvm::yaml::ScalarTraits;
using llvm::yaml::IO;

template <>
struct ScalarTraits<MyCustomType> {
    static void output(const MyCustomType &value, void*,
                      llvm::raw_ostream &out) {
        out << value; // do custom formatting here
    }
    static StringRef input(StringRef scalar, void*, MyCustomType &value) {
        // do custom parsing here. Return the empty string on success,
        // or an error message on failure.
        return StringRef();
    }
    // Determine if this scalar needs quotes.
    static bool mustQuote(StringRef) { return true; }
};
```


Block Scalars

YAML block scalars are string literals that are represented in YAML using the literal block notation, just like the example shown below:

```
text: |
  First line
  Second line
```

The YAML I/O library provides support for translating between YAML block scalars and specific C++ types by allowing you to specialize `BlockScalarTraits` on your data type. The library doesn't provide any built-in support for block scalar I/O for types like `std::string` and `llvm::StringRef` as they are already supported by YAML I/O and use the ordinary scalar notation by default.

`BlockScalarTraits` specializations are very similar to the `ScalarTraits` specialization - YAML I/O will provide the native type and your specialization must create a temporary `llvm::StringRef` when writing, and it will also provide an `llvm::StringRef` that has the value of that block scalar and your specialization must convert that to your native data type when reading. An example of a custom type with an appropriate specialization of `BlockScalarTraits` is shown below:

```
using llvm::yaml::BlockScalarTraits;
using llvm::yaml::IO;

struct MyStringType {
    std::string Str;
};

template <>
struct BlockScalarTraits<MyStringType> {
    static void output(const MyStringType &Value, void *Ctxt,
                      llvm::raw_ostream &OS) {
        OS << Value.Str;
    }

    static StringRef input(StringRef Scalar, void *Ctxt,
                          MyStringType &Value) {
        Value.Str = Scalar.str();
        return StringRef();
    }
};
```

2.20.5 Mappings

To be translated to or from a YAML mapping for your type `T` you must specialize `llvm::yaml::MappingTraits` on `T` and implement the “void mapping(IO &io, T&)” method. If your native data structures use pointers to a class everywhere, you can specialize on the class pointer. Examples:

```
using llvm::yaml::MappingTraits;
using llvm::yaml::IO;

// Example of struct Foo which is used by value
template <>
struct MappingTraits<Foo> {
    static void mapping(IO &io, Foo &foo) {
        io.mapOptional("size", foo.size);
        ...
    }
};
```

```
    }  
};  
  
// Example of struct Bar which is natively always a pointer  
template <>  
struct MappingTraits<Bar*> {  
    static void mapping(IO &io, Bar *&bar) {  
        io.mapOptional("size", bar->size);  
        ...  
    }  
};
```

No Normalization

The mapping() method is responsible, if needed, for normalizing and denormalizing. In a simple case where the native data structure requires no normalization, the mapping method just uses mapOptional() or mapRequired() to bind the struct's fields to YAML key names. For example:

```
using llvm::yaml::MappingTraits;  
using llvm::yaml::IO;  
  
template <>  
struct MappingTraits<Person> {  
    static void mapping(IO &io, Person &info) {  
        io.mapRequired("name", info.name);  
        io.mapOptional("hat-size", info.hatSize);  
    }  
};
```

Normalization

When [de]normalization is required, the mapping() method needs a way to access normalized values as fields. To help with this, there is a template MappingNormalization<> which you can then use to automatically do the normalization and denormalization. The template is used to create a local variable in your mapping() method which contains the normalized keys.

Suppose you have native data type Polar which specifies a position in polar coordinates (distance, angle):

```
struct Polar {  
    float distance;  
    float angle;  
};
```

but you've decided the normalized YAML for should be in x,y coordinates. That is, you want the yaml to look like:

```
x: 10.3  
y: -4.7
```

You can support this by defining a MappingTraits that normalizes the polar coordinates to x,y coordinates when writing YAML and denormalizes x,y coordinates into polar when reading YAML.

```
using llvm::yaml::MappingTraits;  
using llvm::yaml::IO;  
  
template <>
```

```

struct MappingTraits<Polar> {

    class NormalizedPolar {
    public:
        NormalizedPolar(IO &io)
            : x(0.0), y(0.0) {
        }
        NormalizedPolar(IO &, Polar &polar)
            : x(polar.distance * cos(polar.angle)),
              y(polar.distance * sin(polar.angle)) {
        }
        Polar denormalize(IO &) {
            return Polar(sqrt(x*x+y*y), atan(x,y));
        }

        float      x;
        float      y;
    };

    static void mapping(IO &io, Polar &polar) {
        MappingNormalization<NormalizedPolar, Polar> keys(io, polar);

        io.mapRequired("x",    keys->x);
        io.mapRequired("y",    keys->y);
    }
};

```

When writing YAML, the local variable “keys” will be a stack allocated instance of `NormalizedPolar`, constructed from the supplied polar object which initializes its `x` and `y` fields. The `mapRequired()` methods then write out the `x` and `y` values as key/value pairs.

When reading YAML, the local variable “keys” will be a stack allocated instance of `NormalizedPolar`, constructed by the empty constructor. The `mapRequired` methods will find the matching key in the YAML document and fill in the `x` and `y` fields of the `NormalizedPolar` object `keys`. At the end of the `mapping()` method when the local `keys` variable goes out of scope, the `denormalize()` method will automatically be called to convert the read values back to polar coordinates, and then assigned back to the second parameter to `mapping()`.

In some cases, the normalized class may be a subclass of the native type and could be returned by the `denormalize()` method, except that the temporary normalized instance is stack allocated. In these cases, the utility template `MappingNormalizationHeap<>` can be used instead. It just like `MappingNormalization<>` except that it heap allocates the normalized object when reading YAML. It never destroys the normalized object. The `denormalize()` method can then return “this”.

Default values

Within a `mapping()` method, calls to `io.mapRequired()` mean that that key is required to exist when parsing YAML documents, otherwise YAML I/O will issue an error.

On the other hand, keys registered with `io.mapOptional()` are allowed to not exist in the YAML document being read. So what value is put in the field for those optional keys? There are two steps to how those optional fields are filled in. First, the second parameter to the `mapping()` method is a reference to a native class. That native class must have a default constructor. Whatever value the default constructor initially sets for an optional field will be that field’s value. Second, the `mapOptional()` method has an optional third parameter. If provided it is the value that `mapOptional()` should set that field to if the YAML document does not have that key.

There is one important difference between those two ways (default constructor and third parameter to `mapOptional()`). When YAML I/O generates a YAML document, if the `mapOptional()` third parameter is used, if the actual value being

written is the same as (using ==) the default value, then that key/value is not written.

Order of Keys

When writing out a YAML document, the keys are written in the order that the calls to `mapRequired()/mapOptional()` are made in the `mapping()` method. This gives you a chance to write the fields in an order that a human reader of the YAML document would find natural. This may be different than the order of the fields in the native class.

When reading in a YAML document, the keys in the document can be in any order, but they are processed in the order that the calls to `mapRequired()/mapOptional()` are made in the `mapping()` method. That enables some interesting functionality. For instance, if the first field bound is the `cpu` and the second field bound is `flags`, and the flags are `cpu` specific, you can programmatically switch how the flags are converted to and from YAML based on the `cpu`. This works for both reading and writing. For example:

```
using llvm::yaml::MappingTraits;
using llvm::yaml::IO;

struct Info {
    CPUs      cpu;
    uint32_t   flags;
};

template <>
struct MappingTraits<Info> {
    static void mapping(IO &io, Info &info) {
        io.mapRequired("cpu",      info.cpu);
        // flags must come after cpu for this to work when reading yaml
        if ( info.cpu == cpu_x86_64 )
            io.mapRequired("flags",  *(My86_64Flags*)info.flags);
        else
            io.mapRequired("flags",  *(My86Flags*)info.flags);
    }
};
```

Tags

The YAML syntax supports tags as a way to specify the type of a node before it is parsed. This allows dynamic types of nodes. But the YAML I/O model uses static typing, so there are limits to how you can use tags with the YAML I/O model. Recently, we added support to YAML I/O for checking/setting the optional tag on a map. Using this functionality it is even possible to support different mappings, as long as they are convertible.

To check a tag, inside your `mapping()` method you can use `io.mapTag()` to specify what the tag should be. This will also add that tag when writing yaml.

Validation

Sometimes in a yaml map, each key/value pair is valid, but the combination is not. This is similar to something having no syntax errors, but still having semantic errors. To support semantic level checking, YAML I/O allows an optional `validate()` method in a `MappingTraits` template specialization.

When parsing yaml, the `validate()` method is called *after* all key/values in the map have been processed. Any error message returned by the `validate()` method during input will be printed just like a syntax error would be printed. When writing yaml, the `validate()` method is called *before* the yaml key/values are written. Any error during output will trigger an `assert()` because it is a programming error to have invalid struct values.

```
using llvm::yaml::MappingTraits;
using llvm::yaml::IO;

struct Stuff {
    ...
};

template <>
struct MappingTraits<Stuff> {
    static void mapping(IO &io, Stuff &stuff) {
        ...
    }
    staticStringRef validate(IO &io, Stuff &stuff) {
        // Look at all fields in 'stuff' and if there
        // are any bad values return a string describing
        // the error. Otherwise return an empty string.
        return StringRef();
    }
};
```

Flow Mapping

A YAML “flow mapping” is a mapping that uses the inline notation (e.g { x: 1, y: 0 }) when written to YAML. To specify that a type should be written in YAML using flow mapping, your MappingTraits specialization should add “static const bool flow = true;”. For instance:

```
using llvm::yaml::MappingTraits;
using llvm::yaml::IO;

struct Stuff {
    ...
};

template <>
struct MappingTraits<Stuff> {
    static void mapping(IO &io, Stuff &stuff) {
        ...
    }

    static const bool flow = true;
}
```

Flow mappings are subject to line wrapping according to the Output object configuration.

2.20.6 Sequence

To be translated to or from a YAML sequence for your type T you must specialize llvm::yaml::SequenceTraits on T and implement two methods: size_t size(IO &io, T&) and T::value_type& element(IO &io, T&, size_t indx). For example:

```
template <>
struct SequenceTraits<MySeq> {
    static size_t size(IO &io, MySeq &list) { ... }
    static MySeqEl &element(IO &io, MySeq &list, size_t index) { ... }
};
```

The `size()` method returns how many elements are currently in your sequence. The `element()` method returns a reference to the *i*'th element in the sequence. When parsing YAML, the `element()` method may be called with an index one bigger than the current size. Your `element()` method should allocate space for one more element (using default constructor if element is a C++ object) and returns a reference to that new allocated space.

Flow Sequence

A YAML “flow sequence” is a sequence that when written to YAML it uses the inline notation (e.g. `[foo, bar]`). To specify that a sequence type should be written in YAML as a flow sequence, your `SequenceTraits` specialization should add “static const bool flow = true;”. For instance:

```
template <>
struct SequenceTraits<MyList> {
    static size_t size(IO &io, MyList &list) { ... }
    static MyListEl &element(IO &io, MyList &list, size_t index) { ... }

    // The existence of this member causes YAML I/O to use a flow sequence
    static const bool flow = true;
};
```

With the above, if you used `MyList` as the data type in your native data structures, then when converted to YAML, a flow sequence of integers will be used (e.g. `[10, -3, 4]`).

Flow sequences are subject to line wrapping according to the Output object configuration.

Utility Macros

Since a common source of sequences is `std::vector<>`, YAML I/O provides macros: `LLVM_YAML_IS_SEQUENCE_VECTOR()` and `LLVM_YAML_IS_FLOW_SEQUENCE_VECTOR()` which can be used to easily specify `SequenceTraits<>` on a `std::vector` type. YAML I/O does not partial specialize `SequenceTraits` on `std::vector<>` because that would force all vectors to be sequences. An example use of the macros:

```
std::vector<MyType1>;
std::vector<MyType2>;
LLVM_YAML_IS_SEQUENCE_VECTOR(MyType1)
LLVM_YAML_IS_FLOW_SEQUENCE_VECTOR(MyType2)
```

2.20.7 Document List

YAML allows you to define multiple “documents” in a single YAML file. Each new document starts with a left aligned “—” token. The end of all documents is denoted with a left aligned “...” token. Many users of YAML will never have need for multiple documents. The top level node in their YAML schema will be a mapping or sequence. For those cases, the following is not needed. But for cases where you do want multiple documents, you can specify a trait for you document list type. The trait has the same methods as `SequenceTraits` but is named `DocumentListTraits`. For example:

```
template <>
struct DocumentListTraits<MyDocList> {
    static size_t size(IO &io, MyDocList &list) { ... }
    static MyDocType element(IO &io, MyDocList &list, size_t index) { ... }
};
```

2.20.8 User Context Data

When an `llvm::yaml::Input` or `llvm::yaml::Output` object is created their constructors take an optional “context” parameter. This is a pointer to whatever state information you might need.

For instance, in a previous example we showed how the conversion type for a flags field could be determined at runtime based on the value of another field in the mapping. But what if an inner mapping needs to know some field value of an outer mapping? That is where the “context” parameter comes in. You can set values in the context in the outer map’s `mapping()` method and retrieve those values in the inner map’s `mapping()` method.

The context value is just a `void*`. All your traits which use the context and operate on your native data types, need to agree what the context value actually is. It could be a pointer to an object or struct which your various traits use to shared context sensitive information.

2.20.9 Output

The `llvm::yaml::Output` class is used to generate a YAML document from your in-memory data structures, using traits defined on your data types. To instantiate an `Output` object you need an `llvm::raw_ostream`, an optional context pointer and an optional wrapping column:

```
class Output : public IO {
public:
    Output(llvm::raw_ostream &, void *context = NULL, int WrapColumn = 70);
```

Once you have an `Output` object, you can use the C++ stream operator on it to write your native data as YAML. One thing to recall is that a YAML file can contain multiple “documents”. If the top level data structure you are streaming as YAML is a mapping, scalar, or sequence, then `Output` assumes you are generating one document and wraps the mapping output with “---” and trailing “...”.

The `WrapColumn` parameter will cause the flow mappings and sequences to line-wrap when they go over the supplied column. Pass 0 to completely suppress the wrapping.

```
using llvm::yaml::Output;

void dumpMyMapDoc(const MyMapType &info) {
    Output yout(llvm::outs());
    yout << info;
}
```

The above could produce output like:

```
---
name:      Tom
hat-size:  7
...
```

On the other hand, if the top level data structure you are streaming as YAML has a `DocumentListTraits` specialization, then `Output` walks through each element of your `DocumentList` and generates a “—” before the start of each element and ends with a “...”.

```
using llvm::yaml::Output;

void dumpMyMapDoc(const MyDocListType &docList) {
    Output yout(llvm::outs());
    yout << docList;
}
```

The above could produce output like:

```
---
name:      Tom
hat-size:  7
---
name:      Tom
shoe-size: 11
...
```

2.20.10 Input

The `llvm::yaml::Input` class is used to parse YAML document(s) into your native data structures. To instantiate an Input object you need a `StringRef` to the entire YAML file, and optionally a context pointer:

```
class Input : public IO {
public:
    Input(StringRef inputContent, void *context=NULL);
```

Once you have an Input object, you can use the C++ stream operator to read the document(s). If you expect there might be multiple YAML documents in one file, you'll need to specialize `DocumentListTraits` on a list of your document type and stream in that document list type. Otherwise you can just stream in the document type. Also, you can check if there was any syntax errors in the YAML by calling the `error()` method on the Input object. For example:

```
// Reading a single document
using llvm::yaml::Input;

Input yin(mb.getBuffer());

// Parse the YAML file
MyDocType theDoc;
yin >> theDoc;

// Check for error
if ( yin.error() )
    return;
```

```
// Reading multiple documents in one file
using llvm::yaml::Input;

LLVM_YAML_IS_DOCUMENT_LIST_VECTOR(std::vector<MyDocType>)

Input yin(mb.getBuffer());

// Parse the YAML file
std::vector<MyDocType> theDocList;
yin >> theDocList;

// Check for error
if ( yin.error() )
    return;
```


2.21 The Often Misunderstood GEP Instruction

- *Introduction*
- *Address Computation*
 - *What is the first index of the GEP instruction?*
 - *Why is the extra 0 index required?*
 - *What is dereferenced by GEP?*
 - *Why don't GEP x,0,0,1 and GEP x,1 alias?*
 - *Why do GEP x,1,0,0 and GEP x,1 alias?*
 - *Can GEP index into vector elements?*
 - *What effect do address spaces have on GEPs?*
 - *How is GEP different from ptrtoint, arithmetic, and inttoptr?*
 - *I'm writing a backend for a target which needs custom lowering for GEP. How do I do this?*
 - *How does VLA addressing work with GEPs?*
- *Rules*
 - *What happens if an array index is out of bounds?*
 - *Can array indices be negative?*
 - *Can I compare two values computed with GEPs?*
 - *Can I do GEP with a different pointer type than the type of the underlying object?*
 - *Can I cast an object's address to integer and add it to null?*
 - *Can I compute the distance between two objects, and add that value to one address to compute the other address?*
 - *Can I do type-based alias analysis on LLVM IR?*
 - *What happens if a GEP computation overflows?*
 - *How can I tell if my front-end is following the rules?*
- *Rationale*
 - *Why is GEP designed this way?*
 - *Why do struct member indices always use i32?*
 - *What's an uglygep?*
- *Summary*

2.21.1 Introduction

This document seeks to dispel the mystery and confusion surrounding LLVM's `GetElementPtr` (GEP) instruction. Questions about the wily GEP instruction are probably the most frequently occurring questions once a developer gets down to coding with LLVM. Here we lay out the sources of confusion and show that the GEP instruction is really quite simple.

2.21.2 Address Computation

When people are first confronted with the GEP instruction, they tend to relate it to known concepts from other programming paradigms, most notably C array indexing and field selection. GEP closely resembles C array indexing and field selection, however it is a little different and this leads to the following questions.

What is the first index of the GEP instruction?

Quick answer: The index stepping through the second operand.

The confusion with the first index usually arises from thinking about the `GetElementPtr` instruction as if it was a C index operator. They aren't the same. For example, when we write, in "C":

```
AType *Foo;
...
X = &Foo->F;
```

it is natural to think that there is only one index, the selection of the field `F`. However, in this example, `Foo` is a pointer. That pointer must be indexed explicitly in LLVM. C, on the other hand, indices through it transparently. To arrive at the same address location as the C code, you would provide the GEP instruction with two index operands. The first operand indexes through the pointer; the second operand indexes the field `F` of the structure, just as if you wrote:

```
X = &Foo[0].F;
```

Sometimes this question gets rephrased as:

Why is it okay to index through the first pointer, but subsequent pointers won't be dereferenced?

The answer is simply because memory does not have to be accessed to perform the computation. The second operand to the GEP instruction must be a value of a pointer type. The value of the pointer is provided directly to the GEP instruction as an operand without any need for accessing memory. It must, therefore be indexed and requires an index operand. Consider this example:

```
struct munger_struct {
    int f1;
    int f2;
};
void munge(struct munger_struct *P) {
    P[0].f1 = P[1].f1 + P[2].f2;
}
...
munger_struct Array[3];
...
munge(Array);
```

In this "C" example, the front end compiler (Clang) will generate three GEP instructions for the three indices through "P" in the assignment statement. The function argument `P` will be the second operand of each of these GEP instructions. The third operand indexes through that pointer. The fourth operand will be the field offset into the `struct munger_struct` type, for either the `f1` or `f2` field. So, in LLVM assembly the `munge` function looks like:

```
void @munge(%struct.munger_struct* %P) {
entry:
    %tmp = getelementptr %struct.munger_struct, %struct.munger_struct* %P, i32 1, i32 0
    %tmp = load i32* %tmp
    %tmp6 = getelementptr %struct.munger_struct, %struct.munger_struct* %P, i32 2, i32 1
    %tmp7 = load i32* %tmp6
    %tmp8 = add i32 %tmp7, %tmp
```

```

    %tmp9 = getelementptr %struct.munger_struct, %struct.munger_struct* %P, i32 0, i32 0
    store i32 %tmp8, i32* %tmp9
    ret void
}

```

In each case the second operand is the pointer through which the GEP instruction starts. The same is true whether the second operand is an argument, allocated memory, or a global variable.

To make this clear, let's consider a more obtuse example:

```

%MyVar = uninitialized global i32
...
%idx1 = getelementptr i32, i32* %MyVar, i64 0
%idx2 = getelementptr i32, i32* %MyVar, i64 1
%idx3 = getelementptr i32, i32* %MyVar, i64 2

```

These GEP instructions are simply making address computations from the base address of `MyVar`. They compute, as follows (using C syntax):

```

idx1 = (char*) &MyVar + 0
idx2 = (char*) &MyVar + 4
idx3 = (char*) &MyVar + 8

```

Since the type `i32` is known to be four bytes long, the indices 0, 1 and 2 translate into memory offsets of 0, 4, and 8, respectively. No memory is accessed to make these computations because the address of `%MyVar` is passed directly to the GEP instructions.

The obtuse part of this example is in the cases of `%idx2` and `%idx3`. They result in the computation of addresses that point to memory past the end of the `%MyVar` global, which is only one `i32` long, not three `i32`s long. While this is legal in LLVM, it is inadvisable because any load or store with the pointer that results from these GEP instructions would produce undefined results.

Why is the extra 0 index required?

Quick answer: there are no superfluous indices.

This question arises most often when the GEP instruction is applied to a global variable which is always a pointer type. For example, consider this:

```

%MyStruct = uninitialized global { float*, i32 }
...
%idx = getelementptr { float*, i32 }, { float*, i32 }* %MyStruct, i64 0, i32 1

```

The GEP above yields an `i32*` by indexing the `i32` typed field of the structure `%MyStruct`. When people first look at it, they wonder why the `i64 0` index is needed. However, a closer inspection of how globals and GEPs work reveals the need. Becoming aware of the following facts will dispel the confusion:

1. The type of `%MyStruct` is *not* `{ float*, i32 }` but rather `{ float*, i32 }*`. That is, `%MyStruct` is a pointer to a structure containing a pointer to a float and an `i32`.
2. Point #1 is evidenced by noticing the type of the second operand of the GEP instruction (`%MyStruct`) which is `{ float*, i32 }*`.
3. The first index, `i64 0` is required to step over the global variable `%MyStruct`. Since the second argument to the GEP instruction must always be a value of pointer type, the first index steps through that pointer. A value of 0 means 0 elements offset from that pointer.
4. The second index, `i32 1` selects the second field of the structure (the `i32`).

What is dereferenced by GEP?

Quick answer: nothing.

The `GetElementPtr` instruction dereferences nothing. That is, it doesn't access memory in any way. That's what the `Load` and `Store` instructions are for. GEP is only involved in the computation of addresses. For example, consider this:

```
%MyVar = uninitialized global { [40 x i32] * }
...
%idx = getelementptr { [40 x i32] * }, { [40 x i32] * } * %MyVar, i64 0, i32 0, i64 0, ↪
↪ i64 17
```

In this example, we have a global variable, `%MyVar` that is a pointer to a structure containing a pointer to an array of 40 ints. The GEP instruction seems to be accessing the 18th integer of the structure's array of ints. However, this is actually an illegal GEP instruction. It won't compile. The reason is that the pointer in the structure *must* be dereferenced in order to index into the array of 40 ints. Since the GEP instruction never accesses memory, it is illegal.

In order to access the 18th integer in the array, you would need to do the following:

```
%idx = getelementptr { [40 x i32] * }, { [40 x i32] * } * %, i64 0, i32 0
%arr = load [40 x i32] ** %idx
%idx = getelementptr [40 x i32], [40 x i32] * %arr, i64 0, i64 17
```

In this case, we have to load the pointer in the structure with a `load` instruction before we can index into the array. If the example was changed to:

```
%MyVar = uninitialized global { [40 x i32] }
...
%idx = getelementptr { [40 x i32] }, { [40 x i32] } *, i64 0, i32 0, i64 17
```

then everything works fine. In this case, the structure does not contain a pointer and the GEP instruction can index through the global variable, into the first field of the structure and access the 18th `i32` in the array there.

Why don't GEP x,0,0,1 and GEP x,1 alias?

Quick Answer: They compute different address locations.

If you look at the first indices in these GEP instructions you find that they are different (0 and 1), therefore the address computation diverges with that index. Consider this example:

```
%MyVar = global { [10 x i32] }
%idx1 = getelementptr { [10 x i32] }, { [10 x i32] } * %MyVar, i64 0, i32 0, i64 1
%idx2 = getelementptr { [10 x i32] }, { [10 x i32] } * %MyVar, i64 1
```

In this example, `idx1` computes the address of the second integer in the array that is in the structure in `%MyVar`, that is `MyVar+4`. The type of `idx1` is `i32*`. However, `idx2` computes the address of *the next* structure after `%MyVar`. The type of `idx2` is `{ [10 x i32] }*` and its value is equivalent to `MyVar + 40` because it indexes past the ten 4-byte integers in `MyVar`. Obviously, in such a situation, the pointers don't alias.

Why do GEP x,1,0,0 and GEP x,1 alias?

Quick Answer: They compute the same address location.

These two GEP instructions will compute the same address because indexing through the 0th element does not change the address. However, it does change the type. Consider this example:

```
%MyVar = global { [10 x i32] }
%idx1 = getelementptr { [10 x i32] }, { [10 x i32] }* %MyVar, i64 1, i32 0, i64 0
%idx2 = getelementptr { [10 x i32] }, { [10 x i32] }* %MyVar, i64 1
```

In this example, the value of %idx1 is %MyVar+40 and its type is i32*. The value of %idx2 is also MyVar+40 but its type is { [10 x i32] }*.

Can GEP index into vector elements?

This hasn't always been forcefully disallowed, though it's not recommended. It leads to awkward special cases in the optimizers, and fundamental inconsistency in the IR. In the future, it will probably be outright disallowed.

What effect do address spaces have on GEPs?

None, except that the address space qualifier on the second operand pointer type always matches the address space qualifier on the result type.

How is GEP different from ptrtoint, arithmetic, and inttoptr?

It's very similar; there are only subtle differences.

With ptrtoint, you have to pick an integer type. One approach is to pick i64; this is safe on everything LLVM supports (LLVM internally assumes pointers are never wider than 64 bits in many places), and the optimizer will actually narrow the i64 arithmetic down to the actual pointer size on targets which don't support 64-bit arithmetic in most cases. However, there are some cases where it doesn't do this. With GEP you can avoid this problem.

Also, GEP carries additional pointer aliasing rules. It's invalid to take a GEP from one object, address into a different separately allocated object, and dereference it. IR producers (front-ends) must follow this rule, and consumers (optimizers, specifically alias analysis) benefit from being able to rely on it. See the [Rules](#) section for more information.

And, GEP is more concise in common cases.

However, for the underlying integer computation implied, there is no difference.

I'm writing a backend for a target which needs custom lowering for GEP. How do I do this?

You don't. The integer computation implied by a GEP is target-independent. Typically what you'll need to do is make your backend pattern-match expressions trees involving ADD, MUL, etc., which are what GEP is lowered into. This has the advantage of letting your code work correctly in more cases.

GEP does use target-dependent parameters for the size and layout of data types, which targets can customize.

If you require support for addressing units which are not 8 bits, you'll need to fix a lot of code in the backend, with GEP lowering being only a small piece of the overall picture.

How does VLA addressing work with GEPs?

GEPs don't natively support VLAs. LLVM's type system is entirely static, and GEP address computations are guided by an LLVM type.

VLA indices can be implemented as linearized indices. For example, an expression like `X[a][b][c]`, must be effectively lowered into a form like `X[a*m+b*n+c]`, so that it appears to the GEP as a single-dimensional array reference.

This means if you want to write an analysis which understands array indices and you want to support VLAs, your code will have to be prepared to reverse-engineer the linearization. One way to solve this problem is to use the `ScalarEvolution` library, which always presents VLA and non-VLA indexing in the same manner.

2.21.3 Rules

What happens if an array index is out of bounds?

There are two senses in which an array index can be out of bounds.

First, there's the array type which comes from the (static) type of the first operand to the GEP. Indices greater than the number of elements in the corresponding static array type are valid. There is no problem with out of bounds indices in this sense. Indexing into an array only depends on the size of the array element, not the number of elements.

A common example of how this is used is arrays where the size is not known. It's common to use array types with zero length to represent these. The fact that the static type says there are zero elements is irrelevant; it's perfectly valid to compute arbitrary element indices, as the computation only depends on the size of the array element, not the number of elements. Note that zero-sized arrays are not a special case here.

This sense is unconnected with `inbounds` keyword. The `inbounds` keyword is designed to describe low-level pointer arithmetic overflow conditions, rather than high-level array indexing rules.

Analysis passes which wish to understand array indexing should not assume that the static array type bounds are respected.

The second sense of being out of bounds is computing an address that's beyond the actual underlying allocated object.

With the `inbounds` keyword, the result value of the GEP is undefined if the address is outside the actual underlying allocated object and not the address one-past-the-end.

Without the `inbounds` keyword, there are no restrictions on computing out-of-bounds addresses. Obviously, performing a load or a store requires an address of allocated and sufficiently aligned memory. But the GEP itself is only concerned with computing addresses.

Can array indices be negative?

Yes. This is basically a special case of array indices being out of bounds.

Can I compare two values computed with GEPs?

Yes. If both addresses are within the same allocated object, or one-past-the-end, you'll get the comparison result you expect. If either is outside of it, integer arithmetic wrapping may occur, so the comparison may not be meaningful.

Can I do GEP with a different pointer type than the type of the underlying object?

Yes. There are no restrictions on bitcasting a pointer value to an arbitrary pointer type. The types in a GEP serve only to define the parameters for the underlying integer computation. They need not correspond with the actual type of the underlying object.

Furthermore, loads and stores don't have to use the same types as the type of the underlying object. Types in this context serve only to specify memory size and alignment. Beyond that there are merely a hint to the optimizer indicating how the value will likely be used.

Can I cast an object's address to integer and add it to null?

You can compute an address that way, but if you use GEP to do the add, you can't use that pointer to actually access the object, unless the object is managed outside of LLVM.

The underlying integer computation is sufficiently defined; null has a defined value — zero — and you can add whatever value you want to it.

However, it's invalid to access (load from or store to) an LLVM-aware object with such a pointer. This includes `GlobalVariables`, `Allocas`, and objects pointed to by noalias pointers.

If you really need this functionality, you can do the arithmetic with explicit integer instructions, and use `inttoptr` to convert the result to an address. Most of GEP's special aliasing rules do not apply to pointers computed from `ptrtoint`, arithmetic, and `inttoptr` sequences.

Can I compute the distance between two objects, and add that value to one address to compute the other address?

As with arithmetic on null, you can use GEP to compute an address that way, but you can't use that pointer to actually access the object if you do, unless the object is managed outside of LLVM.

Also as above, `ptrtoint` and `inttoptr` provide an alternative way to do this which do not have this restriction.

Can I do type-based alias analysis on LLVM IR?

You can't do type-based alias analysis using LLVM's built-in type system, because LLVM has no restrictions on mixing types in addressing, loads or stores.

LLVM's type-based alias analysis pass uses metadata to describe a different type system (such as the C type system), and performs type-based aliasing on top of that. Further details are in the [language reference](#).

What happens if a GEP computation overflows?

If the GEP lacks the `inbounds` keyword, the value is the result from evaluating the implied two's complement integer computation. However, since there's no guarantee of where an object will be allocated in the address space, such values have limited meaning.

If the GEP has the `inbounds` keyword, the result value is undefined (a "trap value") if the GEP overflows (i.e. wraps around the end of the address space).

As such, there are some ramifications of this for inbounds GEPs: scales implied by array/vector/pointer indices are always known to be "nsw" since they are signed values that are scaled by the element size. These values are also allowed to be negative (e.g. `gep i32 *%P, i32 -1`) but the pointer itself is logically treated as an unsigned value. This means that GEPs have an asymmetric relation between the pointer base (which is treated as unsigned) and the offset applied to it (which is treated as signed). The result of the additions within the offset calculation cannot have signed overflow, but when applied to the base pointer, there can be signed overflow.

How can I tell if my front-end is following the rules?

There is currently no checker for the `getelementptr` rules. Currently, the only way to do this is to manually check each place in your front-end where `GetElementPtr` operators are created.

It's not possible to write a checker which could find all rule violations statically. It would be possible to write a checker which works by instrumenting the code with dynamic checks though. Alternatively, it would be possible to write a static checker which catches a subset of possible problems. However, no such checker exists today.

2.21.4 Rationale

Why is GEP designed this way?

The design of GEP has the following goals, in rough unofficial order of priority:

- Support C, C-like languages, and languages which can be conceptually lowered into C (this covers a lot).
- Support optimizations such as those that are common in C compilers. In particular, GEP is a cornerstone of LLVM's [pointer aliasing model](#).
- Provide a consistent method for computing addresses so that address computations don't need to be a part of load and store instructions in the IR.
- Support non-C-like languages, to the extent that it doesn't interfere with other goals.
- Minimize target-specific information in the IR.

Why do struct member indices always use i32?

The specific type i32 is probably just a historical artifact, however it's wide enough for all practical purposes, so there's been no need to change it. It doesn't necessarily imply i32 address arithmetic; it's just an identifier which identifies a field in a struct. Requiring that all struct indices be the same reduces the range of possibilities for cases where two GEPs are effectively the same but have distinct operand types.

What's an uglygep?

Some LLVM optimizers operate on GEPs by internally lowering them into more primitive integer expressions, which allows them to be combined with other integer expressions and/or split into multiple separate integer expressions. If they've made non-trivial changes, translating back into LLVM IR can involve reverse-engineering the structure of the addressing in order to fit it into the static type of the original first operand. It isn't always possible to fully reconstruct this structure; sometimes the underlying addressing doesn't correspond with the static type at all. In such cases the optimizer instead will emit a GEP with the base pointer casted to a simple address-unit pointer, using the name "uglygep". This isn't pretty, but it's just as valid, and it's sufficient to preserve the pointer aliasing guarantees that GEP provides.

2.21.5 Summary

In summary, here's some things to always remember about the `GetElementPtr` instruction:

1. The GEP instruction never accesses memory, it only provides pointer computations.
2. The second operand to the GEP instruction is always a pointer and it must be indexed.
3. There are no superfluous indices for the GEP instruction.
4. Trailing zero indices are superfluous for pointer aliasing, but not for the types of the pointers.
5. Leading zero indices are not superfluous for pointer aliasing nor the types of the pointers.

2.22 Performance Tips for Frontend Authors

- *Abstract*
- *IR Best Practices*
 - *The Basics*
 - *Use of allocas*
 - *Avoid loads and stores of large aggregate type*
 - *Prefer zext over sext when legal*
 - *Zext GEP indices to machine register width*
 - *When to specify alignment*
 - *Other Things to Consider*
- *Describing Language Specific Properties*
 - *Restricted Operation Semantics*
 - *Describing Aliasing Properties*
 - *Modeling Memory Effects*
 - *Pass Ordering*
 - *I Still Can't Find What I'm Looking For*
- *Adding to this document*

2.22.1 Abstract

The intended audience of this document is developers of language frontends targeting LLVM IR. This document is home to a collection of tips on how to generate IR that optimizes well.

2.22.2 IR Best Practices

As with any optimizer, LLVM has its strengths and weaknesses. In some cases, surprisingly small changes in the source IR can have a large effect on the generated code.

Beyond the specific items on the list below, it's worth noting that the most mature frontend for LLVM is Clang. As a result, the further your IR gets from what Clang might emit, the less likely it is to be effectively optimized. It can often be useful to write a quick C program with the semantics you're trying to model and see what decisions Clang's IRGen makes about what IR to emit. Studying Clang's CodeGen directory can also be a good source of ideas. Note that Clang and LLVM are explicitly version locked so you'll need to make sure you're using a Clang built from the same svn revision or release as the LLVM library you're using. As always, it's *strongly* recommended that you track tip of tree development, particularly during bring up of a new project.

The Basics

1. Make sure that your Modules contain both a data layout specification and target triple. Without these pieces, none of the target specific optimization will be enabled. This can have a major effect on the generated code quality.
2. For each function or global emitted, use the most private linkage type possible (private, internal or linkonce_odr preferably). Doing so will make LLVM's inter-procedural optimizations much more effective.

3. Avoid high in-degree basic blocks (e.g. basic blocks with dozens or hundreds of predecessors). Among other issues, the register allocator is known to perform badly when confronted with such structures. The only exception to this guidance is that a unified return block with high in-degree is fine.

Use of allocas

An `alloca` instruction can be used to represent a function scoped stack slot, but can also represent dynamic frame expansion. When representing function scoped variables or locations, placing `alloca` instructions at the beginning of the entry block should be preferred. In particular, place them before any call instructions. Call instructions might get inlined and replaced with multiple basic blocks. The end result is that a following `alloca` instruction would no longer be in the entry basic block afterward.

The SROA (Scalar Replacement Of Aggregates) and Mem2Reg passes only attempt to eliminate `alloca` instructions that are in the entry basic block. Given SSA is the canonical form expected by much of the optimizer; if `allocas` can not be eliminated by Mem2Reg or SROA, the optimizer is likely to be less effective than it could be.

Avoid loads and stores of large aggregate type

LLVM currently does not optimize well loads and stores of large *aggregate types* (i.e. structs and arrays). As an alternative, consider loading individual fields from memory.

Aggregates that are smaller than the largest (performant) load or store instruction supported by the targeted hardware are well supported. These can be an effective way to represent collections of small packed fields.

Prefer zext over sext when legal

On some architectures (X86_64 is one), sign extension can involve an extra instruction whereas zero extension can be folded into a load. LLVM will try to replace a `sext` with a `zext` when it can be proven safe, but if you have information in your source language about the range of an integer value, it can be profitable to use a `zext` rather than a `sext`.

Alternatively, you can *specify the range of the value using metadata* and LLVM can do the `sext` to `zext` conversion for you.

Zext GEP indices to machine register width

Internally, LLVM often promotes the width of GEP indices to machine register width. When it does so, it will default to using sign extension (`sext`) operations for safety. If your source language provides information about the range of the index, you may wish to manually extend indices to machine register width using a `zext` instruction.

When to specify alignment

LLVM will always generate correct code if you don't specify alignment, but may generate inefficient code. For example, if you are targeting MIPS (or older ARM ISAs) then the hardware does not handle unaligned loads and stores, and so you will enter a trap-and-emulate path if you do a load or store with lower-than-natural alignment. To avoid this, LLVM will emit a slower sequence of loads, shifts and masks (or `load-right` + `load-left` on MIPS) for all cases where the load / store does not have a sufficiently high alignment in the IR.

The alignment is used to guarantee the alignment on `allocas` and globals, though in most cases this is unnecessary (most targets have a sufficiently high default alignment that they'll be fine). It is also used to provide a contract to the back end saying 'either this load/store has this alignment, or it is undefined behavior'. This means that the back end is free to emit instructions that rely on that alignment (and mid-level optimizers are free to perform transforms that require that alignment). For x86, it doesn't make much difference, as almost all instructions are alignment-independent. For MIPS, it can make a big difference.

Note that if your loads and stores are atomic, the backend will be unable to lower an under aligned access into a sequence of natively aligned accesses. As a result, alignment is mandatory for atomic loads and stores.

Other Things to Consider

1. Use `ptrtoint/inttoptr` sparingly (they interfere with pointer aliasing analysis), prefer GEPs
2. Prefer globals over `inttoptr` of a constant address - this gives you dereferencability information. In MCJIT, use `getSymbolAddress` to provide actual address.
3. Be wary of ordered and atomic memory operations. They are hard to optimize and may not be well optimized by the current optimizer. Depending on your source language, you may consider using fences instead.
4. If calling a function which is known to throw an exception (unwind), use an `invoke` with a normal destination which contains an unreachable instruction. This form conveys to the optimizer that the call returns abnormally. For an `invoke` which neither returns normally or requires unwind code in the current function, you can use a `noreturn` call instruction if desired. This is generally not required because the optimizer will convert an `invoke` with an unreachable unwind destination to a call instruction.
5. Use profile metadata to indicate statically known cold paths, even if dynamic profiling information is not available. This can make a large difference in code placement and thus the performance of tight loops.
6. When generating code for loops, try to avoid terminating the header block of the loop earlier than necessary. If the terminator of the loop header block is a loop exiting conditional branch, the effectiveness of LICM will be limited for loads not in the header. (This is due to the fact that LLVM may not know such a load is safe to speculatively execute and thus can't lift an otherwise loop invariant load unless it can prove the exiting condition is not taken.) It can be profitable, in some cases, to emit such instructions into the header even if they are not used along a rarely executed path that exits the loop. This guidance specifically does not apply if the condition which terminates the loop header is itself invariant, or can be easily discharged by inspecting the loop index variables.
7. In hot loops, consider duplicating instructions from small basic blocks which end in highly predictable terminators into their successor blocks. If a hot successor block contains instructions which can be vectorized with the duplicated ones, this can provide a noticeable throughput improvement. Note that this is not always profitable and does involve a potentially large increase in code size.
8. When checking a value against a constant, emit the check using a consistent comparison type. The GVN pass *will* optimize redundant equalities even if the type of comparison is inverted, but GVN only runs late in the pipeline. As a result, you may miss the opportunity to run other important optimizations. Improvements to EarlyCSE to remove this issue are tracked in Bug 23333.
9. Avoid using arithmetic intrinsics unless you are *required* by your source language specification to emit a particular code sequence. The optimizer is quite good at reasoning about general control flow and arithmetic, it is not anywhere near as strong at reasoning about the various intrinsics. If profitable for code generation purposes, the optimizer will likely form the intrinsics itself late in the optimization pipeline. It is *very* rarely profitable to emit these directly in the language frontend. This item explicitly includes the use of the *overflow intrinsics*.
10. Avoid using the *assume intrinsic* until you've established that a) there's no other way to express the given fact and b) that fact is critical for optimization purposes. Assumptions are a great prototyping mechanism, but they can have negative effects on both compile time and optimization effectiveness. The former is fixable with enough effort, but the later is fairly fundamental to their designed purpose.

2.22.3 Describing Language Specific Properties

When translating a source language to LLVM, finding ways to express concepts and guarantees available in your source language which are not natively provided by LLVM IR will greatly improve LLVM's ability to optimize your

code. As an example, C/C++’s ability to mark every add as “no signed wrap (nsw)” goes a long way to assisting the optimizer in reasoning about loop induction variables and thus generating more optimal code for loops.

The LLVM LangRef includes a number of mechanisms for annotating the IR with additional semantic information. It is *strongly* recommended that you become highly familiar with this document. The list below is intended to highlight a couple of items of particular interest, but is by no means exhaustive.

Restricted Operation Semantics

1. Add nsw/nuw flags as appropriate. Reasoning about overflow is generally hard for an optimizer so providing these facts from the frontend can be very impactful.
2. Use fast-math flags on floating point operations if legal. If you don’t need strict IEEE floating point semantics, there are a number of additional optimizations that can be performed. This can be highly impactful for floating point intensive computations.

Describing Aliasing Properties

1. Add noalias/align/dereferenceable/nonnull to function arguments and return values as appropriate
2. Use pointer aliasing metadata, especially tbaa metadata, to communicate otherwise-non-deducible pointer aliasing facts
3. Use inbounds on geps. This can help to disambiguate some aliasing queries.

Modeling Memory Effects

1. Mark functions as readnone/readonly/argmemonly or noreturn/nounwind when known. The optimizer will try to infer these flags, but may not always be able to. Manual annotations are particularly important for external functions that the optimizer can not analyze.
2. Use the lifetime.start/lifetime.end and invariant.start/invariant.end intrinsics where possible. Common profitable uses are for stack like data structures (thus allowing dead store elimination) and for describing life times of allocas (thus allowing smaller stack sizes).
3. Mark invariant locations using !invariant.load and TBAA’s constant flags

Pass Ordering

One of the most common mistakes made by new language frontend projects is to use the existing -O2 or -O3 pass pipelines as is. These pass pipelines make a good starting point for an optimizing compiler for any language, but they have been carefully tuned for C and C++, not your target language. You will almost certainly need to use a custom pass order to achieve optimal performance. A couple specific suggestions:

1. For languages with numerous rarely executed guard conditions (e.g. null checks, type checks, range checks) consider adding an extra execution or two of LoopUnswitch and LICM to your pass order. The standard pass order, which is tuned for C and C++ applications, may not be sufficient to remove all dischargeable checks from loops.
2. If your language uses range checks, consider using the IRCE pass. It is not currently part of the standard pass order.
3. A useful sanity check to run is to run your optimized IR back through the -O2 pipeline again. If you see noticeable improvement in the resulting IR, you likely need to adjust your pass order.

I Still Can't Find What I'm Looking For

If you didn't find what you were looking for above, consider proposing an piece of metadata which provides the optimization hint you need. Such extensions are relatively common and are generally well received by the community. You will need to ensure that your proposal is sufficiently general so that it benefits others if you wish to contribute it upstream.

You should also consider describing the problem you're facing on [llvm-dev](#) and asking for advice. It's entirely possible someone has encountered your problem before and can give good advice. If there are multiple interested parties, that also increases the chances that a metadata extension would be well received by the community as a whole.

2.22.4 Adding to this document

If you run across a case that you feel deserves to be covered here, please send a patch to [llvm-commits](#) for review.

If you have questions on these items, please direct them to [llvm-dev](#). The more relevant context you are able to give to your question, the more likely it is to be answered.

2.23 MCJIT Design and Implementation

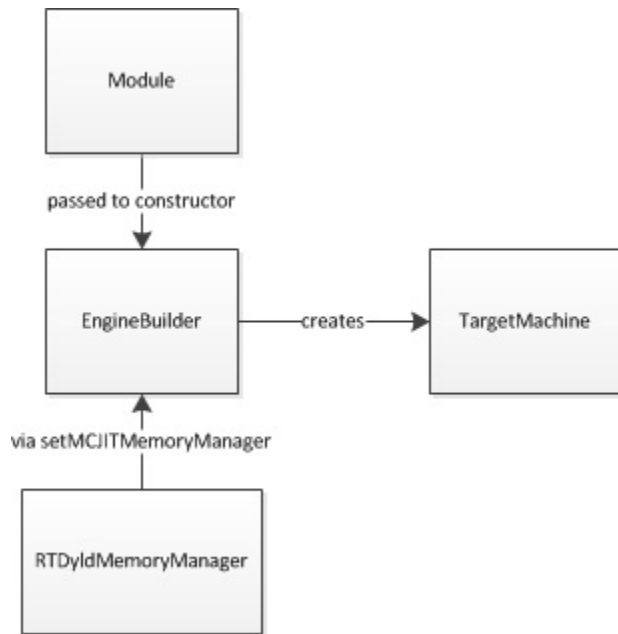
2.23.1 Introduction

This document describes the internal workings of the MCJIT execution engine and the RuntimeDyld component. It is intended as a high level overview of the implementation, showing the flow and interactions of objects throughout the code generation and dynamic loading process.

2.23.2 Engine Creation

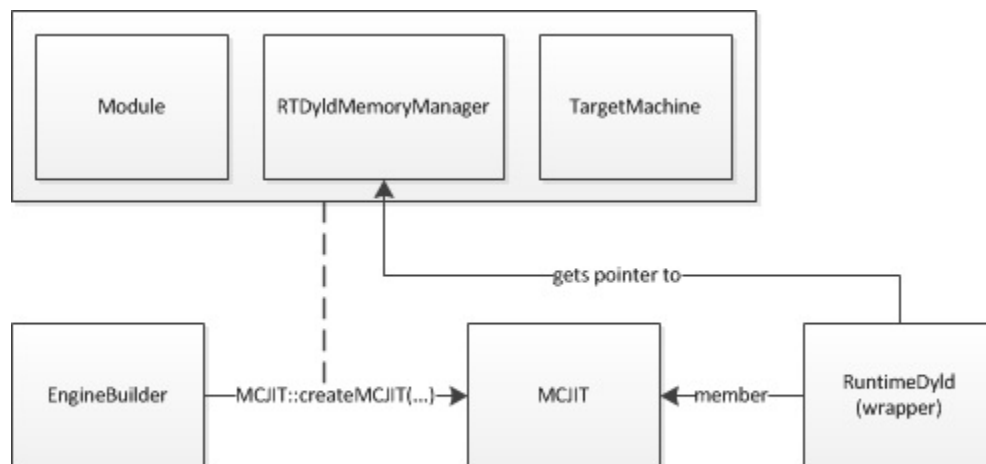
In most cases, an `EngineBuilder` object is used to create an instance of the MCJIT execution engine. The `EngineBuilder` takes an `llvm::Module` object as an argument to its constructor. The client may then set various options that we control the later be passed along to the MCJIT engine, including the selection of MCJIT as the engine type to be created. Of particular interest is the `EngineBuilder::setMCJITMemoryManager` function. If the client does not explicitly create a memory manager at this time, a default memory manager (specifically `SectionMemoryManager`) will be created when the MCJIT engine is instantiated.

Once the options have been set, a client calls `EngineBuilder::create` to create an instance of the MCJIT engine. If the client does not use the form of this function that takes a `TargetMachine` as a parameter, a new `TargetMachine` will be created based on the target triple associated with the `Module` that was used to create the `EngineBuilder`.



`EngineBuilder::create` will call the static `MCJIT::createJIT` function, passing in its pointers to the module, memory manager and target machine objects, all of which will subsequently be owned by the MCJIT object.

The MCJIT class has a member variable, `Dyld`, which contains an instance of the `RuntimeDyld` wrapper class. This member will be used for communications between MCJIT and the actual `RuntimeDyldImpl` object that gets created when an object is loaded.

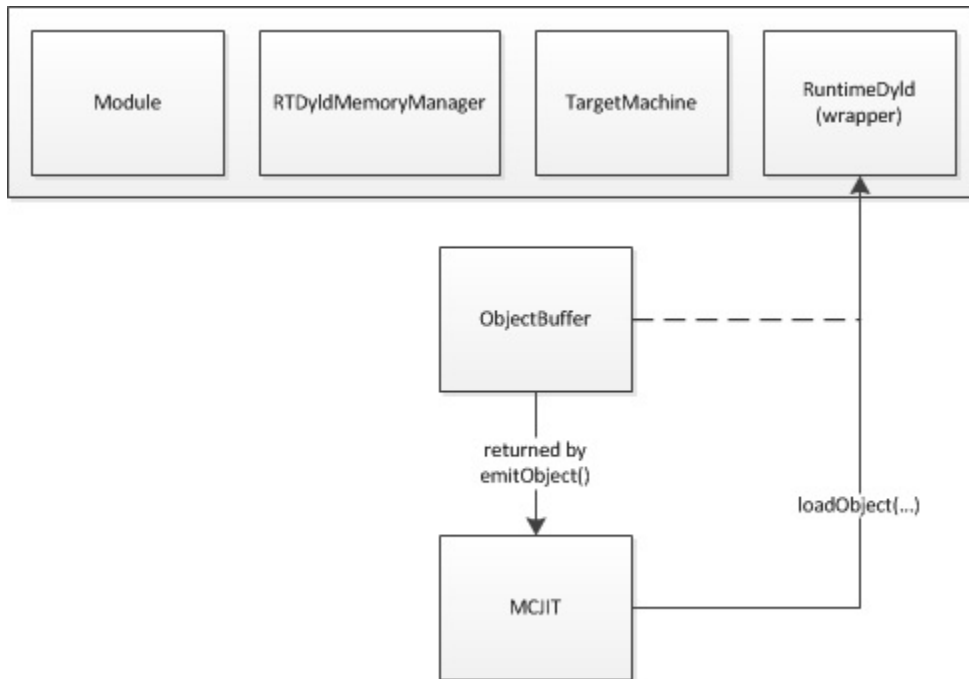


Upon creation, MCJIT holds a pointer to the `Module` object that it received from `EngineBuilder` but it does not immediately generate code for this module. Code generation is deferred until either the `MCJIT::finalizeObject` method is called explicitly or a function such as `MCJIT::getPointerToFunction` is called which requires the code to have been generated.

2.23.3 Code Generation

When code generation is triggered, as described above, MCJIT will first attempt to retrieve an object image from its `ObjectCache` member, if one has been set. If a cached object image cannot be retrieved, MCJIT will call its `emitObject` method. `MCJIT::emitObject` uses a local `PassManager` instance and creates a new `ObjectBufferStream` instance, both

of which it passes to `TargetMachine::addPassesToEmitMC` before calling `PassManager::run` on the Module with which it was created.

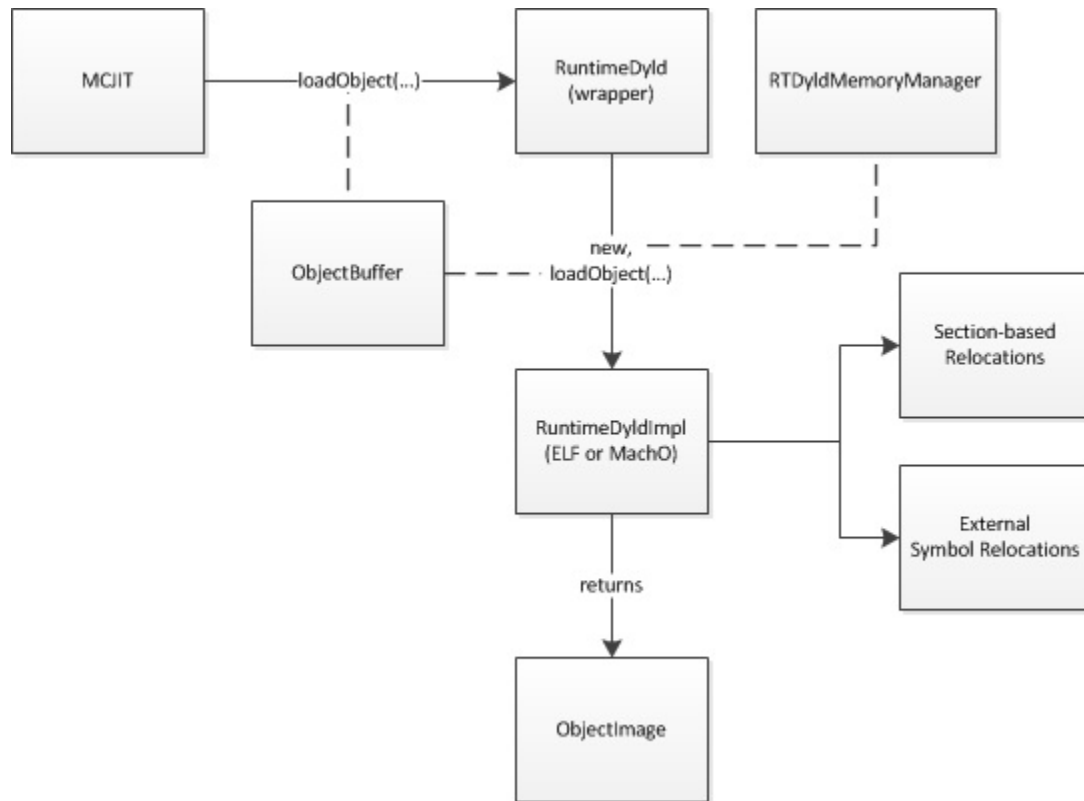


The `PassManager::run` call causes the MC code generation mechanisms to emit a complete relocatable binary object image (either in either ELF or MachO format, depending on the target) into the `ObjectBufferStream` object, which is flushed to complete the process. If an `ObjectCache` is being used, the image will be passed to the `ObjectCache` here.

At this point, the `ObjectBufferStream` contains the raw object image. Before the code can be executed, the code and data sections from this image must be loaded into suitable memory, relocations must be applied and memory permission and code cache invalidation (if required) must be completed.

2.23.4 Object Loading

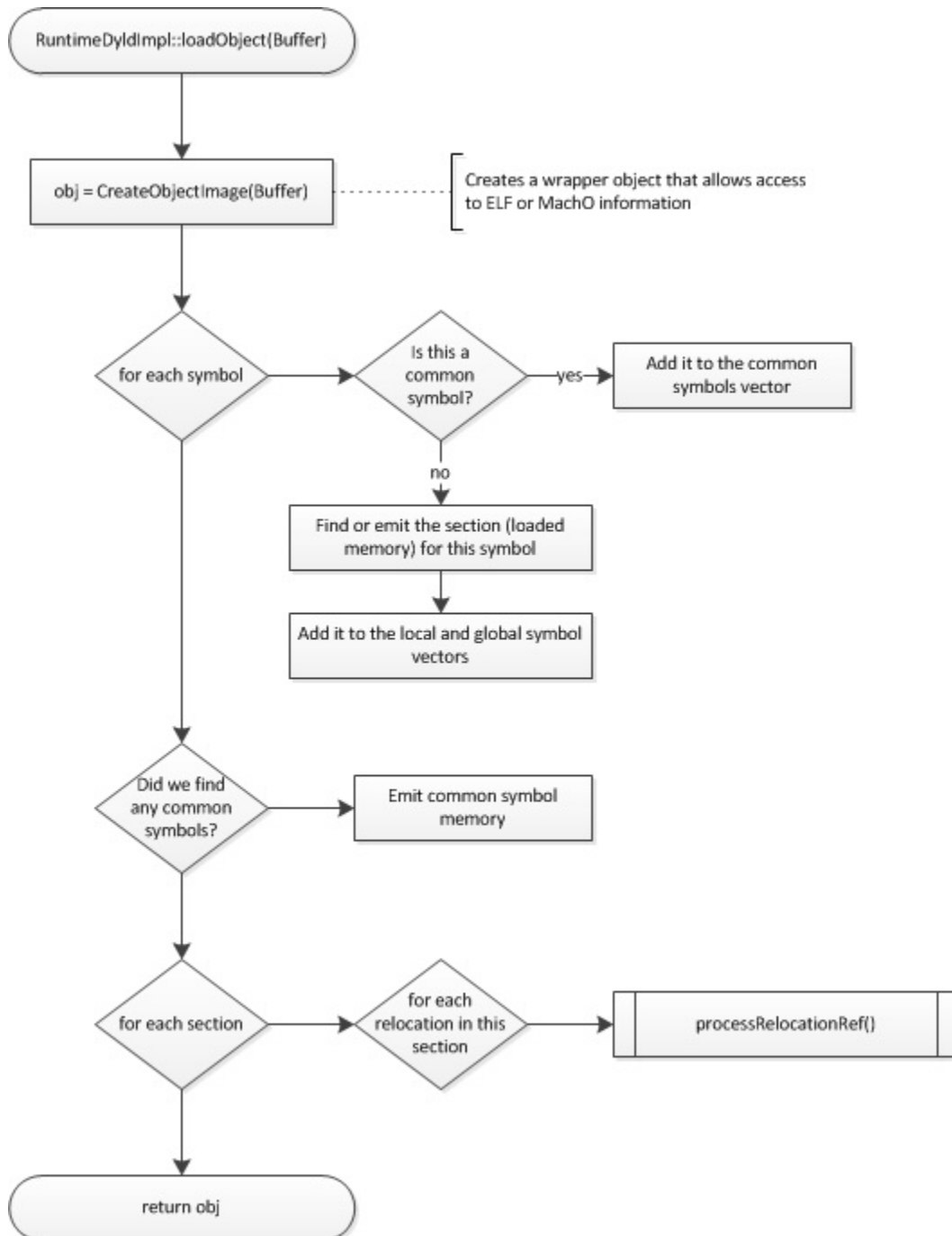
Once an object image has been obtained, either through code generation or having been retrieved from an `ObjectCache`, it is passed to `RuntimeDyld` to be loaded. The `RuntimeDyld` wrapper class examines the object to determine its file format and creates an instance of either `RuntimeDyldELF` or `RuntimeDyldMachO` (both of which derive from the `RuntimeDyldImpl` base class) and calls the `RuntimeDyldImpl::loadObject` method to perform that actual loading.



`RuntimeDyldImpl::loadObject` begins by creating an `ObjectImage` instance from the `ObjectBuffer` it received. `ObjectImage`, which wraps the `ObjectFile` class, is a helper class which parses the binary object image and provides access to the information contained in the format-specific headers, including section, symbol and relocation information.

`RuntimeDyldImpl::loadObject` then iterates through the symbols in the image. Information about common symbols is collected for later use. For each function or data symbol, the associated section is loaded into memory and the symbol is stored in a symbol table map data structure. When the iteration is complete, a section is emitted for the common symbols.

Next, `RuntimeDyldImpl::loadObject` iterates through the sections in the object image and for each section iterates through the relocations for that sections. For each relocation, it calls the format-specific `processRelocationRef` method, which will examine the relocation and store it in one of two data structures, a section-based relocation list map and an external symbol relocation map.



When `RuntimeDyldImpl::loadObject` returns, all of the code and data sections for the object will have been loaded into memory allocated by the memory manager and relocation information will have been prepared, but the relocations have not yet been applied and the generated code is still not ready to be executed.

[Currently (as of August 2013) the MCJIT engine will immediately apply relocations when `loadObject` completes. However, this shouldn't be happening. Because the code may have been generated for a remote target, the client should be given a chance to re-map the section addresses before relocations are applied. It is possible to apply relocations multiple times, but in the case where addresses are to be re-mapped, this first application is wasted effort.]

2.23.5 Address Remapping

At any time after initial code has been generated and before `finalizeObject` is called, the client can remap the address of sections in the object. Typically this is done because the code was generated for an external process and is being mapped into that process' address space. The client remaps the section address by calling `MCJIT::mapSectionAddress`. This should happen before the section memory is copied to its new location.

When `MCJIT::mapSectionAddress` is called, `MCJIT` passes the call on to `RuntimeDyldImpl` (via its `Dyld` member). `RuntimeDyldImpl` stores the new address in an internal data structure but does not update the code at this time, since other sections are likely to change.

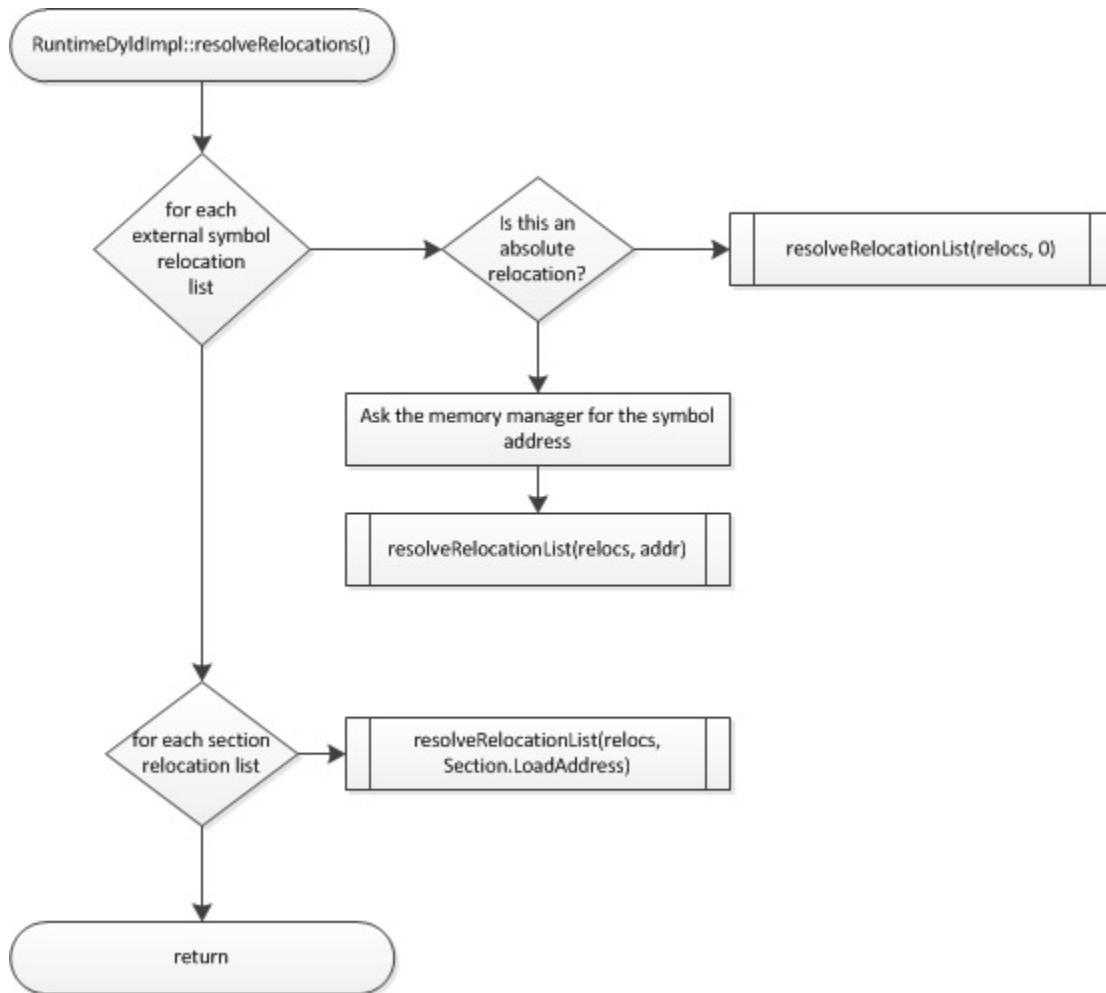
When the client is finished remapping section addresses, it will call `MCJIT::finalizeObject` to complete the remapping process.

2.23.6 Final Preparations

When `MCJIT::finalizeObject` is called, `MCJIT` calls `RuntimeDyld::resolveRelocations`. This function will attempt to locate any external symbols and then apply all relocations for the object.

External symbols are resolved by calling the memory manager's `getPointerToNamedFunction` method. The memory manager will return the address of the requested symbol in the target address space. (Note, this may not be a valid pointer in the host process.) `RuntimeDyld` will then iterate through the list of relocations it has stored which are associated with this symbol and invoke the `resolveRelocation` method which, through an format-specific implementation, will apply the relocation to the loaded section memory.

Next, `RuntimeDyld::resolveRelocations` iterates through the list of sections and for each section iterates through a list of relocations that have been saved which reference that symbol and call `resolveRelocation` for each entry in this list. The relocation list here is a list of relocations for which the symbol associated with the relocation is located in the section associated with the list. Each of these locations will have a target location at which the relocation will be applied that is likely located in a different section.



Once relocations have been applied as described above, MCJIT calls `RuntimeDyld::getEHFrameSection`, and if a non-zero result is returned passes the section data to the memory manager's `registerEHFrames` method. This allows the memory manager to call any desired target-specific functions, such as registering the EH frame information with a debugger.

Finally, MCJIT calls the memory manager's `finalizeMemory` method. In this method, the memory manager will invalidate the target code cache, if necessary, and apply final permissions to the memory pages it has allocated for code and data memory.

2.24 LLVM Community Code of Conduct

Note: This document is currently a **DRAFT** document while it is being discussed by the community.

The LLVM community has always worked to be a welcoming and respectful community, and we want to ensure that doesn't change as we grow and evolve. To that end, we have a few ground rules that we ask people to adhere to:

- *be friendly and patient,*
- *be welcoming,*
- *be considerate,*

- *be respectful,*
- *be careful in the words that you choose and be kind to others,* and
- *when we disagree, try to understand why.*

This isn't an exhaustive list of things that you can't do. Rather, take it in the spirit in which it's intended - a guide to make it easier to communicate and participate in the community.

This code of conduct applies to all spaces managed by the LLVM project or The LLVM Foundation. This includes IRC channels, mailing lists, bug trackers, LLVM events such as the developer meetings and socials, and any other forums created by the project that the community uses for communication. It applies to all of your communication and conduct in these spaces, including emails, chats, things you say, slides, videos, posters, signs, or even t-shirts you display in these spaces. In addition, violations of this code outside these spaces may, in rare cases, affect a person's ability to participate within them, when the conduct amounts to an egregious violation of this code.

If you believe someone is violating the code of conduct, we ask that you report it by emailing conduct@llvm.org. For more details please see our [Reporting Guide](#).

- **Be friendly and patient.**
- **Be welcoming.** We strive to be a community that welcomes and supports people of all backgrounds and identities. This includes, but is not limited to members of any race, ethnicity, culture, national origin, colour, immigration status, social and economic class, educational level, sex, sexual orientation, gender identity and expression, age, size, family status, political belief, religion or lack thereof, and mental and physical ability.
- **Be considerate.** Your work will be used by other people, and you in turn will depend on the work of others. Any decision you take will affect users and colleagues, and you should take those consequences into account. Remember that we're a world-wide community, so you might not be communicating in someone else's primary language.
- **Be respectful.** Not all of us will agree all the time, but disagreement is no excuse for poor behavior and poor manners. We might all experience some frustration now and then, but we cannot allow that frustration to turn into a personal attack. It's important to remember that a community where people feel uncomfortable or threatened is not a productive one. Members of the LLVM community should be respectful when dealing with other members as well as with people outside the LLVM community.
- **Be careful in the words that you choose and be kind to others.** Do not insult or put down other participants. Harassment and other exclusionary behavior aren't acceptable. This includes, but is not limited to:
 - Violent threats or language directed against another person.
 - Discriminatory jokes and language.
 - Posting sexually explicit or violent material.
 - Posting (or threatening to post) other people's personally identifying information ("doxing").
 - Personal insults, especially those using racist or sexist terms.
 - Unwelcome sexual attention.
 - Advocating for, or encouraging, any of the above behavior.

In general, if someone asks you to stop, then stop. Persisting in such behavior after being asked to stop is considered harassment.

- **When we disagree, try to understand why.** Disagreements, both social and technical, happen all the time and LLVM is no exception. It is important that we resolve disagreements and differing views constructively. Remember that we're different. The strength of LLVM comes from its varied community, people from a wide range of backgrounds. Different people have different perspectives on issues. Being unable to understand why someone holds a viewpoint doesn't mean that they're wrong. Don't forget that it is human to err and blaming each other doesn't get us anywhere. Instead, focus on helping to resolve issues and learning from mistakes.

2.24.1 Questions?

If you have questions, please feel free to contact the LLVM Foundation Code of Conduct Advisory Committee by emailing conduct@llvm.org.

(This text is based on the [Django Project Code of Conduct](#), which is in turn based on wording from the [Speak Up!](#) project.)

2.25 Compiling CUDA with clang

- *Introduction*
- *Compiling CUDA Code*
 - *Prerequisites*
 - *Invoking clang*
 - *Flags that control numerical code*
- *Standard library support*
 - *`<math.h>` and `<cmath>`*
 - *`<std::complex>`*
 - *`<algorithm>`*
- *Detecting clang vs NVCC from code*
- *Dialect Differences Between clang and nvcc*
 - *Compilation Models*
 - *Overloading Based on `__host__` and `__device__` Attributes*
 - *Using a Different Class on Host/Device*
- *Optimizations*
- *Publication*
- *Obtaining Help*

2.25.1 Introduction

This document describes how to compile CUDA code with clang, and gives some details about LLVM and clang's CUDA implementations.

This document assumes a basic familiarity with CUDA. Information about CUDA programming can be found in the [CUDA programming guide](#).

2.25.2 Compiling CUDA Code

Prerequisites

CUDA is supported in LLVM 3.9, but it's still in active development, so we recommend you [compile clang/LLVM from HEAD](#).

Before you build CUDA code, you'll need to have installed the appropriate driver for your NVIDIA GPU and the CUDA SDK. See [NVIDIA's CUDA installation guide](#) for details. Note that clang **does not support** the CUDA toolkit as installed by many Linux package managers; you probably need to install NVIDIA's package.

You will need CUDA 7.0, 7.5, or 8.0 to compile with clang.

CUDA compilation is supported on Linux, on MacOS as of 2016-11-18, and on Windows as of 2017-01-05.

Invoking clang

Invoking clang for CUDA compilation works similarly to compiling regular C++. You just need to be aware of a few additional flags.

You can use [this](#) program as a toy example. Save it as `axpy.cu`. (Clang detects that you're compiling CUDA code by noticing that your filename ends with `.cu`. Alternatively, you can pass `-x cuda`.)

To build and run, run the following commands, filling in the parts in angle brackets as described below:

```
$ clang++ axpy.cu -o axpy --cuda-gpu-arch=<GPU arch> \
-L<CUDA install path>/<lib64 or lib> \
-lcudart_static -ldl -lrt -pthread
$ ./axpy
y[0] = 2
y[1] = 4
y[2] = 6
y[3] = 8
```

On MacOS, replace `-lcudart_static` with `-lcudart`; otherwise, you may get “CUDA driver version is insufficient for CUDA runtime version” errors when you run your program.

- `<CUDA install path>` – the directory where you installed CUDA SDK. Typically, `/usr/local/cuda`. Pass e.g. `-L/usr/local/cuda/lib64` if compiling in 64-bit mode; otherwise, pass e.g. `-L/usr/local/cuda/lib`. (In CUDA, the device code and host code always have the same pointer widths, so if you're compiling 64-bit code for the host, you're also compiling 64-bit code for the device.)
- `<GPU arch>` – the [compute capability](#) of your GPU. For example, if you want to run your program on a GPU with compute capability of 3.5, specify `--cuda-gpu-arch=sm_35`.

Note: You cannot pass `compute_XX` as an argument to `--cuda-gpu-arch`; only `sm_XX` is currently supported. However, clang always includes PTX in its binaries, so e.g. a binary compiled with `--cuda-gpu-arch=sm_30` would be forwards-compatible with e.g. `sm_35` GPUs.

You can pass `--cuda-gpu-arch` multiple times to compile for multiple archs.

The `-L` and `-l` flags only need to be passed when linking. When compiling, you may also need to pass `--cuda-path=/path/to/cuda` if you didn't install the CUDA SDK into `/usr/local/cuda`, `/usr/local/cuda-7.0`, or `/usr/local/cuda-7.5`.

Flags that control numerical code

If you're using GPUs, you probably care about making numerical code run fast. GPU hardware allows for more control over numerical operations than most CPUs, but this results in more compiler options for you to juggle.

Flags you may wish to tweak include:

- `-ffp-contract={on, off, fast}` (defaults to `fast` on host and device when compiling CUDA) Controls whether the compiler emits fused multiply-add operations.
 - `off`: never emit `fma` operations, and prevent `ptxas` from fusing multiply and add instructions.
 - `on`: fuse multiplies and adds within a single statement, but never across statements (C11 semantics). Prevent `ptxas` from fusing other multiplies and adds.
 - `fast`: fuse multiplies and adds wherever profitable, even across statements. Doesn't prevent `ptxas` from fusing additional multiplies and adds.

Fused multiply-add instructions can be much faster than the unfused equivalents, but because the intermediate result in an `fma` is not rounded, this flag can affect numerical code.

- `-fcuda-flush-denormals-to-zero` (default: `off`) When this is enabled, floating point operations may flush **denormal** inputs and/or outputs to 0. Operations on denormal numbers are often much slower than the same operations on normal numbers.
- `-fcuda-approx-transcendentals` (default: `off`) When this is enabled, the compiler may emit calls to faster, approximate versions of transcendental functions, instead of using the slower, fully IEEE-compliant versions. For example, this flag allows clang to emit the `ptx sin.approx.f32` instruction.

This is implied by `-ffast-math`.

2.25.3 Standard library support

In clang and nvcc, most of the C++ standard library is not supported on the device side.

`<math.h>` and `<cmath>`

In clang, `math.h` and `cmath` are available and **pass tests** adapted from libc++'s test suite.

In nvcc `math.h` and `cmath` are mostly available. Versions of `::foo` in namespace `std` (e.g. `std::sinf`) are not available, and where the standard calls for overloads that take integral arguments, these are usually not available.

```
#include <math.h>
#include <cmath.h>

// clang is OK with everything in this function.
__device__ void test() {
    std::sin(0.); // nvcc - ok
    std::sin(0);  // nvcc - error, because no std::sin(int) override is available.
    sin(0);       // nvcc - same as above.

    sinf(0.);     // nvcc - ok
    std::sinf(0.); // nvcc - no such function
}
```

<std::complex>

nvcc does not officially support `std::complex`. It's an error to use `std::complex` in `__device__` code, but it often works in `__host__ __device__` code due to nvcc's interpretation of the "wrong-side rule" (see below). However, we have heard from implementers that it's possible to get into situations where nvcc will omit a call to an `std::complex` function, especially when compiling without optimizations.

As of 2016-11-16, clang supports `std::complex` without these caveats. It is tested with libstdc++ 4.8.5 and newer, but is known to work only with libc++ newer than 2016-11-16.

<algorithm>

In C++14, many useful functions from `<algorithm>` (notably, `std::min` and `std::max`) become `constexpr`. You can therefore use these in device code, when compiling with clang.

2.25.4 Detecting clang vs NVCC from code

Although clang's CUDA implementation is largely compatible with NVCC's, you may still want to detect when you're compiling CUDA code specifically with clang.

This is tricky, because NVCC may invoke clang as part of its own compilation process! For example, NVCC uses the host compiler's preprocessor when compiling for device code, and that host compiler may in fact be clang.

When clang is actually compiling CUDA code – rather than being used as a subtool of NVCC's – it defines the `__CUDA__` macro. `__CUDA_ARCH__` is defined only in device mode (but will be defined if NVCC is using clang as a preprocessor). So you can use the following incantations to detect clang CUDA compilation, in host and device modes:

```
#if defined(__clang__) && defined(__CUDA__) && !defined(__CUDA_ARCH__)
// clang compiling CUDA code, host mode.
#endif

#if defined(__clang__) && defined(__CUDA__) && defined(__CUDA_ARCH__)
// clang compiling CUDA code, device mode.
#endif
```

Both clang and nvcc define `__CUDACC__` during CUDA compilation. You can detect NVCC specifically by looking for `__NVCC__`.

2.25.5 Dialect Differences Between clang and nvcc

There is no formal CUDA spec, and clang and nvcc speak slightly different dialects of the language. Below, we describe some of the differences.

This section is painful; hopefully you can skip this section and live your life blissfully unaware.

Compilation Models

Most of the differences between clang and nvcc stem from the different compilation models used by clang and nvcc. nvcc uses *split compilation*, which works roughly as follows:

- Run a preprocessor over the input `.cu` file to split it into two source files: `H`, containing source code for the host, and `D`, containing source code for the device.
- For each GPU architecture `arch` that we're compiling for, do:

- Compile `D` using `nvcc` proper. The result of this is a `ptx` file for `P_arch`.
- Optionally, invoke `ptxas`, the PTX assembler, to generate a file, `S_arch`, containing GPU machine code (SASS) for `arch`.
- Invoke `fatbin` to combine all `P_arch` and `S_arch` files into a single “fat binary” file, `F`.
- Compile `H` using an external host compiler (`gcc`, `clang`, or whatever you like). `F` is packaged up into a header file which is force-included into `H`; `nvcc` generates code that calls into this header to e.g. launch kernels.

`clang` uses *merged parsing*. This is similar to split compilation, except all of the host and device code is present and must be semantically-correct in both compilation steps.

- For each GPU architecture `arch` that we’re compiling for, do:
 - Compile the input `.cu` file for device, using `clang`. `__host__` code is parsed and must be semantically correct, even though we’re not generating code for the host at this time.

The output of this step is a `ptx` file `P_arch`.

- Invoke `ptxas` to generate a SASS file, `S_arch`. Note that, unlike `nvcc`, `clang` always generates SASS code.
- Invoke `fatbin` to combine all `P_arch` and `S_arch` files into a single fat binary file, `F`.
- Compile `H` using `clang`. `__device__` code is parsed and must be semantically correct, even though we’re not generating code for the device at this time.

`F` is passed to this compilation, and `clang` includes it in a special ELF section, where it can be found by tools like `cuobjdump`.

(You may ask at this point, why does `clang` need to parse the input file multiple times? Why not parse it just once, and then use the AST to generate code for the host and each device architecture?)

Unfortunately this can’t work because we have to define different macros during host compilation and during device compilation for each GPU architecture.)

`clang`’s approach allows it to be highly robust to C++ edge cases, as it doesn’t need to decide at an early stage which declarations to keep and which to throw away. But it has some consequences you should be aware of.

Overloading Based on `__host__` and `__device__` Attributes

Let “H”, “D”, and “HD” stand for “`__host__` functions”, “`__device__` functions”, and “`__host__ __device__` functions”, respectively. Functions with no attributes behave the same as H.

`nvcc` does not allow you to create H and D functions with the same signature:

```
// nvcc: error - function "foo" has already been defined
__host__ void foo() {}
__device__ void foo() {}
```

However, `nvcc` allows you to “overload” H and D functions with different signatures:

```
// nvcc: no error
__host__ void foo(int) {}
__device__ void foo() {}
```

In `clang`, the `__host__` and `__device__` attributes are part of a function’s signature, and so it’s legal to have H and D functions with (otherwise) the same signature:

```
// clang: no error
__host__ void foo() {}
__device__ void foo() {}
```

HD functions cannot be overloaded by H or D functions with the same signature:

```
// nvcc: error - function "foo" has already been defined
// clang: error - redefinition of 'foo'
__host__ __device__ void foo() {}
__device__ void foo() {}

// nvcc: no error
// clang: no error
__host__ __device__ void bar(int) {}
__device__ void bar() {}
```

When resolving an overloaded function, clang considers the host/device attributes of the caller and callee. These are used as a tiebreaker during overload resolution. See [IdentifyCUDAPreference](#) for the full set of rules, but at a high level they are:

- D functions prefer to call other Ds. HDs are given lower priority.
- Similarly, H functions prefer to call other Hs, or `__global__` functions (with equal priority). HDs are given lower priority.
- HD functions prefer to call other HDs.

When compiling for device, HDs will call Ds with lower priority than HD, and will call Hs with still lower priority. If it's forced to call an H, the program is malformed if we emit code for this HD function. We call this the “wrong-side rule”, see example below.

The rules are symmetrical when compiling for host.

Some examples:

```
__host__ void foo();
__device__ void foo();

__host__ void bar();
__host__ __device__ void bar();

__host__ void test_host() {
    foo(); // calls H overload
    bar(); // calls H overload
}

__device__ void test_device() {
    foo(); // calls D overload
    bar(); // calls HD overload
}

__host__ __device__ void test_hd() {
    foo(); // calls H overload when compiling for host, otherwise D overload
    bar(); // always calls HD overload
}
```

Wrong-side rule example:

```
__host__ void host_only();
```

```
// We don't codegen inline functions unless they're referenced by a
// non-inline function. inline_hd1() is called only from the host side, so
// does not generate an error. inline_hd2() is called from the device side,
// so it generates an error.
inline __host__ __device__ void inline_hd1() { host_only(); } // no error
inline __host__ __device__ void inline_hd2() { host_only(); } // error

__host__ void host_fn() { inline_hd1(); }
__device__ void device_fn() { inline_hd2(); }

// This function is not inline, so it's always codegen'ed on both the host
// and the device. Therefore, it generates an error.
__host__ __device__ void not_inline_hd() { host_only(); }
```

For the purposes of the wrong-side rule, templated functions also behave like inline functions: They aren't codegen'ed unless they're instantiated (usually as part of the process of invoking them).

clang's behavior with respect to the wrong-side rule matches nvcc's, except nvcc only emits a warning for `not_inline_hd`; device code is allowed to call `not_inline_hd`. In its generated code, nvcc may omit `not_inline_hd`'s call to `host_only` entirely, or it may try to generate code for `host_only` on the device. What you get seems to depend on whether or not the compiler chooses to inline `host_only`.

Member functions, including constructors, may be overloaded using H and D attributes. However, destructors cannot be overloaded.

Using a Different Class on Host/Device

Occasionally you may want to have a class with different host/device versions.

If all of the class's members are the same on the host and device, you can just provide overloads for the class's member functions.

However, if you want your class to have different members on host/device, you won't be able to provide working H and D overloads in both classes. In this case, clang is likely to be unhappy with you.

```
#ifndef __CUDA_ARCH__
struct S {
    __device__ void foo() { /* use device_only */ }
    int device_only;
};
#else
struct S {
    __host__ void foo() { /* use host_only */ }
    double host_only;
};

__device__ void test() {
    S s;
    // clang generates an error here, because during host compilation, we
    // have ifdef'ed away the __device__ overload of S::foo(). The __device__
    // overload must be present *even during host compilation*.
    s.foo();
}
#endif
```

We posit that you don't really want to have classes with different members on H and D. For example, if you were to pass one of these as a parameter to a kernel, it would have a different layout on H and D, so would not work properly.

To make code like this compatible with clang, we recommend you separate it out into two classes. If you need to write code that works on both host and device, consider writing an overloaded wrapper function that returns different types on host and device.

```
struct HostS { ... };
struct DeviceS { ... };

__host__ HostS MakeStruct() { return HostS(); }
__device__ DeviceS MakeStruct() { return DeviceS(); }

// Now host and device code can call MakeStruct().
```

Unfortunately, this idiom isn't compatible with nvcc, because it doesn't allow you to overload based on the H/D attributes. Here's an idiom that works with both clang and nvcc:

```
struct HostS { ... };
struct DeviceS { ... };

#ifdef __NVCC__
    #ifndef __CUDA_ARCH__
        __host__ HostS MakeStruct() { return HostS(); }
    #else
        __device__ DeviceS MakeStruct() { return DeviceS(); }
    #endif
#else
    __host__ HostS MakeStruct() { return HostS(); }
    __device__ DeviceS MakeStruct() { return DeviceS(); }
#endif

// Now host and device code can call MakeStruct().
```

Hopefully you don't have to do this sort of thing often.

2.25.6 Optimizations

Modern CPUs and GPUs are architecturally quite different, so code that's fast on a CPU isn't necessarily fast on a GPU. We've made a number of changes to LLVM to make it generate good GPU code. Among these changes are:

- **Straight-line scalar optimizations** – These reduce redundancy within straight-line code.
- **Aggressive speculative execution** – This is mainly for promoting straight-line scalar optimizations, which are most effective on code along dominator paths.
- **Memory space inference** – In PTX, we can operate on pointers that are in a particular “address space” (global, shared, constant, or local), or we can operate on pointers in the “generic” address space, which can point to anything. Operations in a non-generic address space are faster, but pointers in CUDA are not explicitly annotated with their address space, so it's up to LLVM to infer it where possible.
- **Bypassing 64-bit divides** – This was an existing optimization that we enabled for the PTX backend.

64-bit integer divides are much slower than 32-bit ones on NVIDIA GPUs. Many of the 64-bit divides in our benchmarks have a divisor and dividend which fit in 32-bits at runtime. This optimization provides a fast path for this common case.

- **Aggressive loop unrolling and function inlining** – Loop unrolling and function inlining need to be more aggressive for GPUs than for CPUs because control flow transfer in GPU is more expensive. More aggressive unrolling and inlining also promote other optimizations, such as constant propagation and SROA, which sometimes speed up code by over 10x.

(Programmers can force unrolling and inline using clang’s `loop unrolling pragmas` and `__attribute__((always_inline))`.)

2.25.7 Publication

The team at Google published a paper in CGO 2016 detailing the optimizations they’d made to clang/LLVM. Note that “gpuc” is no longer a meaningful name: The relevant tools are now just vanilla clang/LLVM.

[gpuc: An Open-Source GPGPU Compiler](#)

Jingyue Wu, Artem Belevich, Eli Bendersky, Mark Heffernan, Chris Leary, Jacques Pienaar, Bjarke Roune, Rob Springer, Xuetian Weng, Robert Hundt

Proceedings of the 2016 International Symposium on Code Generation and Optimization (CGO 2016)

[Slides from the CGO talk](#)

[Tutorial given at CGO](#)

2.25.8 Obtaining Help

To obtain help on LLVM in general and its CUDA support, see [the LLVM community](#).

2.26 Reporting Guide

Note: This document is currently a **DRAFT** document while it is being discussed by the community.

If you believe someone is violating the [code of conduct](#) you can always report it to the LLVM Foundation Code of Conduct Advisory Committee by emailing conduct@llvm.org. **All reports will be kept confidential.** This isn’t a public list and only [members](#) of the advisory committee will receive the report.

If you believe anyone is in **physical danger**, please notify appropriate law enforcement first. If you are unsure what law enforcement agency is appropriate, please include this in your report and we will attempt to notify them.

If the violation occurs at an event such as a Developer Meeting and requires immediate attention, you can also reach out to any of the event organizers or staff. Event organizers and staff will be prepared to handle the incident and able to help. If you cannot find one of the organizers, the venue staff can locate one for you. We will also post detailed contact information for specific events as part of each events’ information. In person reports will still be kept confidential exactly as above, but also feel free to (anonymously if needed) email conduct@llvm.org.

Note: The LLVM community has long handled inappropriate behavior on its own, using both private communication and public responses. Nothing in this document is intended to discourage this self enforcement of community norms. Instead, the mechanisms described here are intended to supplement any self enforcement within the community. They provide avenues for handling severe cases or cases where the reporting party does not wish to respond directly for any reason.

2.26.1 Filing a report

Reports can be as formal or informal as needed for the situation at hand. If possible, please include as much information as you can. If you feel comfortable, please consider including:

- Your contact info (so we can get in touch with you if we need to follow up).
- Names (real, nicknames, or pseudonyms) of any individuals involved. If there were other witnesses besides you, please try to include them as well.
- When and where the incident occurred. Please be as specific as possible.
- Your account of what occurred. If there is a publicly available record (e.g. a mailing list archive or a public IRC logger) please include a link.
- Any extra context you believe existed for the incident.
- If you believe this incident is ongoing.
- Any other information you believe we should have.

2.26.2 What happens after you file a report?

You will receive an email from the advisory committee acknowledging receipt within 24 hours (and we will aim to respond much quicker than that).

The advisory committee will immediately meet to review the incident and try to determine:

- What happened and who was involved.
- Whether this event constitutes a code of conduct violation.
- Whether this is an ongoing situation, or if there is a threat to anyone's physical safety.

If this is determined to be an ongoing incident or a threat to physical safety, the working groups' immediate priority will be to protect everyone involved. This means we may delay an "official" response until we believe that the situation has ended and that everyone is physically safe.

The working group will try to contact other parties involved or witnessing the event to gain clarity on what happened and understand any different perspectives.

Once the advisory committee has a complete account of the events they will make a decision as to how to respond. Responses may include:

- Nothing, if we determine no violation occurred or it has already been appropriately resolved.
- Providing either moderation or mediation to ongoing interactions (where appropriate, safe, and desired by both parties).
- A private reprimand from the working group to the individuals involved.
- An imposed vacation (i.e. asking someone to "take a week off" from a mailing list or IRC).
- A public reprimand.
- A permanent or temporary ban from some or all LLVM spaces (mailing lists, IRC, etc.)
- Involvement of relevant law enforcement if appropriate.

If the situation is not resolved within one week, we'll respond within one week to the original reporter with an update and explanation.

Once we've determined our response, we will separately contact the original reporter and other individuals to let them know what actions (if any) we'll be taking. We will take into account feedback from the individuals involved on the appropriateness of our response, but we don't guarantee we'll act on it.

After any incident, the advisory committee will make a report on the situation to the LLVM Foundation board. The board may choose to make a public statement about the incident. If that's the case, the identities of anyone involved will remain confidential unless instructed by those individuals otherwise.

2.26.3 Appealing

Only permanent resolutions (such as bans) or requests for public actions may be appealed. To appeal a decision of the working group, contact the LLVM Foundation board at board@llvm.org with your appeal and the board will review the case.

In general, it is **not** appropriate to appeal a particular decision on a public mailing list. Doing so would involve disclosure of information which would be confidential. Disclosing this kind of information publicly may be considered a separate and (potentially) more serious violation of the Code of Conduct. This is not meant to limit discussion of the Code of Conduct, the advisory board itself, or the appropriateness of responses in general, but **please** refrain from mentioning specific facts about cases without the explicit permission of all parties involved.

2.26.4 Members of the Code of Conduct Advisory Committee

The members serving on the advisory committee are listed here with contact information in case you are more comfortable talking directly to a specific member of the committee.

Note: FIXME: When we form the initial advisory committee, the members names and private contact info need to be added here.

(This text is based on the [Django Project](#) Code of Conduct, which is in turn based on wording from the [Speak Up! project](#).)

2.27 Benchmarking tips

2.27.1 Introduction

For benchmarking a patch we want to reduce all possible sources of noise as much as possible. How to do that is very OS dependent.

Note that low noise is required, but not sufficient. It does not exclude measurement bias. See <https://www.cis.upenn.edu/~cis501/papers/producing-wrong-data.pdf> for example.

2.27.2 General

- Use a high resolution timer, e.g. `perf` under linux.
- Run the benchmark multiple times to be able to recognize noise.
- Disable as many processes or services as possible on the target system.
- Disable frequency scaling, turbo boost and address space randomization (see OS specific section).
- Static link if the OS supports it. That avoids any variation that might be introduced by loading dynamic libraries. This can be done by passing `-DLLVM_BUILD_STATIC=ON` to `cmake`.

- Try to avoid storage. On some systems you can use tmpfs. Putting the program, inputs and outputs on tmpfs avoids touching a real storage system, which can have a pretty big variability.

To mount it (on linux and freebsd at least):

```
mount -t tmpfs -o size=<XX>g none dir_to_mount
```

2.27.3 Linux

- Disable address space randomization:

```
echo 0 > /proc/sys/kernel/randomize_va_space
```

- Set scaling_governor to performance:

```
for i in $(ls /sys/devices/system/cpu/cpu*/cpufreq/scaling_governor)
do
    echo performance > /sys/devices/system/cpu/cpu*/cpufreq/scaling_governor
done
```

- Use <https://github.com/lpechacek/cpuset> to reserve cpus for just the program you are benchmarking. If using perf, leave at least 2 cores so that perf runs in one and your program in another:

```
cset shield -c N1,N2 -k on
```

This will move all threads out of N1 and N2. The `-k on` means that even kernel threads are moved out.

- Disable the SMT pair of the cpus you will use for the benchmark. The pair of cpu N can be found in `/sys/devices/system/cpu/cpuN/topology/thread_siblings_list` and disabled with:

```
echo 0 > /sys/devices/system/cpu/cpuX/online
```

- Run the program with:

```
cset shield --exec -- perf stat -r 10 <cmd>
```

This will run the command after `--` in the isolated cpus. The particular perf command runs the `<cmd>` 10 times and reports statistics.

With these in place you can expect perf variations of less than 0.1%.

Linux Intel

- Disable turbo mode:

```
echo 1 > /sys/devices/system/cpu/intel_pstate/no_turbo
```

2.28 A guide to Dockerfiles for building LLVM

2.28.1 Introduction

You can find a number of sources to build docker images with LLVM components in `llvm/utils/docker`. They can be used by anyone who wants to build the docker images for their own use, or as a starting point for someone who wants to write their own Dockerfiles.

We currently provide Dockerfiles with `debian8` and `nvidia-cuda` base images. We also provide an `example` image, which contains placeholders that one would need to fill out in order to produce Dockerfiles for a new docker image.

Why?

Docker images provide a way to produce binary distributions of software inside a controlled environment. Having Dockerfiles to build docker images inside LLVM repo makes them much more discoverable than putting them into any other place.

Docker basics

If you've never heard about Docker before, you might find this section helpful to get a very basic explanation of it. [Docker](#) is a popular solution for running programs in an isolated and reproducible environment, especially to maintain releases for software deployed to large distributed fleets. It uses linux kernel namespaces and cgroups to provide a lightweight isolation inside currently running linux kernel. A single active instance of dockerized environment is called a *docker container*. A snapshot of a docker container filesystem is called a *docker image*. One can start a container from a prebuilt docker image.

Docker images are built from a so-called *Dockerfile*, a source file written in a specialized language that defines instructions to be used when build the docker image (see [official documentation](#) for more details). A minimal Dockerfile typically contains a base image and a number of RUN commands that have to be executed to build the image. When building a new image, docker will first download your base image, mount its filesystem as read-only and then add a writable overlay on top of it to keep track of all filesystem modifications, performed while building your image. When the build process is finished, a diff between your image's final filesystem state and the base image's filesystem is stored in the resulting image.

2.28.2 Overview

The `llvm/utils/docker` folder contains Dockerfiles and simple bash scripts to serve as a basis for anyone who wants to create their own Docker image with LLVM components, compiled from sources. The sources are checked out from the upstream svn repository when building the image.

Inside each subfolder we host Dockerfiles for two images:

- `build/` image is used to compile LLVM, it installs a system compiler and all build dependencies of LLVM. After the build process is finished, the build image will have an archive with compiled components at `/tmp/clang.tar.gz`.
- `release/` image usually only contains LLVM components, compiled by the `build/` image, and also `libstdc++` and `binutils` to make image minimally useful for C++ development. The assumption is that you usually want clang to be one of the provided components.

To build both of those images, use `build_docker_image.sh` script. It will checkout LLVM sources and build clang in the `build` container, copy results of the build to the local filesystem and then build the `release` container using those. The `build_docker_image.sh` accepts a list of LLVM repositories to checkout, and arguments for CMake invocation.

If you want to write your own docker image, start with an `example/` subfolder. It provides incomplete Dockerfiles with (very few) FIXMEs explaining the steps you need to take in order to make your Dockerfiles functional.

2.28.3 Usage

The `llvm/utils/build_docker_image.sh` script provides a rather high degree of control on how to run the build. It allows you to specify the projects to checkout from svn and provide a list of CMake arguments to use during when building LLVM inside docker container.

Here's a very simple example of getting a docker image with clang binary, compiled by the system compiler in the debian8 image:

```
./llvm/utils/docker/build_docker_image.sh \
  --source debian8 \
  --docker-repository clang-debian8 --docker-tag "staging" \
  -p clang -i install-clang -i install-clang-headers \
  -- \
  -DCMAKE_BUILD_TYPE=Release
```

Note that a build like that doesn't use a 2-stage build process that you probably want for clang. Running a 2-stage build is a little more intricate, this command will do that:

```
# Run a 2-stage build.
# LLVM_TARGETS_TO_BUILD=Native is to reduce stage1 compile time.
# Options, starting with BOOTSTRAP_* are passed to stage2 cmake invocation.
./build_docker_image.sh \
  --source debian8 \
  --docker-repository clang-debian8 --docker-tag "staging" \
  -p clang -i stage2-install-clang -i stage2-install-clang-headers \
  -- \
  -DLLVM_TARGETS_TO_BUILD=Native -DCMAKE_BUILD_TYPE=Release \
  -DBOOTSTRAP_CMAKE_BUILD_TYPE=Release \
  -DCLANG_ENABLE_BOOTSTRAP=ON -DCLANG_BOOTSTRAP_TARGETS="install-clang;install-clang-headers"
```

This will produce two images, a release image `clang-debian8:staging` and a build image `clang-debian8-build:staging` from the latest upstream revision. After the image is built you can run `bash` inside a container based on your image like this:

```
docker run -ti clang-debian8:staging bash
```

Now you can run `bash` commands as you normally would:

```
root@80f351b51825:/# clang -v
clang version 5.0.0 (trunk 305064)
Target: x86_64-unknown-linux-gnu
Thread model: posix
InstalledDir: /bin
Found candidate GCC installation: /usr/lib/gcc/x86_64-linux-gnu/4.8
Found candidate GCC installation: /usr/lib/gcc/x86_64-linux-gnu/4.8.4
Found candidate GCC installation: /usr/lib/gcc/x86_64-linux-gnu/4.9
Found candidate GCC installation: /usr/lib/gcc/x86_64-linux-gnu/4.9.2
Selected GCC installation: /usr/lib/gcc/x86_64-linux-gnu/4.9
Candidate multilib: .;@m64
Selected multilib: .;@m64
```

2.28.4 Which image should I choose?

We currently provide two images: `debian8`-based and `nvidia-cuda`-based. They differ in the base image that they use, i.e. they have a different set of preinstalled binaries. `Debian8` is very minimal, `nvidia-cuda` is larger, but has

preinstalled CUDA libraries and allows to access a GPU, installed on your machine.

If you need a minimal linux distribution with only clang and libstdc++ included, you should try debian8-based image.

If you want to use CUDA libraries and have access to a GPU on your machine, you should choose nvidia-cuda-based image and use [nvidia-docker](#) to run your docker containers. Note that you don't need nvidia-docker to build the images, but you need it in order to have an access to GPU from a docker container that is running the built image.

If you have a different use-case, you could create your own image based on `example/` folder.

Any docker image can be built and run using only the docker binary, i.e. you can run debian8 build on Fedora or any other Linux distribution. You don't need to install CMake, compilers or any other clang dependencies. It is all handled during the build process inside Docker's isolated environment.

2.28.5 Stable build

If you want a somewhat recent and somewhat stable build, use the `branches/google/stable` branch, i.e. the following command will produce a debian8-based image using the latest `google/stable` sources for you:

```
./llvm/utils/docker/build_docker_image.sh \
-s debian8 --d clang-debian8 -t "staging" \
--branch branches/google/stable \
-p clang -i install-clang -i install-clang-headers \
-- \
-DCMAKE_BUILD_TYPE=Release
```

2.28.6 Minimizing docker image size

Due to Docker restrictions we use two images (i.e., build and release folders) for the release image to be as small as possible. It's much easier to achieve that using two images, because Docker would store a filesystem layer for each command in the Dockerfile, i.e. if you install some packages in one command, then remove those in a separate command, the size of the resulting image will still be proportional to the size of an image with installed packages. Therefore, we strive to provide a very simple release image which only copies compiled clang and does not do anything else.

Docker 1.13 added a `--squash` flag that allows to flatten the layers of the image, i.e. remove the parts that were actually deleted. That is an easier way to produce the smallest images possible by using just a single image. We do not use it because as of today the flag is in experimental stage and not everyone may have the latest docker version available. When the flag is out of experimental stage, we should investigate replacing two images approach with just a single image, built using `--squash` flag.

Getting Started with the LLVM System Discusses how to get up and running quickly with the LLVM infrastructure. Everything from unpacking and compilation of the distribution to execution of some tools.

Building LLVM with CMake An addendum to the main Getting Started guide for those using the [CMake build system](#).

How To Build On ARM Notes on building and testing LLVM/Clang on ARM.

How To Cross-Compile Clang/LLVM using Clang/LLVM Notes on cross-building and testing LLVM/Clang.

Getting Started with the LLVM System using Microsoft Visual Studio An addendum to the main Getting Started guide for those using Visual Studio on Windows.

LLVM Tutorial: Table of Contents Tutorials about using LLVM. Includes a tutorial about making a custom language with LLVM.

LLVM Command Guide A reference manual for the LLVM command line utilities ("man" pages for LLVM tools).

LLVM's Analysis and Transform Passes A list of optimizations and analyses implemented in LLVM.

Frequently Asked Questions (FAQ) A list of common questions and problems and their solutions.

Release notes for the current release This describes new features, known bugs, and other limitations.

How to submit an LLVM bug report Instructions for properly submitting information about any bugs you run into in the LLVM system.

Sphinx Quickstart Template A template + tutorial for writing new Sphinx documentation. It is meant to be read in source form.

LLVM Testing Infrastructure Guide A reference manual for using the LLVM testing infrastructure.

How to build the C, C++, ObjC, and ObjC++ front end Instructions for building the clang front-end from source.

The LLVM Lexicon Definition of acronyms, terms and concepts used in LLVM.

How To Add Your Build Configuration To LLVM Buildbot Infrastructure Instructions for adding new builder to LLVM buildbot master.

YAML I/O A reference guide for using LLVM's YAML I/O library.

The Often Misunderstood GEP Instruction Answers to some very frequent questions about LLVM's most frequently misunderstood instruction.

Performance Tips for Frontend Authors A collection of tips for frontend authors on how to generate IR which LLVM is able to effectively optimize.

A guide to Dockerfiles for building LLVM A reference for using Dockerfiles provided with LLVM.

For developers of applications which use LLVM as a library.

3.1 LLVM Atomic Instructions and Concurrency Guide

- *Introduction*
- *Optimization outside atomic*
- *Atomic instructions*
- *Atomic orderings*
 - *NotAtomic*
 - *Unordered*
 - *Monotonic*
 - *Acquire*
 - *Release*
 - *AcquireRelease*
 - *SequentiallyConsistent*
- *Atomics and IR optimization*
- *Atomics and Codegen*
- *Libcalls: `__atomic_*`*
- *Libcalls: `__sync_*`*

3.1.1 Introduction

LLVM supports instructions which are well-defined in the presence of threads and asynchronous signals.

The atomic instructions are designed specifically to provide readable IR and optimized code generation for the following:

- The C++11 `<atomic>` header. ([C++11 draft available here.](#)) ([C11 draft available here.](#))
- Proper semantics for Java-style memory, for both `volatile` and regular shared variables. ([Java Specification](#))
- gcc-compatible `__sync_*` builtins. ([Description](#))
- Other scenarios with atomic semantics, including `static` variables with non-trivial constructors in C++.

Atomic and volatile in the IR are orthogonal; “volatile” is the C/C++ volatile, which ensures that every volatile load and store happens and is performed in the stated order. A couple examples: if a `SequentiallyConsistent` store is immediately followed by another `SequentiallyConsistent` store to the same address, the first store can be erased. This transformation is not allowed for a pair of volatile stores. On the other hand, a non-volatile non-atomic load can be moved across a volatile load freely, but not an `Acquire` load.

This document is intended to provide a guide to anyone either writing a frontend for LLVM or working on optimization passes for LLVM with a guide for how to deal with instructions with special semantics in the presence of concurrency. This is not intended to be a precise guide to the semantics; the details can get extremely complicated and unreadable, and are not usually necessary.

3.1.2 Optimization outside atomic

The basic ‘load’ and ‘store’ allow a variety of optimizations, but can lead to undefined results in a concurrent environment; see *NotAtomic*. This section specifically goes into the one optimizer restriction which applies in concurrent environments, which gets a bit more of an extended description because any optimization dealing with stores needs to be aware of it.

From the optimizer’s point of view, the rule is that if there are not any instructions with atomic ordering involved, concurrency does not matter, with one exception: if a variable might be visible to another thread or signal handler, a store cannot be inserted along a path where it might not execute otherwise. Take the following example:

```
/* C code, for readability; run through clang -O2 -S -emit-llvm to get
   equivalent IR */
int x;
void f(int* a) {
    for (int i = 0; i < 100; i++) {
        if (a[i])
            x += 1;
    }
}
```

The following is equivalent in non-concurrent situations:

```
int x;
void f(int* a) {
    int xtemp = x;
    for (int i = 0; i < 100; i++) {
        if (a[i])
            xtemp += 1;
    }
    x = xtemp;
}
```

However, LLVM is not allowed to transform the former to the latter: it could indirectly introduce undefined behavior if another thread can access `x` at the same time. (This example is particularly of interest because before the concurrency model was implemented, LLVM would perform this transformation.)

Note that speculative loads are allowed; a load which is part of a race returns `undef`, but does not have undefined behavior.

3.1.3 Atomic instructions

For cases where simple loads and stores are not sufficient, LLVM provides various atomic instructions. The exact guarantees provided depend on the ordering; see [Atomic orderings](#).

`load atomic` and `store atomic` provide the same basic functionality as non-atomic loads and stores, but provide additional guarantees in situations where threads and signals are involved.

`cmpxchg` and `atomicrmw` are essentially like an atomic load followed by an atomic store (where the store is conditional for `cmpxchg`), but no other memory operation can happen on any thread between the load and store.

A `fence` provides Acquire and/or Release ordering which is not part of another operation; it is normally used along with Monotonic memory operations. A Monotonic load followed by an Acquire fence is roughly equivalent to an Acquire load, and a Monotonic store following a Release fence is roughly equivalent to a Release store. Sequentially-Consistent fences behave as both an Acquire and a Release fence, and offer some additional complicated guarantees, see the C++11 standard for details.

Frontends generating atomic instructions generally need to be aware of the target to some degree; atomic instructions are guaranteed to be lock-free, and therefore an instruction which is wider than the target natively supports can be impossible to generate.

3.1.4 Atomic orderings

In order to achieve a balance between performance and necessary guarantees, there are six levels of atomicity. They are listed in order of strength; each level includes all the guarantees of the previous level except for Acquire/Release. (See also [LangRef Ordering](#).)

NotAtomic

NotAtomic is the obvious, a load or store which is not atomic. (This isn't really a level of atomicity, but is listed here for comparison.) This is essentially a regular load or store. If there is a race on a given memory location, loads from that location return `undef`.

Relevant standard This is intended to match shared variables in C/C++, and to be used in any other context where memory access is necessary, and a race is impossible. (The precise definition is in [LangRef Memory Model](#).)

Notes for frontends The rule is essentially that all memory accessed with basic loads and stores by multiple threads should be protected by a lock or other synchronization; otherwise, you are likely to run into undefined behavior. If your frontend is for a “safe” language like Java, use `Unordered` to load and store any shared variable. Note that NotAtomic volatile loads and stores are not properly atomic; do not try to use them as a substitute. (Per the C/C++ standards, volatile does provide some limited guarantees around asynchronous signals, but atomics are generally a better solution.)

Notes for optimizers Introducing loads to shared variables along a codepath where they would not otherwise exist is allowed; introducing stores to shared variables is not. See [Optimization outside atomic](#).

Notes for code generation The one interesting restriction here is that it is not allowed to write to bytes outside of the bytes relevant to a store. This is mostly relevant to unaligned stores: it is not allowed in general to convert an unaligned store into two aligned stores of the same width as the unaligned store. Backends are also expected to

generate an i8 store as an i8 store, and not an instruction which writes to surrounding bytes. (If you are writing a backend for an architecture which cannot satisfy these restrictions and cares about concurrency, please send an email to llvm-dev.)

Unordered

Unordered is the lowest level of atomicity. It essentially guarantees that races produce somewhat sane results instead of having undefined behavior. It also guarantees the operation to be lock-free, so it does not depend on the data being part of a special atomic structure or depend on a separate per-process global lock. Note that code generation will fail for unsupported atomic operations; if you need such an operation, use explicit locking.

Relevant standard This is intended to match the Java memory model for shared variables.

Notes for frontends This cannot be used for synchronization, but is useful for Java and other “safe” languages which need to guarantee that the generated code never exhibits undefined behavior. Note that this guarantee is cheap on common platforms for loads of a native width, but can be expensive or unavailable for wider loads, like a 64-bit store on ARM. (A frontend for Java or other “safe” languages would normally split a 64-bit store on ARM into two 32-bit unordered stores.)

Notes for optimizers In terms of the optimizer, this prohibits any transformation that transforms a single load into multiple loads, transforms a store into multiple stores, narrows a store, or stores a value which would not be stored otherwise. Some examples of unsafe optimizations are narrowing an assignment into a bitfield, re-materializing a load, and turning loads and stores into a memcpy call. Reordering unordered operations is safe, though, and optimizers should take advantage of that because unordered operations are common in languages that need them.

Notes for code generation These operations are required to be atomic in the sense that if you use unordered loads and unordered stores, a load cannot see a value which was never stored. A normal load or store instruction is usually sufficient, but note that an unordered load or store cannot be split into multiple instructions (or an instruction which does multiple memory operations, like LDRD on ARM without LPAE, or not naturally-aligned LDRD on LPAE ARM).

Monotonic

Monotonic is the weakest level of atomicity that can be used in synchronization primitives, although it does not provide any general synchronization. It essentially guarantees that if you take all the operations affecting a specific address, a consistent ordering exists.

Relevant standard This corresponds to the C++11/C11 `memory_order_relaxed`; see those standards for the exact definition.

Notes for frontends If you are writing a frontend which uses this directly, use with caution. The guarantees in terms of synchronization are very weak, so make sure these are only used in a pattern which you know is correct. Generally, these would either be used for atomic operations which do not protect other memory (like an atomic counter), or along with a `fence`.

Notes for optimizers In terms of the optimizer, this can be treated as a read+write on the relevant memory location (and alias analysis will take advantage of that). In addition, it is legal to reorder non-atomic and Unordered loads around Monotonic loads. CSE/DSE and a few other optimizations are allowed, but Monotonic operations are unlikely to be used in ways which would make those optimizations useful.

Notes for code generation Code generation is essentially the same as that for unordered for loads and stores. No fences are required. `cmpxchg` and `atomicrmw` are required to appear as a single operation.

Acquire

Acquire provides a barrier of the sort necessary to acquire a lock to access other memory with normal loads and stores.

Relevant standard This corresponds to the C++11/C11 `memory_order_acquire`. It should also be used for C++11/C11 `memory_order_consume`.

Notes for frontends If you are writing a frontend which uses this directly, use with caution. Acquire only provides a semantic guarantee when paired with a Release operation.

Notes for optimizers Optimizers not aware of atomics can treat this like a nothrow call. It is also possible to move stores from before an Acquire load or read-modify-write operation to after it, and move non-Acquire loads from before an Acquire operation to after it.

Notes for code generation Architectures with weak memory ordering (essentially everything relevant today except x86 and SPARC) require some sort of fence to maintain the Acquire semantics. The precise fences required varies widely by architecture, but for a simple implementation, most architectures provide a barrier which is strong enough for everything (`dmb` on ARM, `sync` on PowerPC, etc.). Putting such a fence after the equivalent Monotonic operation is sufficient to maintain Acquire semantics for a memory operation.

Release

Release is similar to Acquire, but with a barrier of the sort necessary to release a lock.

Relevant standard This corresponds to the C++11/C11 `memory_order_release`.

Notes for frontends If you are writing a frontend which uses this directly, use with caution. Release only provides a semantic guarantee when paired with a Acquire operation.

Notes for optimizers Optimizers not aware of atomics can treat this like a nothrow call. It is also possible to move loads from after a Release store or read-modify-write operation to before it, and move non-Release stores from after an Release operation to before it.

Notes for code generation See the section on Acquire; a fence before the relevant operation is usually sufficient for Release. Note that a store-store fence is not sufficient to implement Release semantics; store-store fences are generally not exposed to IR because they are extremely difficult to use correctly.

AcquireRelease

AcquireRelease (`acq_rel` in IR) provides both an Acquire and a Release barrier (for fences and operations which both read and write memory).

Relevant standard This corresponds to the C++11/C11 `memory_order_acq_rel`.

Notes for frontends If you are writing a frontend which uses this directly, use with caution. Acquire only provides a semantic guarantee when paired with a Release operation, and vice versa.

Notes for optimizers In general, optimizers should treat this like a nothrow call; the possible optimizations are usually not interesting.

Notes for code generation This operation has Acquire and Release semantics; see the sections on Acquire and Release.

SequentiallyConsistent

SequentiallyConsistent (`seq_cst` in IR) provides Acquire semantics for loads and Release semantics for stores. Additionally, it guarantees that a total ordering exists between all SequentiallyConsistent operations.

Relevant standard This corresponds to the C++11/C11 `memory_order_seq_cst`, Java `volatile`, and the gcc-compatible `__sync_*` builtins which do not specify otherwise.

Notes for frontends If a frontend is exposing atomic operations, these are much easier to reason about for the programmer than other kinds of operations, and using them is generally a practical performance tradeoff.

Notes for optimizers Optimizers not aware of atomics can treat this like a nothrow call. For SequentiallyConsistent loads and stores, the same reorderings are allowed as for Acquire loads and Release stores, except that SequentiallyConsistent operations may not be reordered.

Notes for code generation SequentiallyConsistent loads minimally require the same barriers as Acquire operations and SequentiallyConsistent stores require Release barriers. Additionally, the code generator must enforce ordering between SequentiallyConsistent stores followed by SequentiallyConsistent loads. This is usually done by emitting either a full fence before the loads or a full fence after the stores; which is preferred varies by architecture.

3.1.5 Atomics and IR optimization

Predicates for optimizer writers to query:

- `isSimple()`: A load or store which is not volatile or atomic. This is what, for example, `memcpyopt` would check for operations it might transform.
- `isUnordered()`: A load or store which is not volatile and at most Unordered. This would be checked, for example, by LICM before hoisting an operation.
- `mayReadFromMemory()/mayWriteToMemory()`: Existing predicate, but note that they return true for any operation which is volatile or at least Monotonic.
- `isStrongerThan / isAtLeastOrStrongerThan`: These are predicates on orderings. They can be useful for passes that are aware of atomics, for example to do DSE across a single atomic access, but not across a release-acquire pair (see `MemoryDependencyAnalysis` for an example of this)
- Alias analysis: Note that AA will return `ModRef` for anything Acquire or Release, and for the address accessed by any Monotonic operation.

To support optimizing around atomic operations, make sure you are using the right predicates; everything should work if that is done. If your pass should optimize some atomic operations (Unordered operations in particular), make sure it doesn't replace an atomic load or store with a non-atomic operation.

Some examples of how optimizations interact with various kinds of atomic operations:

- `memcpyopt`: An atomic operation cannot be optimized into part of a `memcpy`/`memset`, including unordered loads/stores. It can pull operations across some atomic operations.
- LICM: Unordered loads/stores can be moved out of a loop. It just treats monotonic operations like a read+write to a memory location, and anything stricter than that like a nothrow call.
- DSE: Unordered stores can be DSE'ed like normal stores. Monotonic stores can be DSE'ed in some cases, but it's tricky to reason about, and not especially important. It is possible in some case for DSE to operate across a stronger atomic operation, but it is fairly tricky. DSE delegates this reasoning to `MemoryDependencyAnalysis` (which is also used by other passes like GVN).
- Folding a load: Any atomic load from a constant global can be constant-folded, because it cannot be observed. Similar reasoning allows `sroa` with atomic loads and stores.

3.1.6 Atomics and Codegen

Atomic operations are represented in the SelectionDAG with `ATOMIC_*` opcodes. On architectures which use barrier instructions for all atomic ordering (like ARM), appropriate fences can be emitted by the AtomicExpand Codegen pass if `setInsertFencesForAtomic()` was used.

The `MachineMemOperand` for all atomic operations is currently marked as volatile; this is not correct in the IR sense of volatile, but CodeGen handles anything marked volatile very conservatively. This should get fixed at some point.

One very important property of the atomic operations is that if your backend supports any inline lock-free atomic operations of a given size, you should support *ALL* operations of that size in a lock-free manner.

When the target implements atomic `cmpxchg` or LL/SC instructions (as most do) this is trivial: all the other operations can be implemented on top of those primitives. However, on many older CPUs (e.g. ARMv5, SparcV8, Intel 80386) there are atomic load and store instructions, but no `cmpxchg` or LL/SC. As it is invalid to implement `atomic load` using the native instruction, but `cmpxchg` using a library call to a function that uses a mutex, `atomic load` must *also* expand to a library call on such architectures, so that it can remain atomic with regards to a simultaneous `cmpxchg`, by using the same mutex.

`AtomicExpandPass` can help with that: it will expand all atomic operations to the proper `__atomic_*` libcalls for any size above the maximum set by `setMaxAtomicSizeInBitsSupported` (which defaults to 0).

On x86, all atomic loads generate a `MOV`. SequentiallyConsistent stores generate an `XCHG`, other stores generate a `MOV`. SequentiallyConsistent fences generate an `MFENCE`, other fences do not cause any code to be generated. `cmpxchg` uses the `LOCK CMPXCHG` instruction. `atomicrmw xchg` uses `XCHG`, `atomicrmw add` and `atomicrmw sub` use `XADD`, and all other `atomicrmw` operations generate a loop with `LOCK CMPXCHG`. Depending on the users of the result, some `atomicrmw` operations can be translated into operations like `LOCK AND`, but that does not work in general.

On ARM (before v8), MIPS, and many other RISC architectures, Acquire, Release, and SequentiallyConsistent semantics require barrier instructions for every such operation. Loads and stores generate normal instructions. `cmpxchg` and `atomicrmw` can be represented using a loop with LL/SC-style instructions which take some sort of exclusive lock on a cache line (`LDREX` and `STREX` on ARM, etc.).

It is often easiest for backends to use `AtomicExpandPass` to lower some of the atomic constructs. Here are some lowerings it can do:

- `cmpxchg` -> loop with load-linked/store-conditional by overriding `shouldExpandAtomicCmpXchgInIR()`, `emitLoadLinked()`, `emitStoreConditional()`
- large loads/stores -> ll-sc/`cmpxchg` by overriding `shouldExpandAtomicStoreInIR()/shouldExpandAtomicLoadInIR()`
- strong atomic accesses -> monotonic accesses + fences by overriding `shouldInsertFencesForAtomic()`, `emitLeadingFence()`, and `emitTrailingFence()`
- atomic rmw -> loop with `cmpxchg` or load-linked/store-conditional by overriding `expandAtomicRMWInIR()`
- expansion to `__atomic_*` libcalls for unsupported sizes.

For an example of all of these, look at the ARM backend.

3.1.7 Libcalls: `__atomic_*`

There are two kinds of atomic library calls that are generated by LLVM. Please note that both sets of library functions somewhat confusingly share the names of builtin functions defined by clang. Despite this, the library functions are not directly related to the builtins: it is *not* the case that `__atomic_*` builtins lower to `__atomic_*` library calls and `__sync_*` builtins lower to `__sync_*` library calls.

The first set of library functions are named `__atomic_*`. This set has been “standardized” by GCC, and is described below. (See also [GCC’s documentation](#))

LLVM’s `AtomicExpandPass` will translate atomic operations on data sizes above `MaxAtomicSizeInBitsSupported` into calls to these functions.

There are four generic functions, which can be called with data of any size or alignment:

```
void __atomic_load(size_t size, void *ptr, void *ret, int ordering)
void __atomic_store(size_t size, void *ptr, void *val, int ordering)
void __atomic_exchange(size_t size, void *ptr, void *val, void *ret, int ordering)
bool __atomic_compare_exchange(size_t size, void *ptr, void *expected, void *desired,
    ↪ int success_order, int failure_order)
```

There are also size-specialized versions of the above functions, which can only be used with *naturally-aligned* pointers of the appropriate size. In the signatures below, “N” is one of 1, 2, 4, 8, and 16, and “iN” is the appropriate integer type of that size; if no such integer type exists, the specialization cannot be used:

```
iN __atomic_load_N(iN *ptr, iN val, int ordering)
void __atomic_store_N(iN *ptr, iN val, int ordering)
iN __atomic_exchange_N(iN *ptr, iN val, int ordering)
bool __atomic_compare_exchange_N(iN *ptr, iN *expected, iN desired, int success_order,
    ↪ int failure_order)
```

Finally there are some read-modify-write functions, which are only available in the size-specific variants (any other sizes use a `__atomic_compare_exchange` loop):

```
iN __atomic_fetch_add_N(iN *ptr, iN val, int ordering)
iN __atomic_fetch_sub_N(iN *ptr, iN val, int ordering)
iN __atomic_fetch_and_N(iN *ptr, iN val, int ordering)
iN __atomic_fetch_or_N(iN *ptr, iN val, int ordering)
iN __atomic_fetch_xor_N(iN *ptr, iN val, int ordering)
iN __atomic_fetch_nand_N(iN *ptr, iN val, int ordering)
```

This set of library functions have some interesting implementation requirements to take note of:

- They support all sizes and alignments – including those which cannot be implemented natively on any existing hardware. Therefore, they will certainly use mutexes in for some sizes/alignments.
- As a consequence, they cannot be shipped in a statically linked compiler-support library, as they have state which must be shared amongst all DSOs loaded in the program. They must be provided in a shared library used by all objects.
- The set of atomic sizes supported lock-free must be a superset of the sizes any compiler can emit. That is: if a new compiler introduces support for inline-lock-free atomics of size N, the `__atomic_*` functions must also have a lock-free implementation for size N. This is a requirement so that code produced by an old compiler (which will have called the `__atomic_*` function) interoperates with code produced by the new compiler (which will use native the atomic instruction).

Note that it’s possible to write an entirely target-independent implementation of these library functions by using the compiler atomic builtins themselves to implement the operations on naturally-aligned pointers of supported sizes, and a generic mutex implementation otherwise.

3.1.8 Libcalls: `__sync_*`

Some targets or OS/target combinations can support lock-free atomics, but for various reasons, it is not practical to emit the instructions inline.

There’s two typical examples of this.

Some CPUs support multiple instruction sets which can be switched back and forth on function-call boundaries. For example, MIPS supports the MIPS16 ISA, which has a smaller instruction encoding than the usual MIPS32 ISA. ARM, similarly, has the Thumb ISA. In MIPS16 and earlier versions of Thumb, the atomic instructions are not encodable. However, those instructions are available via a function call to a function with the longer encoding.

Additionally, a few OS/target pairs provide kernel-supported lock-free atomics. ARM/Linux is an example of this: the kernel [provides](#) a function which on older CPUs contains a “magically-restartable” atomic sequence (which looks atomic so long as there’s only one CPU), and contains actual atomic instructions on newer multicore models. This sort of functionality can typically be provided on any architecture, if all CPUs which are missing atomic compare-and-swap support are uniprocessor (no SMP). This is almost always the case. The only common architecture without that property is SPARC – SPARCV8 SMP systems were common, yet it doesn’t support any sort of compare-and-swap operation.

In either of these cases, the Target in LLVM can claim support for atomics of an appropriate size, and then implement some subset of the operations via libcalls to a `__sync_*` function. Such functions *must* not use locks in their implementation, because unlike the `__atomic_*` routines used by AtomicExpandPass, these may be mixed-and-matched with native instructions by the target lowering.

Further, these routines do not need to be shared, as they are stateless. So, there is no issue with having multiple copies included in one binary. Thus, typically these routines are implemented by the statically-linked compiler runtime support library.

LLVM will emit a call to an appropriate `__sync_*` routine if the target ISelLowering code has set the corresponding `ATOMIC_CMPXCHG`, `ATOMIC_SWAP`, or `ATOMIC_LOAD_*` operation to “Expand”, and if it has opted-into the availability of those library functions via a call to `initSyncLibcalls()`.

The full set of functions that may be called by LLVM is (for N being 1, 2, 4, 8, or 16):

```
iN __sync_val_compare_and_swap_N(iN *ptr, iN expected, iN desired)
iN __sync_lock_test_and_set_N(iN *ptr, iN val)
iN __sync_fetch_and_add_N(iN *ptr, iN val)
iN __sync_fetch_and_sub_N(iN *ptr, iN val)
iN __sync_fetch_and_and_N(iN *ptr, iN val)
iN __sync_fetch_and_or_N(iN *ptr, iN val)
iN __sync_fetch_and_xor_N(iN *ptr, iN val)
iN __sync_fetch_and_nand_N(iN *ptr, iN val)
iN __sync_fetch_and_max_N(iN *ptr, iN val)
iN __sync_fetch_and_umax_N(iN *ptr, iN val)
iN __sync_fetch_and_min_N(iN *ptr, iN val)
iN __sync_fetch_and_umin_N(iN *ptr, iN val)
```

This list doesn’t include any function for atomic load or store; all known architectures support atomic loads and stores directly (possibly by emitting a fence on either side of a normal load or store.)

There’s also, somewhat separately, the possibility to lower `ATOMIC_FENCE` to `__sync_synchronize()`. This may happen or not happen independent of all the above, controlled purely by `setOperationAction(ISD::ATOMIC_FENCE, ...)`.

3.2 LLVM Coding Standards

- *Introduction*
- *Languages, Libraries, and Standards*
 - *C++ Standard Versions*

- *C++ Standard Library*
- *Supported C++11 Language and Library Features*
- *Other Languages*
- *Mechanical Source Issues*
 - *Source Code Formatting*
 - * *Commenting*
 - *File Headers*
 - *Class overviews*
 - *Method information*
 - * *Comment Formatting*
 - * *Doxygen Use in Documentation Comments*
 - * *#include Style*
 - * *Source Code Width*
 - * *Use Spaces Instead of Tabs*
 - * *Indent Code Consistently*
 - *Format Lambdas Like Blocks Of Code*
 - *Braced Initializer Lists*
 - *Language and Compiler Issues*
 - * *Treat Compiler Warnings Like Errors*
 - * *Write Portable Code*
 - * *Do not use RTTI or Exceptions*
 - * *Do not use Static Constructors*
 - * *Use of class and struct Keywords*
 - * *Do not use Braced Initializer Lists to Call a Constructor*
 - * *Use auto Type Deduction to Make Code More Readable*
 - * *Beware unnecessary copies with auto*
- *Style Issues*
 - *The High-Level Issues*
 - * *A Public Header File is a Module*
 - * *#include as Little as Possible*
 - * *Keep “Internal” Headers Private*
 - * *Use Early Exits and continue to Simplify Code*
 - * *Don’t use else after a return*
 - * *Turn Predicate Loops into Predicate Functions*
 - *The Low-Level Issues*

- * *Name Types, Functions, Variables, and Enumerators Properly*
- * *Assert Liberally*
- * *Do Not Use `using namespace std`*
- * *Provide a Virtual Method Anchor for Classes in Headers*
- * *Don't use default labels in fully covered switches over enumerations*
- * *Don't evaluate `end()` every time through a loop*
- * *`#include <iostream>` is Forbidden*
- * *Use `raw_ostream`*
- * *Avoid `std::endl`*
- * *Don't use `inline` when defining a function in a class definition*
- *Microscopic Details*
 - * *Spaces Before Parentheses*
 - * *Prefer Preincrement*
 - * *Namespace Indentation*
 - * *Anonymous Namespaces*

- *See Also*

3.2.1 Introduction

This document attempts to describe a few coding standards that are being used in the LLVM source tree. Although no coding standards should be regarded as absolute requirements to be followed in all instances, coding standards are particularly important for large-scale code bases that follow a library-based design (like LLVM).

While this document may provide guidance for some mechanical formatting issues, whitespace, or other “microscopic details”, these are not fixed standards. Always follow the golden rule:

If you are extending, enhancing, or bug fixing already implemented code, use the style that is already being used so that the source is uniform and easy to follow.

Note that some code bases (e.g. `libc++`) have really good reasons to deviate from the coding standards. In the case of `libc++`, this is because the naming and other conventions are dictated by the C++ standard. If you think there is a specific good reason to deviate from the standards here, please bring it up on the LLVM-dev mailing list.

There are some conventions that are not uniformly followed in the code base (e.g. the naming convention). This is because they are relatively new, and a lot of code was written before they were put in place. Our long term goal is for the entire codebase to follow the convention, but we explicitly *do not* want patches that do large-scale reformatting of existing code. On the other hand, it is reasonable to rename the methods of a class if you're about to change it in some other way. Just do the reformatting as a separate commit from the functionality change.

The ultimate goal of these guidelines is to increase the readability and maintainability of our common source base. If you have suggestions for topics to be included, please mail them to [Chris](#).

3.2.2 Languages, Libraries, and Standards

Most source code in LLVM and other LLVM projects using these coding standards is C++ code. There are some places where C code is used either due to environment restrictions, historical restrictions, or due to third-party source

code imported into the tree. Generally, our preference is for standards conforming, modern, and portable C++ code as the implementation language of choice.

C++ Standard Versions

LLVM, Clang, and LLD are currently written using C++11 conforming code, although we restrict ourselves to features which are available in the major toolchains supported as host compilers. The LLDB project is even more aggressive in the set of host compilers supported and thus uses still more features. Regardless of the supported features, code is expected to (when reasonable) be standard, portable, and modern C++11 code. We avoid unnecessary vendor-specific extensions, etc.

C++ Standard Library

Use the C++ standard library facilities whenever they are available for a particular task. LLVM and related projects emphasize and rely on the standard library facilities for as much as possible. Common support libraries providing functionality missing from the standard library for which there are standard interfaces or active work on adding standard interfaces will often be implemented in the LLVM namespace following the expected standard interface.

There are some exceptions such as the standard I/O streams library which are avoided. Also, there is much more detailed information on these subjects in the *LLVM Programmer's Manual*.

Supported C++11 Language and Library Features

While LLVM, Clang, and LLD use C++11, not all features are available in all of the toolchains which we support. The set of features supported for use in LLVM is the intersection of those supported in the minimum requirements described in the *Getting Started with the LLVM System* page, section *Software*. The ultimate definition of this set is what build bots with those respective toolchains accept. Don't argue with the build bots. However, we have some guidance below to help you know what to expect.

Each toolchain provides a good reference for what it accepts:

- Clang: http://clang.llvm.org/cxx_status.html
- GCC: <http://gcc.gnu.org/projects/cxx0x.html>
- MSVC: <http://msdn.microsoft.com/en-us/library/hh567368.aspx>

In most cases, the MSVC list will be the dominating factor. Here is a summary of the features that are expected to work. Features not on this list are unlikely to be supported by our host compilers.

- Rvalue references: [N2118](#)
 - But *not* Rvalue references for `*this` or member qualifiers ([N2439](#))
- Static assert: [N1720](#)
- `auto` type deduction: [N1984](#), [N1737](#)
- Trailing return types: [N2541](#)
- Lambdas: [N2927](#)
 - But *not* lambdas with default arguments.
- `decltype`: [N2343](#)
- Nested closing right angle brackets: [N1757](#)
- Extern templates: [N1987](#)

- `nullptr`: [N2431](#)
- Strongly-typed and forward declarable enums: [N2347](#), [N2764](#)
- Local and unnamed types as template arguments: [N2657](#)
- Range-based for-loop: [N2930](#)
 - But `{}` are required around inner `do {} while()` loops. As a result, `{}` are required around function-like macros inside range-based for loops.
- `override` and `final`: [N2928](#), [N3206](#), [N3272](#)
- Atomic operations and the C++11 memory model: [N2429](#)
- Variadic templates: [N2242](#)
- Explicit conversion operators: [N2437](#)
- Defaulted and deleted functions: [N2346](#)
- Initializer lists: [N2627](#)
- Delegating constructors: [N1986](#)
- Default member initializers (non-static data member initializers): [N2756](#)
 - Feel free to use these wherever they make sense and where the `=` syntax is allowed. Don't use braced initialization syntax.

The supported features in the C++11 standard libraries are less well tracked, but also much greater. Most of the standard libraries implement most of C++11's library. The most likely lowest common denominator is Linux support. For `libc++`, the support is just poorly tested and undocumented but expected to be largely complete. YMMV. For `libstdc++`, the support is documented in detail in [the libstdc++ manual](#). There are some very minor missing facilities that are unlikely to be common problems, and there are a few larger gaps that are worth being aware of:

- Not all of the type traits are implemented
- No regular expression library.
- While most of the atomics library is well implemented, the fences are missing. Fortunately, they are rarely needed.
- The locale support is incomplete.

Other than these areas you should assume the standard library is available and working as expected until some build bot tells you otherwise. If you're in an uncertain area of one of the above points, but you cannot test on a Linux system, your best approach is to minimize your use of these features, and watch the Linux build bots to find out if your usage triggered a bug. For example, if you hit a type trait which doesn't work we can then add support to LLVM's traits header to emulate it.

Other Languages

Any code written in the Go programming language is not subject to the formatting rules below. Instead, we adopt the formatting rules enforced by the [gofmt](#) tool.

Go code should strive to be idiomatic. Two good sets of guidelines for what this means are [Effective Go](#) and [Go Code Review Comments](#).

3.2.3 Mechanical Source Issues

Source Code Formatting

Commenting

Comments are one critical part of readability and maintainability. Everyone knows they should comment their code, and so should you. When writing comments, write them as English prose, which means they should use proper capitalization, punctuation, etc. Aim to describe what the code is trying to do and why, not *how* it does it at a micro level. Here are a few critical things to document:

File Headers

Every source file should have a header on it that describes the basic purpose of the file. If a file does not have a header, it should not be checked into the tree. The standard header looks like this:

```
//==== llvm/Instruction.h - Instruction class definition -----*- C++ -*-====//
//
//      The LLVM Compiler Infrastructure
//
// This file is distributed under the University of Illinois Open Source
// License. See LICENSE.TXT for details.
//
//=====//
///
/// \file
/// This file contains the declaration of the Instruction class, which is the
/// base class for all of the VM instructions.
///
//=====//
```

A few things to note about this particular format: The “`-*- C++ -*-`” string on the first line is there to tell Emacs that the source file is a C++ file, not a C file (Emacs assumes `.h` files are C files by default).

Note: This tag is not necessary in `.cpp` files. The name of the file is also on the first line, along with a very short description of the purpose of the file. This is important when printing out code and flipping through lots of pages.

The next section in the file is a concise note that defines the license that the file is released under. This makes it perfectly clear what terms the source code can be distributed under and should not be modified in any way.

The main body is a `doxygen` comment (identified by the `///` comment marker instead of the usual `/**`) describing the purpose of the file. The first sentence (or a passage beginning with `\brief`) is used as an abstract. Any additional information should be separated by a blank line. If an algorithm is being implemented or something tricky is going on, a reference to the paper where it is published should be included, as well as any notes or *gotchas* in the code to watch out for.

Class overviews

Classes are one fundamental part of a good object oriented design. As such, a class definition should have a comment block that explains what the class is used for and how it works. Every non-trivial class is expected to have a `doxygen` comment block.

Method information

Methods defined in a class (as well as any global functions) should also be documented properly. A quick note about what it does and a description of the borderline behaviour is all that is necessary here (unless something particularly tricky or insidious is going on). The hope is that people can figure out how to use your interfaces without reading the code itself.

Good things to talk about here are what happens when something unexpected happens: does the method return null? Abort? Format your hard disk?

Comment Formatting

In general, prefer C++ style comments (`//` for normal comments, `///` for doxygen documentation comments). They take less space, require less typing, don't have nesting problems, etc. There are a few cases when it is useful to use C style (`/* */`) comments however:

1. When writing C code: Obviously if you are writing C code, use C style comments.
2. When writing a header file that may be `#included` by a C source file.
3. When writing a source file that is used by a tool that only accepts C style comments.

Commenting out large blocks of code is discouraged, but if you really have to do this (for documentation purposes or as a suggestion for debug printing), use `#if 0` and `#endif`. These nest properly and are better behaved in general than C style comments.

Doxygen Use in Documentation Comments

Use the `\file` command to turn the standard file header into a file-level comment.

Include descriptive paragraphs for all public interfaces (public classes, member and non-member functions). Don't just restate the information that can be inferred from the API name. The first sentence (or a paragraph beginning with `\brief`) is used as an abstract. Try to use a single sentence as the `\brief` adds visual clutter. Put detailed discussion into separate paragraphs.

To refer to parameter names inside a paragraph, use the `\p name` command. Don't use the `\arg name` command since it starts a new paragraph that contains documentation for the parameter.

Wrap non-inline code examples in `\code ... \endcode`.

To document a function parameter, start a new paragraph with the `\param name` command. If the parameter is used as an out or an in/out parameter, use the `\param [out] name` or `\param [in,out] name` command, respectively.

To describe function return value, start a new paragraph with the `\returns` command.

A minimal documentation comment:

```
/// Sets the xzyzy property to \p Baz.
void setXzyzy(bool Baz);
```

A documentation comment that uses all Doxygen features in a preferred way:

```
/// Does foo and bar.
///
/// Does not do foo the usual way if \p Baz is true.
///
/// Typical usage:
```

```
/// \code
///   fooBar(false, "quux", Res);
/// \endcode
///
/// \param Quux kind of foo to do.
/// \param [out] Result filled with bar sequence on foo success.
///
/// \returns true on success.
bool fooBar(bool Baz, StringRef Quux, std::vector<int> &Result);
```

Don't duplicate the documentation comment in the header file and in the implementation file. Put the documentation comments for public APIs into the header file. Documentation comments for private APIs can go to the implementation file. In any case, implementation files can include additional comments (not necessarily in Doxygen markup) to explain implementation details as needed.

Don't duplicate function or class name at the beginning of the comment. For humans it is obvious which function or class is being documented; automatic documentation processing tools are smart enough to bind the comment to the correct declaration.

Wrong:

```
// In Something.h:

/// Something - An abstraction for some complicated thing.
class Something {
public:
    /// fooBar - Does foo and bar.
    void fooBar();
};

// In Something.cpp:

/// fooBar - Does foo and bar.
void Something::fooBar() { ... }
```

Correct:

```
// In Something.h:

/// An abstraction for some complicated thing.
class Something {
public:
    /// Does foo and bar.
    void fooBar();
};

// In Something.cpp:

/// Builds a B-tree in order to do foo. See paper by...
void Something::fooBar() { ... }
```

It is not required to use additional Doxygen features, but sometimes it might be a good idea to do so.

Consider:

- adding comments to any narrow namespace containing a collection of related functions or types;
- using top-level groups to organize a collection of related functions at namespace scope where the grouping is smaller than the namespace;

- using member groups and additional comments attached to member groups to organize within a class.

For example:

```
class Something {
    /// \name Functions that do Foo.
    /// @{
    void fooBar();
    void fooBaz();
    /// @}
    ...
};
```

#include Style

Immediately after the *header file comment* (and include guards if working on a header file), the *minimal list of #includes* required by the file should be listed. We prefer these #includes to be listed in this order:

1. Main Module Header
2. Local/Private Headers
3. LLVM project/subproject headers (clang/..., lldb/..., llvm/..., etc)
4. System #includes

and each category should be sorted lexicographically by the full path.

The *Main Module Header* file applies to .cpp files which implement an interface defined by a .h file. This #include should always be included **first** regardless of where it lives on the file system. By including a header file first in the .cpp files that implement the interfaces, we ensure that the header does not have any hidden dependencies which are not explicitly #included in the header, but should be. It is also a form of documentation in the .cpp file to indicate where the interfaces it implements are defined.

LLVM project and subproject headers should be grouped from most specific to least specific, for the same reasons described above. For example, LLDB depends on both clang and LLVM, and clang depends on LLVM. So an LLDB source file should include lldb headers first, followed by clang headers, followed by llvm headers, to reduce the possibility (for example) of an LLDB header accidentally picking up a missing include due to the previous inclusion of that header in the main source file or some earlier header file. clang should similarly include its own headers before including llvm headers. This rule applies to all LLVM subprojects.

Source Code Width

Write your code to fit within 80 columns of text. This helps those of us who like to print out code and look at your code in an xterm without resizing it.

The longer answer is that there must be some limit to the width of the code in order to reasonably allow developers to have multiple files side-by-side in windows on a modest display. If you are going to pick a width limit, it is somewhat arbitrary but you might as well pick something standard. Going with 90 columns (for example) instead of 80 columns wouldn't add any significant value and would be detrimental to printing out code. Also many other projects have standardized on 80 columns, so some people have already configured their editors for it (vs something else, like 90 columns).

This is one of many contentious issues in coding standards, but it is not up for debate.

Use Spaces Instead of Tabs

In all cases, prefer spaces to tabs in source files. People have different preferred indentation levels, and different styles of indentation that they like; this is fine. What isn't fine is that different editors/viewers expand tabs out to different tab stops. This can cause your code to look completely unreadable, and it is not worth dealing with.

As always, follow the *Golden Rule* above: follow the style of existing code if you are modifying and extending it. If you like four spaces of indentation, **DO NOT** do that in the middle of a chunk of code with two spaces of indentation. Also, do not reindent a whole source file: it makes for incredible diffs that are absolutely worthless.

Indent Code Consistently

Okay, in your first year of programming you were told that indentation is important. If you didn't believe and internalize this then, now is the time. Just do it. With the introduction of C++11, there are some new formatting challenges that merit some suggestions to help have consistent, maintainable, and tool-friendly formatting and indentation.

Format Lambdas Like Blocks Of Code

When formatting a multi-line lambda, format it like a block of code, that's what it is. If there is only one multi-line lambda in a statement, and there are no expressions lexically after it in the statement, drop the indent to the standard two space indent for a block of code, as if it were an if-block opened by the preceding part of the statement:

```
std::sort(foo.begin(), foo.end(), [&](Foo a, Foo b) -> bool {
    if (a.blah < b.blah)
        return true;
    if (a.baz < b.baz)
        return true;
    return a.bam < b.bam;
});
```

To take best advantage of this formatting, if you are designing an API which accepts a continuation or single callable argument (be it a functor, or a `std::function`), it should be the last argument if at all possible.

If there are multiple multi-line lambdas in a statement, or there is anything interesting after the lambda in the statement, indent the block two spaces from the indent of the []:

```
dyn_switch(V->stripPointerCasts(),
    [] (PHINode *PN) {
        // process phis...
    },
    [] (SelectInst *SI) {
        // process selects...
    },
    [] (LoadInst *LI) {
        // process loads...
    },
    [] (AllocaInst *AI) {
        // process allocas...
    });
```

Braced Initializer Lists

With C++11, there are significantly more uses of braced lists to perform initialization. These allow you to easily construct aggregate temporaries in expressions among other niceness. They now have a natural way of ending up nested within each other and within function calls in order to build up aggregates (such as option structs) from local variables. To make matters worse, we also have many more uses of braces in an expression context that are *not* performing initialization.

The historically common formatting of braced initialization of aggregate variables does not mix cleanly with deep nesting, general expression contexts, function arguments, and lambdas. We suggest new code use a simple rule for formatting braced initialization lists: act as-if the braces were parentheses in a function call. The formatting rules exactly match those already well understood for formatting nested function calls. Examples:

```
foo({a, b, c}, {1, 2, 3});

llvm::Constant *Mask[] = {
    llvm::ConstantInt::get(llvm::Type::getInt32Ty(llvm::Context()), 0),
    llvm::ConstantInt::get(llvm::Type::getInt32Ty(llvm::Context()), 1),
    llvm::ConstantInt::get(llvm::Type::getInt32Ty(llvm::Context()), 2)};
```

This formatting scheme also makes it particularly easy to get predictable, consistent, and automatic formatting with tools like [Clang Format](#).

Language and Compiler Issues

Treat Compiler Warnings Like Errors

If your code has compiler warnings in it, something is wrong — you aren’t casting values correctly, you have “questionable” constructs in your code, or you are doing something legitimately wrong. Compiler warnings can cover up legitimate errors in output and make dealing with a translation unit difficult.

It is not possible to prevent all warnings from all compilers, nor is it desirable. Instead, pick a standard compiler (like `gcc`) that provides a good thorough set of warnings, and stick to it. At least in the case of `gcc`, it is possible to work around any spurious errors by changing the syntax of the code slightly. For example, a warning that annoys me occurs when I write code like this:

```
if (V = getValue()) {
    ...
}
```

`gcc` will warn me that I probably want to use the `==` operator, and that I probably mistyped it. In most cases, I haven’t, and I really don’t want the spurious errors. To fix this particular problem, I rewrite the code like this:

```
if ((V = getValue())) {
    ...
}
```

which shuts `gcc` up. Any `gcc` warning that annoys you can be fixed by massaging the code appropriately.

Write Portable Code

In almost all cases, it is possible and within reason to write completely portable code. If there are cases where it isn’t possible to write portable code, isolate it behind a well defined (and well documented) interface.

In practice, this means that you shouldn't assume much about the host compiler (and Visual Studio tends to be the lowest common denominator). If advanced features are used, they should only be an implementation detail of a library which has a simple exposed API, and preferably be buried in `libSystem`.

Do not use RTTI or Exceptions

In an effort to reduce code and executable size, LLVM does not use RTTI (e.g. `dynamic_cast<>;`) or exceptions. These two language features violate the general C++ principle of “*you only pay for what you use*”, causing executable bloat even if exceptions are never used in the code base, or if RTTI is never used for a class. Because of this, we turn them off globally in the code.

That said, LLVM does make extensive use of a hand-rolled form of RTTI that use templates like `isa<>`, `cast<>`, and `dyn_cast<>`. This form of RTTI is opt-in and can be *added to any class*. It is also substantially more efficient than `dynamic_cast<>`.

Do not use Static Constructors

Static constructors and destructors (e.g. global variables whose types have a constructor or destructor) should not be added to the code base, and should be removed wherever possible. Besides [well known problems](#) where the order of initialization is undefined between globals in different source files, the entire concept of static constructors is at odds with the common use case of LLVM as a library linked into a larger application.

Consider the use of LLVM as a JIT linked into another application (perhaps for [OpenGL](#), [custom languages](#), [shaders in movies](#), etc). Due to the design of static constructors, they must be executed at startup time of the entire application, regardless of whether or how LLVM is used in that larger application. There are two problems with this:

- The time to run the static constructors impacts startup time of applications — a critical time for GUI apps, among others.
- The static constructors cause the app to pull many extra pages of memory off the disk: both the code for the constructor in each `.o` file and the small amount of data that gets touched. In addition, touched/dirty pages put more pressure on the VM system on low-memory machines.

We would really like for there to be zero cost for linking in an additional LLVM target or other library into an application, but static constructors violate this goal.

That said, LLVM unfortunately does contain static constructors. It would be a [great project](#) for someone to purge all static constructors from LLVM, and then enable the `-Wglobal-constructors` warning flag (when building with Clang) to ensure we do not regress in the future.

Use of `class` and `struct` Keywords

In C++, the `class` and `struct` keywords can be used almost interchangeably. The only difference is when they are used to declare a class: `class` makes all members private by default while `struct` makes all members public by default.

Unfortunately, not all compilers follow the rules and some will generate different symbols based on whether `class` or `struct` was used to declare the symbol (e.g., MSVC). This can lead to problems at link time.

- All declarations and definitions of a given `class` or `struct` must use the same keyword. For example:

```
class Foo;

// Breaks mangling in MSVC.
struct Foo { int Data; };
```


- As a rule of thumb, `struct` should be kept to structures where *all* members are declared public.

```
// Foo feels like a class... this is strange.
struct Foo {
private:
    int Data;
public:
    Foo() : Data(0) { }
    int getData() const { return Data; }
    void setData(int D) { Data = D; }
};

// Bar isn't POD, but it does look like a struct.
struct Bar {
    int Data;
    Bar() : Data(0) { }
};
```

Do not use Braced Initializer Lists to Call a Constructor

In C++11 there is a “generalized initialization syntax” which allows calling constructors using braced initializer lists. Do not use these to call constructors with any interesting logic or if you care that you’re calling some *particular* constructor. Those should look like function calls using parentheses rather than like aggregate initialization. Similarly, if you need to explicitly name the type and call its constructor to create a temporary, don’t use a braced initializer list. Instead, use a braced initializer list (without any type for temporaries) when doing aggregate initialization or something notionally equivalent. Examples:

```
class Foo {
public:
    // Construct a Foo by reading data from the disk in the whizbang format, ...
    Foo(std::string filename);

    // Construct a Foo by looking up the Nth element of some global data ...
    Foo(int N);

    // ...
};

// The Foo constructor call is very deliberate, no braces.
std::fill(foo.begin(), foo.end(), Foo("name"));

// The pair is just being constructed like an aggregate, use braces.
bar_map.insert({my_key, my_value});
```

If you use a braced initializer list when initializing a variable, use an equals before the open curly brace:

```
int data[] = {0, 1, 2, 3};
```

Use auto Type Deduction to Make Code More Readable

Some are advocating a policy of “almost always `auto`” in C++11, however LLVM uses a more moderate stance. Use `auto` if and only if it makes the code more readable or easier to maintain. Don’t “almost always” use `auto`, but do use `auto` with initializers like `cast<Foo>(...)` or other places where the type is already obvious from the

context. Another time when `auto` works well for these purposes is when the type would have been abstracted away anyways, often behind a container's typedef such as `std::vector<T>::iterator`.

Beware unnecessary copies with `auto`

The convenience of `auto` makes it easy to forget that its default behavior is a copy. Particularly in range-based `for` loops, careless copies are expensive.

As a rule of thumb, use `auto &` unless you need to copy the result, and use `auto *` when copying pointers.

```
// Typically there's no reason to copy.
for (const auto &Val : Container) { observe(Val); }
for (auto &Val : Container) { Val.change(); }

// Remove the reference if you really want a new copy.
for (auto Val : Container) { Val.change(); saveSomewhere(Val); }

// Copy pointers, but make it clear that they're pointers.
for (const auto *Ptr : Container) { observe(*Ptr); }
for (auto *Ptr : Container) { Ptr->change(); }
```

3.2.4 Style Issues

The High-Level Issues

A Public Header File is a Module

C++ doesn't do too well in the modularity department. There is no real encapsulation or data hiding (unless you use expensive protocol classes), but it is what we have to work with. When you write a public header file (in the LLVM source tree, they live in the top level "include" directory), you are defining a module of functionality.

Ideally, modules should be completely independent of each other, and their header files should only `#include` the absolute minimum number of headers possible. A module is not just a class, a function, or a namespace: it's a collection of these that defines an interface. This interface may be several functions, classes, or data structures, but the important issue is how they work together.

In general, a module should be implemented by one or more `.cpp` files. Each of these `.cpp` files should include the header that defines their interface first. This ensures that all of the dependences of the module header have been properly added to the module header itself, and are not implicit. System headers should be included after user headers for a translation unit.

`#include` as Little as Possible

`#include` hurts compile time performance. Don't do it unless you have to, especially in header files.

But wait! Sometimes you need to have the definition of a class to use it, or to inherit from it. In these cases go ahead and `#include` that header file. Be aware however that there are many cases where you don't need to have the full definition of a class. If you are using a pointer or reference to a class, you don't need the header file. If you are simply returning a class instance from a prototyped function or method, you don't need it. In fact, for most cases, you simply don't need the definition of a class. And not `#include`ing speeds up compilation.

It is easy to try to go too overboard on this recommendation, however. You **must** include all of the header files that you are using — you can include them either directly or indirectly through another header file. To make sure that you don't accidentally forget to include a header file in your module header, make sure to include your module header **first**

in the implementation file (as mentioned above). This way there won't be any hidden dependencies that you'll find out about later.

Keep “Internal” Headers Private

Many modules have a complex implementation that causes them to use more than one implementation (.cpp) file. It is often tempting to put the internal communication interface (helper classes, extra functions, etc) in the public module header file. Don't do this!

If you really need to do something like this, put a private header file in the same directory as the source files, and include it locally. This ensures that your private interface remains private and undisturbed by outsiders.

Note: It's okay to put extra implementation methods in a public class itself. Just make them private (or protected) and all is well.

Use Early Exits and `continue` to Simplify Code

When reading code, keep in mind how much state and how many previous decisions have to be remembered by the reader to understand a block of code. Aim to reduce indentation where possible when it doesn't make it more difficult to understand the code. One great way to do this is by making use of early exits and the `continue` keyword in long loops. As an example of using an early exit from a function, consider this “bad” code:

```
Value *doSomething(Instruction *I) {
    if (!isa<TerminatorInst>(I) &&
        I->hasOneUse() && doOtherThing(I)) {
        ... some long code ....
    }

    return 0;
}
```

This code has several problems if the body of the 'if' is large. When you're looking at the top of the function, it isn't immediately clear that this *only* does interesting things with non-terminator instructions, and only applies to things with the other predicates. Second, it is relatively difficult to describe (in comments) why these predicates are important because the `if` statement makes it difficult to lay out the comments. Third, when you're deep within the body of the code, it is indented an extra level. Finally, when reading the top of the function, it isn't clear what the result is if the predicate isn't true; you have to read to the end of the function to know that it returns null.

It is much preferred to format the code like this:

```
Value *doSomething(Instruction *I) {
    // Terminators never need 'something' done to them because ...
    if (isa<TerminatorInst>(I))
        return 0;

    // We conservatively avoid transforming instructions with multiple uses
    // because goats like cheese.
    if (!I->hasOneUse())
        return 0;

    // This is really just here for example.
    if (!doOtherThing(I))
        return 0;
}
```

```
... some long code ....  
}
```

This fixes these problems. A similar problem frequently happens in `for` loops. A silly example is something like this:

```
for (BasicBlock::iterator II = BB->begin(), E = BB->end(); II != E; ++II) {  
    if (BinaryOperator *BO = dyn_cast<BinaryOperator>(II)) {  
        Value *LHS = BO->getOperand(0);  
        Value *RHS = BO->getOperand(1);  
        if (LHS != RHS) {  
            ...  
        }  
    }  
}
```

When you have very, very small loops, this sort of structure is fine. But if it exceeds more than 10-15 lines, it becomes difficult for people to read and understand at a glance. The problem with this sort of code is that it gets very nested very quickly. Meaning that the reader of the code has to keep a lot of context in their brain to remember what is going immediately on in the loop, because they don't know if/when the `if` conditions will have `elses` etc. It is strongly preferred to structure the loop like this:

```
for (BasicBlock::iterator II = BB->begin(), E = BB->end(); II != E; ++II) {  
    BinaryOperator *BO = dyn_cast<BinaryOperator>(II);  
    if (!BO) continue;  
  
    Value *LHS = BO->getOperand(0);  
    Value *RHS = BO->getOperand(1);  
    if (LHS == RHS) continue;  
  
    ...  
}
```

This has all the benefits of using early exits for functions: it reduces nesting of the loop, it makes it easier to describe why the conditions are true, and it makes it obvious to the reader that there is no `else` coming up that they have to push context into their brain for. If a loop is large, this can be a big understandability win.

Don't use `else` after a `return`

For similar reasons above (reduction of indentation and easier reading), please do not use `'else'` or `'else if'` after something that interrupts control flow — like `return`, `break`, `continue`, `goto`, etc. For example, this is *bad*:

```
case 'J': {  
    if (Signed) {  
        Type = Context.getsigjmp_bufType();  
        if (Type.isNull()) {  
            Error = ASTContext::GE_Missing_sigjmp_buf;  
            return QualType();  
        } else {  
            break;  
        }  
    } else {  
        Type = Context.getjmp_bufType();  
        if (Type.isNull()) {  
            Error = ASTContext::GE_Missing_jmp_buf;  
        }  
    }  
}
```

```

    return QualType();
  } else {
    break;
  }
}

```

It is better to write it like this:

```

case 'J':
  if (Signed) {
    Type = Context.getsigjmp_bufType();
    if (Type.isNull()) {
      Error = ASTContext::GE_Missing_sigjmp_buf;
      return QualType();
    }
  } else {
    Type = Context.getjmp_bufType();
    if (Type.isNull()) {
      Error = ASTContext::GE_Missing_jmp_buf;
      return QualType();
    }
  }
  break;

```

Or better yet (in this case) as:

```

case 'J':
  if (Signed)
    Type = Context.getsigjmp_bufType();
  else
    Type = Context.getjmp_bufType();

  if (Type.isNull()) {
    Error = Signed ? ASTContext::GE_Missing_sigjmp_buf :
                  ASTContext::GE_Missing_jmp_buf;
    return QualType();
  }
  break;

```

The idea is to reduce indentation and the amount of code you have to keep track of when reading the code.

Turn Predicate Loops into Predicate Functions

It is very common to write small loops that just compute a boolean value. There are a number of ways that people commonly write these, but an example of this sort of thing is:

```

bool FoundFoo = false;
for (unsigned I = 0, E = BarList.size(); I != E; ++I)
  if (BarList[I]->isFoo()) {
    FoundFoo = true;
    break;
  }

if (FoundFoo) {

```

```
...  
}
```

This sort of code is awkward to write, and is almost always a bad sign. Instead of this sort of loop, we strongly prefer to use a predicate function (which may be *static*) that uses *early exits* to compute the predicate. We prefer the code to be structured like this:

```
/// \returns true if the specified list has an element that is a foo.  
static bool containsFoo(const std::vector<Bar*> &List) {  
    for (unsigned I = 0, E = List.size(); I != E; ++I)  
        if (List[I]->isFoo())  
            return true;  
    return false;  
}  
...  
  
if (containsFoo(BarList)) {  
    ...  
}
```

There are many reasons for doing this: it reduces indentation and factors out code which can often be shared by other code that checks for the same predicate. More importantly, it *forces you to pick a name* for the function, and forces you to write a comment for it. In this silly example, this doesn't add much value. However, if the condition is complex, this can make it a lot easier for the reader to understand the code that queries for this predicate. Instead of being faced with the in-line details of how we check to see if the BarList contains a foo, we can trust the function name and continue reading with better locality.

The Low-Level Issues

Name Types, Functions, Variables, and Enumerators Properly

Poorly-chosen names can mislead the reader and cause bugs. We cannot stress enough how important it is to use *descriptive* names. Pick names that match the semantics and role of the underlying entities, within reason. Avoid abbreviations unless they are well known. After picking a good name, make sure to use consistent capitalization for the name, as inconsistency requires clients to either memorize the APIs or to look it up to find the exact spelling.

In general, names should be in camel case (e.g. `TextFileReader` and `isLValue()`). Different kinds of declarations have different rules:

- **Type names** (including classes, structs, enums, typedefs, etc) should be nouns and start with an upper-case letter (e.g. `TextFileReader`).
- **Variable names** should be nouns (as they represent state). The name should be camel case, and start with an upper case letter (e.g. `Leader` or `Boats`).
- **Function names** should be verb phrases (as they represent actions), and command-like function should be imperative. The name should be camel case, and start with a lower case letter (e.g. `openFile()` or `isFoo()`).
- **Enum declarations** (e.g. `enum Foo { ... }`) are types, so they should follow the naming conventions for types. A common use for enums is as a discriminator for a union, or an indicator of a subclass. When an enum is used for something like this, it should have a `Kind` suffix (e.g. `ValueKind`).
- **Enumerators** (e.g. `enum { Foo, Bar }`) and **public member variables** should start with an upper-case letter, just like types. Unless the enumerators are defined in their own small namespace or inside a class, enumerators should have a prefix corresponding to the enum declaration name. For example, `enum ValueKind { ... }`; may contain enumerators like `VK_Argument`, `VK_BasicBlock`, etc. Enumerators that are just convenience constants are exempt from the requirement for a prefix. For instance:

```
enum {
    MaxSize = 42,
    Density = 12
};
```

As an exception, classes that mimic STL classes can have member names in STL’s style of lower-case words separated by underscores (e.g. `begin()`, `push_back()`, and `empty()`). Classes that provide multiple iterators should add a singular prefix to `begin()` and `end()` (e.g. `global_begin()` and `use_begin()`).

Here are some examples of good and bad names:

```
class VehicleMaker {
    ...
    Factory<Tire> F;           // Bad -- abbreviation and non-descriptive.
    Factory<Tire> Factory;     // Better.
    Factory<Tire> TireFactory; // Even better -- if VehicleMaker has more than one
                             // kind of factories.
};

Vehicle makeVehicle(VehicleType Type) {
    VehicleMaker M;           // Might be OK if having a short life-span.
    Tire Tmp1 = M.makeTire(); // Bad -- 'Tmp1' provides no information.
    Light Headlight = M.makeLight("head"); // Good -- descriptive.
    ...
}
```

Assert Liberally

Use the “`assert`” macro to its fullest. Check all of your preconditions and assumptions, you never know when a bug (not necessarily even yours) might be caught early by an assertion, which reduces debugging time dramatically. The “`<cassert>`” header file is probably already included by the header files you are using, so it doesn’t cost anything to use it.

To further assist with debugging, make sure to put some kind of error message in the assertion statement, which is printed if the assertion is tripped. This helps the poor debugger make sense of why an assertion is being made and enforced, and hopefully what to do about it. Here is one complete example:

```
inline Value *getOperand(unsigned I) {
    assert(I < Operands.size() && "getOperand() out of range!");
    return Operands[I];
}
```

Here are more examples:

```
assert(Ty->isPointerType() && "Can't allocate a non-pointer type!");

assert((Opcode == Shl || Opcode == Shr) && "ShiftInst Opcode invalid!");

assert(idx < getNumSuccessors() && "Successor # out of range!");

assert(V1.getType() == V2.getType() && "Constant types must be identical!");

assert(isa<PHINode>(Succ->front()) && "Only works on PHId BBs!");
```

You get the idea.

In the past, asserts were used to indicate a piece of code that should not be reached. These were typically of the form:

```
assert(0 && "Invalid radix for integer literal");
```

This has a few issues, the main one being that some compilers might not understand the assertion, or warn about a missing return in builds where assertions are compiled out.

Today, we have something much better: `llvm_unreachable`:

```
llvm_unreachable("Invalid radix for integer literal");
```

When assertions are enabled, this will print the message if it's ever reached and then exit the program. When assertions are disabled (i.e. in release builds), `llvm_unreachable` becomes a hint to compilers to skip generating code for this branch. If the compiler does not support this, it will fall back to the “abort” implementation.

Another issue is that values used only by assertions will produce an “unused value” warning when assertions are disabled. For example, this code will warn:

```
unsigned Size = V.size();  
assert(Size > 42 && "Vector smaller than it should be");  
  
bool NewToSet = Myset.insert(Value);  
assert(NewToSet && "The value shouldn't be in the set yet");
```

These are two interesting different cases. In the first case, the call to `V.size()` is only useful for the assert, and we don't want it executed when assertions are disabled. Code like this should move the call into the assert itself. In the second case, the side effects of the call must happen whether the assert is enabled or not. In this case, the value should be cast to void to disable the warning. To be specific, it is preferred to write the code like this:

```
assert(V.size() > 42 && "Vector smaller than it should be");  
  
bool NewToSet = Myset.insert(Value); (void)NewToSet;  
assert(NewToSet && "The value shouldn't be in the set yet");
```

Do Not Use using namespace std

In LLVM, we prefer to explicitly prefix all identifiers from the standard namespace with an “`std::`” prefix, rather than rely on “`using namespace std;`”.

In header files, adding a `'using namespace XXX'` directive pollutes the namespace of any source file that `#includes` the header. This is clearly a bad thing.

In implementation files (e.g. `.cpp` files), the rule is more of a stylistic rule, but is still important. Basically, using explicit namespace prefixes makes the code **clearer**, because it is immediately obvious what facilities are being used and where they are coming from. And **more portable**, because namespace clashes cannot occur between LLVM code and other namespaces. The portability rule is important because different standard library implementations expose different symbols (potentially ones they shouldn't), and future revisions to the C++ standard will add more symbols to the `std` namespace. As such, we never use `'using namespace std;'` in LLVM.

The exception to the general rule (i.e. it's not an exception for the `std` namespace) is for implementation files. For example, all of the code in the LLVM project implements code that lives in the `'llvm'` namespace. As such, it is ok, and actually clearer, for the `.cpp` files to have a `'using namespace llvm;'` directive at the top, after the `#includes`. This reduces indentation in the body of the file for source editors that indent based on braces, and keeps the conceptual context cleaner. The general form of this rule is that any `.cpp` file that implements code in any namespace may use that namespace (and its parents'), but should not use any others.

Provide a Virtual Method Anchor for Classes in Headers

If a class is defined in a header file and has a vtable (either it has virtual methods or it derives from classes with virtual methods), it must always have at least one out-of-line virtual method in the class. Without this, the compiler will copy the vtable and RTTI into every `.o` file that `#includes` the header, bloating `.o` file sizes and increasing link times.

Don't use default labels in fully covered switches over enumerations

`-Wswitch` warns if a switch, without a default label, over an enumeration does not cover every enumeration value. If you write a default label on a fully covered switch over an enumeration then the `-Wswitch` warning won't fire when new elements are added to that enumeration. To help avoid adding these kinds of defaults, Clang has the warning `-Wcovered-switch-default` which is off by default but turned on when building LLVM with a version of Clang that supports the warning.

A knock-on effect of this stylistic requirement is that when building LLVM with GCC you may get warnings related to “control may reach end of non-void function” if you return from each case of a covered switch-over-enum because GCC assumes that the enum expression may take any representable value, not just those of individual enumerators. To suppress this warning, use `llvm_unreachable` after the switch.

Don't evaluate `end()` every time through a loop

Because C++ doesn't have a standard “foreach” loop (though it can be emulated with macros and may be coming in C++0x) we end up writing a lot of loops that manually iterate from begin to end on a variety of containers or through other data structures. One common mistake is to write a loop in this style:

```
BasicBlock *BB = ...
for (BasicBlock::iterator I = BB->begin(); I != BB->end(); ++I)
    ... use I ...
```

The problem with this construct is that it evaluates “`BB->end()`” every time through the loop. Instead of writing the loop like this, we strongly prefer loops to be written so that they evaluate it once before the loop starts. A convenient way to do this is like so:

```
BasicBlock *BB = ...
for (BasicBlock::iterator I = BB->begin(), E = BB->end(); I != E; ++I)
    ... use I ...
```

The observant may quickly point out that these two loops may have different semantics: if the container (a basic block in this case) is being mutated, then “`BB->end()`” may change its value every time through the loop and the second loop may not in fact be correct. If you actually do depend on this behavior, please write the loop in the first form and add a comment indicating that you did it intentionally.

Why do we prefer the second form (when correct)? Writing the loop in the first form has two problems. First it may be less efficient than evaluating it at the start of the loop. In this case, the cost is probably minor — a few extra loads every time through the loop. However, if the base expression is more complex, then the cost can rise quickly. I've seen loops where the end expression was actually something like: “`SomeMap[X]->end()`” and map lookups really aren't cheap. By writing it in the second form consistently, you eliminate the issue entirely and don't even have to think about it.

The second (even bigger) issue is that writing the loop in the first form hints to the reader that the loop is mutating the container (a fact that a comment would handily confirm!). If you write the loop in the second form, it is immediately obvious without even looking at the body of the loop that the container isn't being modified, which makes it easier to read the code and understand what it does.

While the second form of the loop is a few extra keystrokes, we do strongly prefer it.

#include <iostream> is Forbidden

The use of `#include <iostream>` in library files is hereby **forbidden**, because many common implementations transparently inject a *static constructor* into every translation unit that includes it.

Note that using the other stream headers (`<sstream>` for example) is not problematic in this regard — just `<iostream>`. However, `raw_ostream` provides various APIs that are better performing for almost every use than `std::ostream` style APIs.

Note: New code should always use *raw_ostream* for writing, or the `llvm::MemoryBuffer` API for reading files.

Use raw_ostream

LLVM includes a lightweight, simple, and efficient stream implementation in `llvm/Support/raw_ostream.h`, which provides all of the common features of `std::ostream`. All new code should use `raw_ostream` instead of `ostream`.

Unlike `std::ostream`, `raw_ostream` is not a template and can be forward declared as `class raw_ostream`. Public headers should generally not include the `raw_ostream` header, but use forward declarations and constant references to `raw_ostream` instances.

Avoid std::endl

The `std::endl` modifier, when used with `iostreams` outputs a newline to the output stream specified. In addition to doing this, however, it also flushes the output stream. In other words, these are equivalent:

```
std::cout << std::endl;
std::cout << '\n' << std::flush;
```

Most of the time, you probably have no reason to flush the output stream, so it's better to use a literal `'\n'`.

Don't use inline when defining a function in a class definition

A member function defined in a class definition is implicitly inline, so don't put the `inline` keyword in this case.

Don't:

```
class Foo {
public:
    inline void bar() {
        // ...
    }
};
```

Do:

```
class Foo {
public:
    void bar() {
        // ...
    }
};
```

Microscopic Details

This section describes preferred low-level formatting guidelines along with reasoning on why we prefer them.

Spaces Before Parentheses

We prefer to put a space before an open parenthesis only in control flow statements, but not in normal function call expressions and function-like macros. For example, this is good:

```
if (X) ...
for (I = 0; I != 100; ++I) ...
while (LLVMRocks) ...

somefunc(42);
assert(3 != 4 && "laws of math are failing me");

A = foo(42, 92) + bar(X);
```

and this is bad:

```
if(X) ...
for(I = 0; I != 100; ++I) ...
while(LLVMRocks) ...

somefunc (42);
assert (3 != 4 && "laws of math are failing me");

A = foo (42, 92) + bar (X);
```

The reason for doing this is not completely arbitrary. This style makes control flow operators stand out more, and makes expressions flow better. The function call operator binds very tightly as a postfix operator. Putting a space after a function name (as in the last example) makes it appear that the code might bind the arguments of the left-hand-side of a binary operator with the argument list of a function and the name of the right side. More specifically, it is easy to misread the “A” example as:

```
A = foo ((42, 92) + bar) (X);
```

when skimming through the code. By avoiding a space in a function, we avoid this misinterpretation.

Prefer Preincrement

Hard fast rule: Preincrement (++X) may be no slower than postincrement (X++) and could very well be a lot faster than it. Use preincrementation whenever possible.

The semantics of postincrement include making a copy of the value being incremented, returning it, and then preincrementing the “work value”. For primitive types, this isn’t a big deal. But for iterators, it can be a huge issue (for example, some iterators contains stack and set objects in them... copying an iterator could invoke the copy ctor’s of these as well). In general, get in the habit of always using preincrement, and you won’t have a problem.

Namespace Indentation

In general, we strive to reduce indentation wherever possible. This is useful because we want code to *fit into 80 columns* without wrapping horribly, but also because it makes it easier to understand the code. To facilitate this and

avoid some insanely deep nesting on occasion, don't indent namespaces. If it helps readability, feel free to add a comment indicating what namespace is being closed by a `}`. For example:

```
namespace llvm {
namespace knowledge {

/// This class represents things that Smith can have an intimate
/// understanding of and contains the data associated with it.
class Grokable {
...
public:
    explicit Grokable() { ... }
    virtual ~Grokable() = 0;

    ...
};

} // end namespace knowledge
} // end namespace llvm
```

Feel free to skip the closing comment when the namespace being closed is obvious for any reason. For example, the outer-most namespace in a header file is rarely a source of confusion. But namespaces both anonymous and named in source files that are being closed half way through the file probably could use clarification.

Anonymous Namespaces

After talking about namespaces in general, you may be wondering about anonymous namespaces in particular. Anonymous namespaces are a great language feature that tells the C++ compiler that the contents of the namespace are only visible within the current translation unit, allowing more aggressive optimization and eliminating the possibility of symbol name collisions. Anonymous namespaces are to C++ as “static” is to C functions and global variables. While “static” is available in C++, anonymous namespaces are more general: they can make entire classes private to a file.

The problem with anonymous namespaces is that they naturally want to encourage indentation of their body, and they reduce locality of reference: if you see a random function definition in a C++ file, it is easy to see if it is marked static, but seeing if it is in an anonymous namespace requires scanning a big chunk of the file.

Because of this, we have a simple guideline: make anonymous namespaces as small as possible, and only use them for class declarations. For example, this is good:

```
namespace {
class StringSort {
...
public:
    StringSort(...)
    bool operator<(const char *RHS) const;
};
} // end anonymous namespace

static void runHelper() {
    ...
}

bool StringSort::operator<(const char *RHS) const {
    ...
}
```

This is bad:

```
namespace {
class StringSort {
...
public:
    StringSort(...)
    bool operator<(const char *RHS) const;
};

void runHelper() {
    ...
}

bool StringSort::operator<(const char *RHS) const {
    ...
}

} // end anonymous namespace
```

This is bad specifically because if you’re looking at “runHelper” in the middle of a large C++ file, that you have no immediate way to tell if it is local to the file. When it is marked static explicitly, this is immediately obvious. Also, there is no reason to enclose the definition of “operator<” in the namespace just because it was declared there.

3.2.5 See Also

A lot of these comments and recommendations have been culled from other sources. Two particularly important books for our work are:

1. [Effective C++](#) by Scott Meyers. Also interesting and useful are “More Effective C++” and “Effective STL” by the same author.
2. [Large-Scale C++ Software Design](#) by John Lakos

If you get some free time, and you haven’t read them: do so, you might learn something.

3.3 CommandLine 2.0 Library Manual

- *Introduction*
- *Quick Start Guide*
 - *Boolean Arguments*
 - *Argument Aliases*
 - *Selecting an alternative from a set of possibilities*
 - *Named Alternatives*
 - *Parsing a list of options*
 - *Collecting options as a set of flags*
 - *Adding freeform text to help output*

- *Grouping options into categories*
- *Reference Guide*
 - *Positional Arguments*
 - * *Specifying positional options with hyphens*
 - * *Determining absolute position with getPosition()*
 - * *The `cl::ConsumeAfter` modifier*
 - *Internal vs External Storage*
 - *Option Attributes*
 - *Option Modifiers*
 - * *Hiding an option from `-help` output*
 - * *Controlling the number of occurrences required and allowed*
 - * *Controlling whether or not a value must be specified*
 - * *Controlling other formatting options*
 - * *Miscellaneous option modifiers*
 - * *Response files*
 - *Top-Level Classes and Functions*
 - * *The `cl::getRegisteredOptions` function*
 - * *The `cl::ParseCommandLineOptions` function*
 - * *The `cl::ParseEnvironmentOptions` function*
 - * *The `cl::SetVersionPrinter` function*
 - * *The `cl::opt` class*
 - * *The `cl::list` class*
 - * *The `cl::bits` class*
 - * *The `cl::alias` class*
 - * *The `cl::extrahelp` class*
 - * *The `cl::OptionCategory` class*
 - *Builtin parsers*
- *Extension Guide*
 - *Writing a custom parser*
 - *Exploiting external storage*
 - *Dynamically adding command line options*

3.3.1 Introduction

This document describes the CommandLine argument processing library. It will show you how to use it, and what it can do. The CommandLine library uses a declarative approach to specifying the command line options that your program takes. By default, these options declarations implicitly hold the value parsed for the option declared (of

course this *can be changed*).

Although there are a **lot** of command line argument parsing libraries out there in many different languages, none of them fit well with what I needed. By looking at the features and problems of other libraries, I designed the CommandLine library to have the following features:

1. **Speed:** The CommandLine library is very quick and uses little resources. The parsing time of the library is directly proportional to the number of arguments parsed, not the number of options recognized. Additionally, command line argument values are captured transparently into user defined global variables, which can be accessed like any other variable (and with the same performance).
2. **Type Safe:** As a user of CommandLine, you don't have to worry about remembering the type of arguments that you want (is it an int? a string? a bool? an enum?) and keep casting it around. Not only does this help prevent error prone constructs, it also leads to dramatically cleaner source code.
3. **No subclasses required:** To use CommandLine, you instantiate variables that correspond to the arguments that you would like to capture, you don't subclass a parser. This means that you don't have to write **any** boilerplate code.
4. **Globally accessible:** Libraries can specify command line arguments that are automatically enabled in any tool that links to the library. This is possible because the application doesn't have to keep a list of arguments to pass to the parser. This also makes supporting *dynamically loaded options* trivial.
5. **Cleaner:** CommandLine supports enum and other types directly, meaning that there is less error and more security built into the library. You don't have to worry about whether your integral command line argument accidentally got assigned a value that is not valid for your enum type.
6. **Powerful:** The CommandLine library supports many different types of arguments, from simple *boolean flags* to *scalars arguments* (*strings*, *integers*, *enums*, *doubles*), to *lists of arguments*. This is possible because CommandLine is...
7. **Extensible:** It is very simple to add a new argument type to CommandLine. Simply specify the parser that you want to use with the command line option when you declare it. *Custom parsers* are no problem.
8. **Labor Saving:** The CommandLine library cuts down on the amount of grunt work that you, the user, have to do. For example, it automatically provides a `-help` option that shows the available command line options for your tool. Additionally, it does most of the basic correctness checking for you.
9. **Capable:** The CommandLine library can handle lots of different forms of options often found in real programs. For example, *positional* arguments, `ls` style *grouping* options (to allow processing `'ls -lad'` naturally), `ld` style *prefix* options (to parse `'-lmalloc -L/usr/lib'`), and interpreter style options.

This document will hopefully let you jump in and start using CommandLine in your utility quickly and painlessly. Additionally it should be a simple reference manual to figure out how stuff works.

3.3.2 Quick Start Guide

This section of the manual runs through a simple CommandLine'ification of a basic compiler tool. This is intended to show you how to jump into using the CommandLine library in your own program, and show you some of the cool things it can do.

To start out, you need to include the CommandLine header file into your program:

```
#include "llvm/Support/CommandLine.h"
```

Additionally, you need to add this as the first line of your main program:

```
int main(int argc, char **argv) {
    cl::ParseCommandLineOptions(argc, argv);
```

```
...
}
```

... which actually parses the arguments and fills in the variable declarations.

Now that you are ready to support command line arguments, we need to tell the system which ones we want, and what type of arguments they are. The `CommandLine` library uses a declarative syntax to model command line arguments with the global variable declarations that capture the parsed values. This means that for every command line option that you would like to support, there should be a global variable declaration to capture the result. For example, in a compiler, we would like to support the Unix-standard ‘`-o <filename>`’ option to specify where to put the output. With the `CommandLine` library, this is represented like this:

```
cl::opt<string> OutputFilename("o", cl::desc("Specify output filename"), cl::value_
↳ desc("filename"));
```

This declares a global variable “`OutputFilename`” that is used to capture the result of the “`o`” argument (first parameter). We specify that this is a simple scalar option by using the “`cl::opt`” template (as opposed to the “`cl::list`” template), and tell the `CommandLine` library that the data type that we are parsing is a string.

The second and third parameters (which are optional) are used to specify what to output for the “`-help`” option. In this case, we get a line that looks like this:

```
USAGE: compiler [options]

OPTIONS:
  -help                - display available options (-help-hidden for more)
  -o <filename>        - Specify output filename
```

Because we specified that the command line option should parse using the `string` data type, the variable declared is automatically usable as a real string in all contexts that a normal C++ string object may be used. For example:

```
...
std::ofstream Output(OutputFilename.c_str());
if (Output.good()) ...
...
```

There are many different options that you can use to customize the command line option handling library, but the above example shows the general interface to these options. The options can be specified in any order, and are specified with helper functions like `cl::desc(...)`, so there are no positional dependencies to remember. The available options are discussed in detail in the [Reference Guide](#).

Continuing the example, we would like to have our compiler take an input filename as well as an output filename, but we do not want the input filename to be specified with a hyphen (ie, not `-filename.c`). To support this style of argument, the `CommandLine` library allows for *positional* arguments to be specified for the program. These positional arguments are filled with command line parameters that are not in option form. We use this feature like this:

```
cl::opt<string> InputFilename(cl::Positional, cl::desc("<input file>"), cl::init("-
↳ "));
```

This declaration indicates that the first positional argument should be treated as the input filename. Here we use the `cl::init` option to specify an initial value for the command line option, which is used if the option is not specified (if you do not specify a `cl::init` modifier for an option, then the default constructor for the data type is used to initialize the value). Command line options default to being optional, so if we would like to require that the user always specify an input filename, we would add the `cl::Required` flag, and we could eliminate the `cl::init` modifier, like this:

```
cl::opt<string> InputFilename(cl::Positional, cl::desc("<input file>"), cl::Required);
```


Again, the CommandLine library does not require the options to be specified in any particular order, so the above declaration is equivalent to:

```
cl::opt<string> InputFilename(cl::Positional, cl::Required, cl::desc("<input file>"));
```

By simply adding the `cl::Required` flag, the CommandLine library will automatically issue an error if the argument is not specified, which shifts all of the command line option verification code out of your application into the library. This is just one example of how using flags can alter the default behaviour of the library, on a per-option basis. By adding one of the declarations above, the `-help` option synopsis is now extended to:

```
USAGE: compiler [options] <input file>

OPTIONS:
  -help          - display available options (-help-hidden for more)
  -o <filename>  - Specify output filename
```

... indicating that an input filename is expected.

Boolean Arguments

In addition to input and output filenames, we would like the compiler example to support three boolean flags: “-f” to force writing binary output to a terminal, “-quiet” to enable quiet mode, and “-q” for backwards compatibility with some of our users. We can support these by declaring options of boolean type like this:

```
cl::opt<bool> Force ("f", cl::desc("Enable binary output on terminals"));
cl::opt<bool> Quiet ("quiet", cl::desc("Don't print informational messages"));
cl::opt<bool> Quiet2("q", cl::desc("Don't print informational messages"), cl::Hidden);
```

This does what you would expect: it declares three boolean variables (“Force”, “Quiet”, and “Quiet2”) to recognize these options. Note that the “-q” option is specified with the “`cl::Hidden`” flag. This modifier prevents it from being shown by the standard “-help” output (note that it is still shown in the “-help-hidden” output).

The CommandLine library uses a *different parser* for different data types. For example, in the string case, the argument passed to the option is copied literally into the content of the string variable... we obviously cannot do that in the boolean case, however, so we must use a smarter parser. In the case of the boolean parser, it allows no options (in which case it assigns the value of true to the variable), or it allows the values “true” or “false” to be specified, allowing any of the following inputs:

```
compiler -f          # No value, 'Force' == true
compiler -f=true     # Value specified, 'Force' == true
compiler -f=TRUE     # Value specified, 'Force' == true
compiler -f=FALSE    # Value specified, 'Force' == false
```

... you get the idea. The *bool parser* just turns the string values into boolean values, and rejects things like ‘compiler -f=foo’. Similarly, the *float*, *double*, and *int* parsers work like you would expect, using the ‘strtol’ and ‘strtod’ C library calls to parse the string value into the specified data type.

With the declarations above, “compiler -help” emits this:

```
USAGE: compiler [options] <input file>

OPTIONS:
  -f          - Enable binary output on terminals
  -o          - Override output filename
  -quiet      - Don't print informational messages
  -help       - display available options (-help-hidden for more)
```

and “`compiler -help-hidden`” prints this:

```
USAGE: compiler [options] <input file>

OPTIONS:
  -f      - Enable binary output on terminals
  -o      - Override output filename
  -q      - Don't print informational messages
  -quiet  - Don't print informational messages
  -help   - display available options (-help-hidden for more)
```

This brief example has shown you how to use the ‘`cl::opt`’ class to parse simple scalar command line arguments. In addition to simple scalar arguments, the `CommandLine` library also provides primitives to support `CommandLine` option *aliases*, and *lists* of options.

Argument Aliases

So far, the example works well, except for the fact that we need to check the quiet condition like this now:

```
...
  if (!Quiet && !Quiet2) printInformationalMessage(...);
...
```

... which is a real pain! Instead of defining two values for the same condition, we can use the “`cl::alias`” class to make the “`-q`” option an **alias** for the “`-quiet`” option, instead of providing a value itself:

```
cl::opt<bool> Force ("f", cl::desc("Overwrite output files"));
cl::opt<bool> Quiet ("quiet", cl::desc("Don't print informational messages"));
cl::alias      QuietA("q", cl::desc("Alias for -quiet"), cl::aliasopt(Quiet));
```

The third line (which is the only one we modified from above) defines a “`-q`” alias that updates the “`Quiet`” variable (as specified by the `cl::aliasopt` modifier) whenever it is specified. Because aliases do not hold state, the only thing the program has to query is the `Quiet` variable now. Another nice feature of aliases is that they automatically hide themselves from the `-help` output (although, again, they are still visible in the `-help-hidden` output).

Now the application code can simply use:

```
...
  if (!Quiet) printInformationalMessage(...);
...
```

... which is much nicer! The “`cl::alias`” can be used to specify an alternative name for any variable type, and has many uses.

Selecting an alternative from a set of possibilities

So far we have seen how the `CommandLine` library handles builtin types like `std::string`, `bool` and `int`, but how does it handle things it doesn’t know about, like enums or ‘`int*`’s?

The answer is that it uses a table-driven generic parser (unless you specify your own parser, as described in the *Extension Guide*). This parser maps literal strings to whatever type is required, and requires you to tell it what this mapping should be.

Let’s say that we would like to add four optimization levels to our optimizer, using the standard flags “`-g`”, “`-O0`”, “`-O1`”, and “`-O2`”. We could easily implement this with boolean options like above, but there are several problems with this strategy:

1. A user could specify more than one of the options at a time, for example, “`compiler -O3 -O2`”. The `CommandLine` library would not be able to catch this erroneous input for us.
2. We would have to test 4 different variables to see which ones are set.
3. This doesn’t map to the numeric levels that we want... so we cannot easily see if some level \geq “-O1” is enabled.

To cope with these problems, we can use an enum value, and have the `CommandLine` library fill it in with the appropriate level directly, which is used like this:

```
enum OptLevel {
    g, O1, O2, O3
};

cl::opt<OptLevel> OptimizationLevel(cl::desc("Choose optimization level:"),
    cl::values(
        clEnumVal(g, "No optimizations, enable debugging"),
        clEnumVal(O1, "Enable trivial optimizations"),
        clEnumVal(O2, "Enable default optimizations"),
        clEnumVal(O3, "Enable expensive optimizations")));

...
if (OptimizationLevel >= O2) doPartialRedundancyElimination(...);
...
```

This declaration defines a variable “`OptimizationLevel`” of the “`OptLevel`” enum type. This variable can be assigned any of the values that are listed in the declaration. The `CommandLine` library enforces that the user can only specify one of the options, and it ensure that only valid enum values can be specified. The “`clEnumVal`” macros ensure that the command line arguments matched the enum values. With this option added, our help output now is:

```
USAGE: compiler [options] <input file>

OPTIONS:
  Choose optimization level:
    -g          - No optimizations, enable debugging
    -O1         - Enable trivial optimizations
    -O2         - Enable default optimizations
    -O3         - Enable expensive optimizations
    -f          - Enable binary output on terminals
    -help       - display available options (-help-hidden for more)
    -o <filename> - Specify output filename
    -quiet      - Don't print informational messages
```

In this case, it is sort of awkward that flag names correspond directly to enum names, because we probably don’t want a enum definition named “`g`” in our program. Because of this, we can alternatively write this example like this:

```
enum OptLevel {
    Debug, O1, O2, O3
};

cl::opt<OptLevel> OptimizationLevel(cl::desc("Choose optimization level:"),
    cl::values(
        clEnumValN(Debug, "g", "No optimizations, enable debugging"),
        clEnumVal(O1, "Enable trivial optimizations"),
        clEnumVal(O2, "Enable default optimizations"),
        clEnumVal(O3, "Enable expensive optimizations")));

...
```

```
if (OptimizationLevel == Debug) outputDebugInfo(...);
...
```

By using the “`clEnumValN`” macro instead of “`clEnumVal`”, we can directly specify the name that the flag should get. In general a direct mapping is nice, but sometimes you can’t or don’t want to preserve the mapping, which is when you would use it.

Named Alternatives

Another useful argument form is a named alternative style. We shall use this style in our compiler to specify different debug levels that can be used. Instead of each debug level being its own switch, we want to support the following options, of which only one can be specified at a time: “`--debug-level=none`”, “`--debug-level=quick`”, “`--debug-level=detailed`”. To do this, we use the exact same format as our optimization level flags, but we also specify an option name. For this case, the code looks like this:

```
enum DebugLev {
    nodebuginfo, quick, detailed
};

// Enable Debug Options to be specified on the command line
cl::opt<DebugLev> DebugLevel("debug_level", cl::desc("Set the debugging level:"),
    cl::values(
        clEnumValN(nodebuginfo, "none", "disable debug information"),
        clEnumVal(quick, "enable quick debug information"),
        clEnumVal(detailed, "enable detailed debug information")));
```

This definition defines an enumerated command line variable of type “`enum DebugLev`”, which works exactly the same way as before. The difference here is just the interface exposed to the user of your program and the help output by the “`-help`” option:

```
USAGE: compiler [options] <input file>

OPTIONS:
  Choose optimization level:
    -g             - No optimizations, enable debugging
    -O1            - Enable trivial optimizations
    -O2            - Enable default optimizations
    -O3            - Enable expensive optimizations
  -debug_level    - Set the debugging level:
    =none         - disable debug information
    =quick        - enable quick debug information
    =detailed     - enable detailed debug information
  -f             - Enable binary output on terminals
  -help          - display available options (-help-hidden for more)
  -o <filename> - Specify output filename
  -quiet         - Don't print informational messages
```

Again, the only structural difference between the debug level declaration and the optimization level declaration is that the debug level declaration includes an option name (“`debug_level`”), which automatically changes how the library processes the argument. The `CommandLine` library supports both forms so that you can choose the form most appropriate for your application.

Parsing a list of options

Now that we have the standard run-of-the-mill argument types out of the way, let's get a little wild and crazy. Let's say that we want our optimizer to accept a **list** of optimizations to perform, allowing duplicates. For example, we might want to run: “`compiler -dce -constprop -inline -dce -strip`”. In this case, the order of the arguments and the number of appearances is very important. This is what the “`cl::list`” template is for. First, start by defining an enum of the optimizations that you would like to perform:

```
enum Opts {
    // 'inline' is a C++ keyword, so name it 'inlining'
    dce, constprop, inlining, strip
};
```

Then define your “`cl::list`” variable:

```
cl::list<Opts> OptimizationList(cl::desc("Available Optimizations:"),
    cl::values(
        clEnumVal(dce, "Dead Code Elimination"),
        clEnumVal(constprop, "Constant Propagation"),
        clEnumValN(inlining, "inline", "Procedure Integration"),
        clEnumVal(strip, "Strip Symbols")));
```

This defines a variable that is conceptually of the type “`std::vector<enum Opts>`”. Thus, you can access it with standard vector methods:

```
for (unsigned i = 0; i != OptimizationList.size(); ++i)
    switch (OptimizationList[i])
        ...
```

... to iterate through the list of options specified.

Note that the “`cl::list`” template is completely general and may be used with any data types or other arguments that you can use with the “`cl::opt`” template. One especially useful way to use a list is to capture all of the positional arguments together if there may be more than one specified. In the case of a linker, for example, the linker takes several ‘.o’ files, and needs to capture them into a list. This is naturally specified as:

```
...
cl::list<std::string> InputFileNames(cl::Positional, cl::desc("<Input files>"),
    ↪cl::OneOrMore);
...
```

This variable works just like a “`vector<string>`” object. As such, accessing the list is simple, just like above. In this example, we used the `cl::OneOrMore` modifier to inform the CommandLine library that it is an error if the user does not specify any .o files on our command line. Again, this just reduces the amount of checking we have to do.

Collecting options as a set of flags

Instead of collecting sets of options in a list, it is also possible to gather information for enum values in a **bit vector**. The representation used by the `cl::bits` class is an unsigned integer. An enum value is represented by a 0/1 in the enum’s ordinal value bit position. 1 indicating that the enum was specified, 0 otherwise. As each specified value is parsed, the resulting enum’s bit is set in the option’s bit vector:

```
bits |= 1 << (unsigned)enum;
```

Options that are specified multiple times are redundant. Any instances after the first are discarded.

Reworking the above list example, we could replace `cl::list` with `cl::bits`:

```
cl::bits<Opts> OptimizationBits(cl::desc("Available Optimizations:"),
    cl::values(
        clEnumVal(dce, "Dead Code Elimination"),
        clEnumVal(constprop, "Constant Propagation"),
        clEnumValN(inlining, "inline", "Procedure Integration"),
        clEnumVal(strip, "Strip Symbols")));
```

To test to see if `constprop` was specified, we can use the `cl::bits::isSet` function:

```
if (OptimizationBits.isSet(constprop)) {
    ...
}
```

It's also possible to get the raw bit vector using the `cl::bits::getBits` function:

```
unsigned bits = OptimizationBits.getBits();
```

Finally, if external storage is used, then the location specified must be of **type** `unsigned`. In all other ways a `cl::bits` option is equivalent to a `cl::list` option.

Adding freeform text to help output

As our program grows and becomes more mature, we may decide to put summary information about what it does into the help output. The help output is styled to look similar to a Unix man page, providing concise information about a program. Unix man pages, however often have a description about what the program does. To add this to your CommandLine program, simply pass a third argument to the `cl::ParseCommandLineOptions` call in main. This additional argument is then printed as the overview information for your program, allowing you to include any additional information that you want. For example:

```
int main(int argc, char **argv) {
    cl::ParseCommandLineOptions(argc, argv, " CommandLine compiler example\n\n"
                                " This program blah blah blah...\n");
    ...
}
```

would yield the help output:

```
**OVERVIEW: CommandLine compiler example

This program blah blah blah...**

USAGE: compiler [options] <input file>

OPTIONS:
    ...
    -help                - display available options (-help-hidden for more)
    -o <filename>        - Specify output filename
```

Grouping options into categories

If our program has a large number of options it may become difficult for users of our tool to navigate the output of `-help`. To alleviate this problem we can put our options into categories. This can be done by declaring option categories (`cl::OptionCategory` objects) and then placing our options into these categories using the `cl::cat` option attribute. For example:

```
cl::OptionCategory StageSelectionCat("Stage Selection Options",
                                     "These control which stages are run.");

cl::opt<bool> Preprocessor("E", cl::desc("Run preprocessor stage."),
                          cl::cat(StageSelectionCat));

cl::opt<bool> NoLink("c", cl::desc("Run all stages except linking."),
                   cl::cat(StageSelectionCat));
```

The output of `-help` will become categorized if an option category is declared. The output looks something like

```
OVERVIEW: This is a small program to demo the LLVM CommandLine API
USAGE: Sample [options]

OPTIONS:

  General options:

    -help                - Display available options (-help-hidden for more)
    -help-list           - Display list of available options (-help-list-hidden for
    ↪more)

  Stage Selection Options:
  These control which stages are run.

    -E                  - Run preprocessor stage.
    -c                  - Run all stages except linking.
```

In addition to the behaviour of `-help` changing when an option category is declared, the command line option `-help-list` becomes visible which will print the command line options as uncategorized list.

Note that Options that are not explicitly categorized will be placed in the `cl::GeneralCategory` category.

3.3.3 Reference Guide

Now that you know the basics of how to use the CommandLine library, this section will give you the detailed information you need to tune how command line options work, as well as information on more “advanced” command line option processing capabilities.

Positional Arguments

Positional arguments are those arguments that are not named, and are not specified with a hyphen. Positional arguments should be used when an option is specified by its position alone. For example, the standard Unix `grep` tool takes a regular expression argument, and an optional filename to search through (which defaults to standard input if a filename is not specified). Using the CommandLine library, this would be specified as:

```
cl::opt<string> Regex (cl::Positional, cl::desc("<regular expression>"),
    ↪cl::Required);
cl::opt<string> Filename(cl::Positional, cl::desc("<input file>"), cl::init("-"));
```

Given these two option declarations, the `-help` output for our `grep` replacement would look like this:

```
USAGE: spiffygrep [options] <regular expression> <input file>
```

OPTIONS:

```
-help - display available options (-help-hidden for more)
```

... and the resultant program could be used just like the standard `grep` tool.

Positional arguments are sorted by their order of construction. This means that command line options will be ordered according to how they are listed in a `.cpp` file, but will not have an ordering defined if the positional arguments are defined in multiple `.cpp` files. The fix for this problem is simply to define all of your positional arguments in one `.cpp` file.

Specifying positional options with hyphens

Sometimes you may want to specify a value to your positional argument that starts with a hyphen (for example, searching for `-foo` in a file). At first, you will have trouble doing this, because it will try to find an argument named `-foo`, and will fail (and single quotes will not save you). Note that the system `grep` has the same problem:

```
$ spiffygrep '-foo' test.txt
Unknown command line argument '-foo'. Try: spiffygrep -help'

$ grep '-foo' test.txt
grep: illegal option -- f
grep: illegal option -- o
grep: illegal option -- o
Usage: grep -hblcnsviw pattern file . . .
```

The solution for this problem is the same for both your tool and the system version: use the `--` marker. When the user specifies `--` on the command line, it is telling the program that all options after the `--` should be treated as positional arguments, not options. Thus, we can use it like this:

```
$ spiffygrep -- -foo test.txt
...output...
```

Determining absolute position with `getPosition()`

Sometimes an option can affect or modify the meaning of another option. For example, consider `gcc`'s `-x LANG` option. This tells `gcc` to ignore the suffix of subsequent positional arguments and force the file to be interpreted as if it contained source code in language `LANG`. In order to handle this properly, you need to know the absolute position of each argument, especially those in lists, so their interaction(s) can be applied correctly. This is also useful for options like `-llibname` which is actually a positional argument that starts with a dash.

So, generally, the problem is that you have two `cl::list` variables that interact in some way. To ensure the correct interaction, you can use the `cl::list::getPosition(optnum)` method. This method returns the absolute position (as found on the command line) of the `optnum` item in the `cl::list`.

The idiom for usage is like this:

```
static cl::list<std::string> Files(cl::Positional, cl::OneOrMore);
static cl::list<std::string> Libraries("l", cl::ZeroOrMore);

int main(int argc, char**argv) {
    // ...
    std::vector<std::string>::iterator fileIt = Files.begin();
    std::vector<std::string>::iterator libIt  = Libraries.begin();
    unsigned libPos = 0, filePos = 0;
```



```

while ( 1 ) {
    if ( libIt != Libraries.end() )
        libPos = Libraries.getPosition( libIt - Libraries.begin() );
    else
        libPos = 0;
    if ( fileIt != Files.end() )
        filePos = Files.getPosition( fileIt - Files.begin() );
    else
        filePos = 0;

    if ( filePos != 0 && (libPos == 0 || filePos < libPos) ) {
        // Source File Is next
        ++fileIt;
    }
    else if ( libPos != 0 && (filePos == 0 || libPos < filePos) ) {
        // Library is next
        ++libIt;
    }
    else
        break; // we're done with the list
}
}

```

Note that, for compatibility reasons, the `cl::opt` also supports an unsigned `getPosition()` option that will provide the absolute position of that option. You can apply the same approach as above with a `cl::opt` and a `cl::list` option as you can with two lists.

The `cl::ConsumeAfter` modifier

The `cl::ConsumeAfter` *formatting option* is used to construct programs that use “interpreter style” option processing. With this style of option processing, all arguments specified after the last positional argument are treated as special interpreter arguments that are not interpreted by the command line argument.

As a concrete example, lets say we are developing a replacement for the standard Unix Bourne shell (`/bin/sh`). To run `/bin/sh`, first you specify options to the shell itself (like `-x` which turns on trace output), then you specify the name of the script to run, then you specify arguments to the script. These arguments to the script are parsed by the Bourne shell command line option processor, but are not interpreted as options to the shell itself. Using the `CommandLine` library, we would specify this as:

```

cl::opt<string> Script(cl::Positional, cl::desc("<input script>"), cl::init("-"));
cl::list<string> Argv(cl::ConsumeAfter, cl::desc("<program arguments>..."));
cl::opt<bool> Trace("x", cl::desc("Enable trace output"));

```

which automatically provides the help output:

```

USAGE: spiffysh [options] <input script> <program arguments>...

OPTIONS:
  -help - display available options (-help-hidden for more)
  -x    - Enable trace output

```

At runtime, if we run our new shell replacement as ``spiffysh -x test.sh -a -x -y bar``, the `Trace` variable will be set to true, the `Script` variable will be set to `"test.sh"`, and the `Argv` list will contain `["-a", "-x", "-y", "bar"]`, because they were specified after the last positional argument (which is the script name).

There are several limitations to when `cl::ConsumeAfter` options can be specified. For example, only one

`cl::ConsumeAfter` can be specified per program, there must be at least one *positional argument* specified, there must not be any *cl::list* positional arguments, and the `cl::ConsumeAfter` option should be a *cl::list* option.

Internal vs External Storage

By default, all command line options automatically hold the value that they parse from the command line. This is very convenient in the common case, especially when combined with the ability to define command line options in the files that use them. This is called the internal storage model.

Sometimes, however, it is nice to separate the command line option processing code from the storage of the value parsed. For example, lets say that we have a ‘-debug’ option that we would like to use to enable debug information across the entire body of our program. In this case, the boolean value controlling the debug code should be globally accessible (in a header file, for example) yet the command line option processing code should not be exposed to all of these clients (requiring lots of .cpp files to `#include CommandLine.h`).

To do this, set up your .h file with your option, like this for example:

```
// DebugFlag.h - Get access to the '-debug' command line option
//

// DebugFlag - This boolean is set to true if the '-debug' command line option
// is specified. This should probably not be referenced directly, instead, use
// the DEBUG macro below.
//
extern bool DebugFlag;

// DEBUG macro - This macro should be used by code to emit debug information.
// In the '-debug' option is specified on the command line, and if this is a
// debug build, then the code specified as the option to the macro will be
// executed. Otherwise it will not be.
#ifdef NDEBUG
#define DEBUG(X)
#else
#define DEBUG(X) do { if (DebugFlag) { X; } } while (0)
#endif
```

This allows clients to blissfully use the `DEBUG()` macro, or the `DebugFlag` explicitly if they want to. Now we just need to be able to set the `DebugFlag` boolean when the option is set. To do this, we pass an additional argument to our command line argument processor, and we specify where to fill in with the *cl::location* attribute:

```
bool DebugFlag; // the actual value
static cl::opt<bool, true> // The parser
Debug("debug", cl::desc("Enable debug output"), cl::Hidden, cl::location(DebugFlag));
```

In the above example, we specify “true” as the second argument to the *cl::opt* template, indicating that the template should not maintain a copy of the value itself. In addition to this, we specify the *cl::location* attribute, so that `DebugFlag` is automatically set.

Option Attributes

This section describes the basic attributes that you can specify on options.

- The option name attribute (which is required for all options, except *positional options*) specifies what the option name is. This option is specified in simple double quotes:

```
cl::opt<bool> Quiet("quiet");
```

- The **cl::desc** attribute specifies a description for the option to be shown in the `-help` output for the program. This attribute supports multi-line descriptions with lines separated by ‘`\n`’.
- The **cl::value_desc** attribute specifies a string that can be used to fine tune the `-help` output for a command line option. Look [here](#) for an example.
- The **cl::init** attribute specifies an initial value for a [scalar](#) option. If this attribute is not specified then the command line option value defaults to the value created by the default constructor for the type.

Warning: If you specify both **cl::init** and **cl::location** for an option, you must specify **cl::location** first, so that when the command-line parser sees **cl::init**, it knows where to put the initial value. (You will get an error at runtime if you don’t put them in the right order.)

- The **cl::location** attribute where to store the value for a parsed command line option if using external storage. See the section on [Internal vs External Storage](#) for more information.
- The **cl::aliasopt** attribute specifies which option a [cl::alias](#) option is an alias for.
- The **cl::values** attribute specifies the string-to-value mapping to be used by the generic parser. It takes a list of (option, value, description) triplets that specify the option name, the value mapped to, and the description shown in the `-help` for the tool. Because the generic parser is used most frequently with enum values, two macros are often useful:
 1. The **clEnumVal** macro is used as a nice simple way to specify a triplet for an enum. This macro automatically makes the option name be the same as the enum name. The first option to the macro is the enum, the second is the description for the command line option.
 2. The **clEnumValN** macro is used to specify macro options where the option name doesn’t equal the enum name. For this macro, the first argument is the enum value, the second is the flag name, and the second is the description.

You will get a compile time error if you try to use `cl::values` with a parser that does not support it.

- The **cl::multi_val** attribute specifies that this option takes has multiple values (example: `-sectalign segname sectname sectvalue`). This attribute takes one unsigned argument - the number of values for the option. This attribute is valid only on `cl::list` options (and will fail with compile error if you try to use it with other option types). It is allowed to use all of the usual modifiers on multi-valued options (besides `cl::ValueDisallowed`, obviously).
- The **cl::cat** attribute specifies the option category that the option belongs to. The category should be a [cl::OptionCategory](#) object.

Option Modifiers

Option modifiers are the flags and expressions that you pass into the constructors for [cl::opt](#) and [cl::list](#). These modifiers give you the ability to tweak how options are parsed and how `-help` output is generated to fit your application well.

These options fall into five main categories:

1. Hiding an option from `-help` output
2. Controlling the number of occurrences required and allowed
3. Controlling whether or not a value must be specified

4. Controlling other formatting options
5. Miscellaneous option modifiers

It is not possible to specify two options from the same category (you'll get a runtime error) to a single option, except for options in the miscellaneous category. The `CommandLine` library specifies defaults for all of these settings that are the most useful in practice and the most common, which mean that you usually shouldn't have to worry about these.

Hiding an option from `-help` output

The `cl::NotHidden`, `cl::Hidden`, and `cl::ReallyHidden` modifiers are used to control whether or not an option appears in the `-help` and `-help-hidden` output for the compiled program:

- The **`cl::NotHidden`** modifier (which is the default for `cl::opt` and `cl::list` options) indicates the option is to appear in both help listings.
- The **`cl::Hidden`** modifier (which is the default for `cl::alias` options) indicates that the option should not appear in the `-help` output, but should appear in the `-help-hidden` output.
- The **`cl::ReallyHidden`** modifier indicates that the option should not appear in any help output.

Controlling the number of occurrences required and allowed

This group of options is used to control how many time an option is allowed (or required) to be specified on the command line of your program. Specifying a value for this setting allows the `CommandLine` library to do error checking for you.

The allowed values for this option group are:

- The **`cl::Optional`** modifier (which is the default for the `cl::opt` and `cl::alias` classes) indicates that your program will allow either zero or one occurrence of the option to be specified.
- The **`cl::ZeroOrMore`** modifier (which is the default for the `cl::list` class) indicates that your program will allow the option to be specified zero or more times.
- The **`cl::Required`** modifier indicates that the specified option must be specified exactly one time.
- The **`cl::OneOrMore`** modifier indicates that the option must be specified at least one time.
- The **`cl::ConsumeAfter`** modifier is described in the *Positional arguments section*.

If an option is not specified, then the value of the option is equal to the value specified by the `cl::init` attribute. If the `cl::init` attribute is not specified, the option value is initialized with the default constructor for the data type.

If an option is specified multiple times for an option of the `cl::opt` class, only the last value will be retained.

Controlling whether or not a value must be specified

This group of options is used to control whether or not the option allows a value to be present. In the case of the `CommandLine` library, a value is either specified with an equal sign (e.g. `-index-depth=17`) or as a trailing string (e.g. `-o a.out`).

The allowed values for this option group are:

- The **`cl::ValueOptional`** modifier (which is the default for `bool` typed options) specifies that it is acceptable to have a value, or not. A boolean argument can be enabled just by appearing on the command line, or it can have an explicit `-foo=true`. If an option is specified with this mode, it is illegal for the value to be provided without the equal sign. Therefore `-foo true` is illegal. To get this behavior, you must use the `cl::ValueRequired` modifier.

- The **cl::ValueRequired** modifier (which is the default for all other types except for *unnamed alternatives using the generic parser*) specifies that a value must be provided. This mode informs the command line library that if an option is not provided with an equal sign, that the next argument provided must be the value. This allows things like ‘-o a.out’ to work.
- The **cl::ValueDisallowed** modifier (which is the default for *unnamed alternatives using the generic parser*) indicates that it is a runtime error for the user to specify a value. This can be provided to disallow users from providing options to boolean options (like ‘-foo=true’).

In general, the default values for this option group work just like you would want them to. As mentioned above, you can specify the **cl::ValueDisallowed** modifier to a boolean argument to restrict your command line parser. These options are mostly useful when *extending the library*.

Controlling other formatting options

The formatting option group is used to specify that the command line option has special abilities and is otherwise different from other command line arguments. As usual, you can only specify one of these arguments at most.

- The **cl::NormalFormatting** modifier (which is the default all options) specifies that this option is “normal”.
- The **cl::Positional** modifier specifies that this is a positional argument that does not have a command line option associated with it. See the *Positional Arguments* section for more information.
- The **cl::ConsumeAfter** modifier specifies that this option is used to capture “interpreter style” arguments. See *this section for more information*.
- The **cl::Prefix** modifier specifies that this option prefixes its value. With ‘Prefix’ options, the equal sign does not separate the value from the option name specified. Instead, the value is everything after the prefix, including any equal sign if present. This is useful for processing odd arguments like `-lmalloc` and `-L/usr/lib` in a linker tool or `-DNAME=value` in a compiler tool. Here, the ‘l’, ‘D’ and ‘L’ options are normal string (or list) options, that have the **cl::Prefix** modifier added to allow the CommandLine library to recognize them. Note that **cl::Prefix** options must not have the **cl::ValueDisallowed** modifier specified.
- The **cl::Grouping** modifier is used to implement Unix-style tools (like `ls`) that have lots of single letter arguments, but only require a single dash. For example, the ‘`ls -labF`’ command actually enables four different options, all of which are single letters. Note that **cl::Grouping** options cannot have values.

The CommandLine library does not restrict how you use the **cl::Prefix** or **cl::Grouping** modifiers, but it is possible to specify ambiguous argument settings. Thus, it is possible to have multiple letter options that are prefix or grouping options, and they will still work as designed.

To do this, the CommandLine library uses a greedy algorithm to parse the input option into (potentially multiple) prefix and grouping options. The strategy basically looks like this:

```

parse(string OrigInput) {
1. string input = OrigInput;
2. if (isOption(input)) return getOption(input).parse(); // Normal option
3. while (!isOption(input) && !input.empty()) input.pop_back(); // Remove the last
   ↳ letter
4. if (input.empty()) return error(); // No matching option
5. if (getOption(input).isPrefix())
   return getOption(input).parse(input);
6. while (!input.empty()) { // Must be grouping options
   getOption(input).parse();
   OrigInput.erase(OrigInput.begin(), OrigInput.begin()+input.length());
   input = OrigInput;
   while (!isOption(input) && !input.empty()) input.pop_back();
}
}

```

```
7. if (!OrigInput.empty()) error();  
  
}
```

Miscellaneous option modifiers

The miscellaneous option modifiers are the only flags where you can specify more than one flag from the set: they are not mutually exclusive. These flags specify boolean properties that modify the option.

- The **`cl::CommaSeparated`** modifier indicates that any commas specified for an option's value should be used to split the value up into multiple values for the option. For example, these two options are equivalent when `cl::CommaSeparated` is specified: `"-foo=a -foo=b -foo=c"` and `"-foo=a,b,c"`. This option only makes sense to be used in a case where the option is allowed to accept one or more values (i.e. it is a [`cl::list`](#) option).
- The **`cl::PositionalEatsArgs`** modifier (which only applies to positional arguments, and only makes sense for lists) indicates that positional argument should consume any strings after it (including strings that start with a `"-"`) up until another recognized positional argument. For example, if you have two "eating" positional arguments, `"pos1"` and `"pos2"`, the string `"-pos1 -foo -bar baz -pos2 -bork"` would cause the `"-foo -bar -baz"` strings to be applied to the `"-pos1"` option and the `"-bork"` string to be applied to the `"-pos2"` option.
- The **`cl::Sink`** modifier is used to handle unknown options. If there is at least one option with `cl::Sink` modifier specified, the parser passes unrecognized option strings to it as values instead of signaling an error. As with `cl::CommaSeparated`, this modifier only makes sense with a [`cl::list`](#) option.

So far, these are the only three miscellaneous option modifiers.

Response files

Some systems, such as certain variants of Microsoft Windows and some older Unices have a relatively low limit on command-line length. It is therefore customary to use the so-called 'response files' to circumvent this restriction. These files are mentioned on the command-line (using the `"@file"`) syntax. The program reads these files and inserts the contents into `argv`, thereby working around the command-line length limits. Response files are enabled by an optional fourth argument to [`cl::ParseEnvironmentOptions`](#) and [`cl::ParseCommandLineOptions`](#).

Top-Level Classes and Functions

Despite all of the built-in flexibility, the `CommandLine` option library really only consists of one function ([`cl::ParseCommandLineOptions`](#)) and three main classes: [`cl::opt`](#), [`cl::list`](#), and [`cl::alias`](#). This section describes these three classes in detail.

The `cl::getRegisteredOptions` function

The `cl::getRegisteredOptions` function is designed to give a programmer access to declared non-positional command line options so that how they appear in `-help` can be modified prior to calling [`cl::ParseCommandLineOptions`](#). Note this method should not be called during any static initialisation because it cannot be guaranteed that all options will have been initialised. Hence it should be called from `main`.

This function can be used to gain access to options declared in libraries that the tool writer may not have direct access to.

The function retrieves a *StringMap* that maps the option string (e.g. `-help`) to an `Option*`.

Here is an example of how the function could be used:

```
using namespace llvm;
int main(int argc, char **argv) {
    cl::OptionCategory AnotherCategory("Some options");

    StringMap<cl::Option*> &Map = cl::getRegisteredOptions();

    //Unhide useful option and put it in a different category
    assert(Map.count("print-all-options") > 0);
    Map["print-all-options"]->setHiddenFlag(cl::NotHidden);
    Map["print-all-options"]->setCategory(AnotherCategory);

    //Hide an option we don't want to see
    assert(Map.count("enable-no-infs-fp-math") > 0);
    Map["enable-no-infs-fp-math"]->setHiddenFlag(cl::Hidden);

    //Change --version to --show-version
    assert(Map.count("version") > 0);
    Map["version"]->setArgStr("show-version");

    //Change --help description
    assert(Map.count("help") > 0);
    Map["help"]->setDescription("Shows help");

    cl::ParseCommandLineOptions(argc, argv, "This is a small program to demo the LLVM_
↳CommandLine API");
    ...
}
```

The `cl::ParseCommandLineOptions` function

The `cl::ParseCommandLineOptions` function is designed to be called directly from `main`, and is used to fill in the values of all of the command line option variables once `argc` and `argv` are available.

The `cl::ParseCommandLineOptions` function requires two parameters (`argc` and `argv`), but may also take an optional third parameter which holds *additional extra text* to emit when the `-help` option is invoked, and a fourth boolean parameter that enables *response files*.

The `cl::ParseEnvironmentOptions` function

The `cl::ParseEnvironmentOptions` function has mostly the same effects as *cl::ParseCommandLineOptions*, except that it is designed to take values for options from an environment variable, for those cases in which reading the command line is not convenient or desired. It fills in the values of all the command line option variables just like *cl::ParseCommandLineOptions* does.

It takes four parameters: the name of the program (since `argv` may not be available, it can't just look in `argv[0]`), the name of the environment variable to examine, the optional *additional extra text* to emit when the `-help` option is invoked, and the boolean switch that controls whether *response files* should be read.

`cl::ParseEnvironmentOptions` will break the environment variable's value up into words and then process them using *cl::ParseCommandLineOptions*. **Note:** Currently `cl::ParseEnvironmentOptions` does not support quoting, so an environment variable containing `-option "foo bar"` will be parsed as three words, `-option`, `"foo`, and `bar"`, which is different from what you would get from the shell with the same input.

The `cl::SetVersionPrinter` function

The `cl::SetVersionPrinter` function is designed to be called directly from `main` and *before* `cl::ParseCommandLineOptions`. Its use is optional. It simply arranges for a function to be called in response to the `--version` option instead of having the `CommandLine` library print out the usual version string for LLVM. This is useful for programs that are not part of LLVM but wish to use the `CommandLine` facilities. Such programs should just define a small function that takes no arguments and returns `void` and that prints out whatever version information is appropriate for the program. Pass the address of that function to `cl::SetVersionPrinter` to arrange for it to be called when the `--version` option is given by the user.

The `cl::opt` class

The `cl::opt` class is the class used to represent scalar command line options, and is the one used most of the time. It is a templated class which can take up to three arguments (all except for the first have default values though):

```
namespace cl {  
    template <class DataType, bool ExternalStorage = false,  
              class ParserClass = parser<DataType> >  
        class opt;  
}
```

The first template argument specifies what underlying data type the command line argument is, and is used to select a default parser implementation. The second template argument is used to specify whether the option should contain the storage for the option (the default) or whether external storage should be used to contain the value parsed for the option (see *Internal vs External Storage* for more information).

The third template argument specifies which parser to use. The default value selects an instantiation of the `parser` class based on the underlying data type of the option. In general, this default works well for most applications, so this option is only used when using a *custom parser*.

The `cl::list` class

The `cl::list` class is the class used to represent a list of command line options. It too is a templated class which can take up to three arguments:

```
namespace cl {  
    template <class DataType, class Storage = bool,  
              class ParserClass = parser<DataType> >  
        class list;  
}
```

This class works the exact same as the `cl::opt` class, except that the second argument is the **type** of the external storage, not a boolean value. For this class, the marker type 'bool' is used to indicate that internal storage should be used.

The `cl::bits` class

The `cl::bits` class is the class used to represent a list of command line options in the form of a bit vector. It is also a templated class which can take up to three arguments:

```
namespace cl {  
    template <class DataType, class Storage = bool,  
              class ParserClass = parser<DataType> >
```



```
class bits;
}
```

This class works the exact same as the `cl::list` class, except that the second argument must be of **type** `unsigned` if external storage is used.

The `cl::alias` class

The `cl::alias` class is a nontemplated class that is used to form aliases for other arguments.

```
namespace cl {
  class alias;
}
```

The `cl::aliasopt` attribute should be used to specify which option this is an alias for. Alias arguments default to being `cl::Hidden`, and use the aliased options parser to do the conversion from string to data.

The `cl::extrahelp` class

The `cl::extrahelp` class is a nontemplated class that allows extra help text to be printed out for the `-help` option.

```
namespace cl {
  struct extrahelp;
}
```

To use the `extrahelp`, simply construct one with a `const char*` parameter to the constructor. The text passed to the constructor will be printed at the bottom of the help message, verbatim. Note that multiple `cl::extrahelp` **can** be used, but this practice is discouraged. If your tool needs to print additional help information, put all that help into a single `cl::extrahelp` instance.

For example:

```
cl::extrahelp("\nADDITIONAL HELP:\n\n This is the extra help\n");
```

The `cl::OptionCategory` class

The `cl::OptionCategory` class is a simple class for declaring option categories.

```
namespace cl {
  class OptionCategory;
}
```

An option category must have a name and optionally a description which are passed to the constructor as `const char*`.

Note that declaring an option category and associating it with an option before parsing options (e.g. statically) will change the output of `-help` from uncategorized to categorized. If an option category is declared but not associated with an option then it will be hidden from the output of `-help` but will be shown in the output of `-help-hidden`.

Builtin parsers

Parsers control how the string value taken from the command line is translated into a typed value, suitable for use in a C++ program. By default, the CommandLine library uses an instance of `parser<type>` if the command line option specifies that it uses values of type ‘type’. Because of this, custom option processing is specified with specializations of the ‘parser’ class.

The CommandLine library provides the following builtin parser specializations, which are sufficient for most applications. It can, however, also be extended to work with new data types and new ways of interpreting the same data. See the [Writing a Custom Parser](#) for more details on this type of library extension.

- The generic `parser<t>` parser can be used to map strings values to any data type, through the use of the [cl::values](#) property, which specifies the mapping information. The most common use of this parser is for parsing enum values, which allows you to use the CommandLine library for all of the error checking to make sure that only valid enum values are specified (as opposed to accepting arbitrary strings). Despite this, however, the generic parser class can be used for any data type.
- The **`parser<bool>` specialization** is used to convert boolean strings to a boolean value. Currently accepted strings are “true”, “TRUE”, “True”, “1”, “false”, “FALSE”, “False”, and “0”.
- The **`parser<boolOrDefault>` specialization** is used for cases where the value is boolean, but we also need to know whether the option was specified at all. `boolOrDefault` is an enum with 3 values, `BOU_UNSET`, `BOU_TRUE` and `BOU_FALSE`. This parser accepts the same strings as “**`parser<bool>`”.**
- The **`parser<string>` specialization** simply stores the parsed string into the string value specified. No conversion or modification of the data is performed.
- The **`parser<int>` specialization** uses the C `strtol` function to parse the string input. As such, it will accept a decimal number (with an optional ‘+’ or ‘-’ prefix) which must start with a non-zero digit. It accepts octal numbers, which are identified with a ‘0’ prefix digit, and hexadecimal numbers with a prefix of ‘0x’ or ‘0X’.
- The **`parser<double>` and `parser<float>` specializations** use the standard C `strtod` function to convert floating point strings into floating point values. As such, a broad range of string formats is supported, including exponential notation (ex: `1.7e15`) and properly supports locales.

3.3.4 Extension Guide

Although the CommandLine library has a lot of functionality built into it already (as discussed previously), one of its true strengths lie in its extensibility. This section discusses how the CommandLine library works under the covers and illustrates how to do some simple, common, extensions.

Writing a custom parser

One of the simplest and most common extensions is the use of a custom parser. As [discussed previously](#), parsers are the portion of the CommandLine library that turns string input from the user into a particular parsed data type, validating the input in the process.

There are two ways to use a new parser:

1. Specialize the [cl::parser](#) template for your custom data type.

This approach has the advantage that users of your custom data type will automatically use your custom parser whenever they define an option with a value type of your data type. The disadvantage of this approach is that it doesn’t work if your fundamental data type is something that is already supported.

2. Write an independent class, using it explicitly from options that need it.

This approach works well in situations where you would line to parse an option using special syntax for a not-very-special data-type. The drawback of this approach is that users of your parser have to be aware that they are using your parser instead of the builtin ones.

To guide the discussion, we will discuss a custom parser that accepts file sizes, specified with an optional unit after the numeric size. For example, we would like to parse “102kb”, “41M”, “1G” into the appropriate integer value. In this case, the underlying data type we want to parse into is ‘unsigned’. We choose approach #2 above because we don’t want to make this the default for all unsigned options.

To start out, we declare our new `FileSizeParser` class:

```
struct FileSizeParser : public cl::parser<unsigned> {
    // parse - Return true on error.
    bool parse(cl::Option &O, StringRef ArgName, const std::string &ArgValue,
              unsigned &Val);
};
```

Our new class inherits from the `cl::parser` template class to fill in the default, boiler plate code for us. We give it the data type that we parse into, the last argument to the `parse` method, so that clients of our custom parser know what object type to pass in to the `parse` method. (Here we declare that we parse into ‘unsigned’ variables.)

For most purposes, the only method that must be implemented in a custom parser is the `parse` method. The `parse` method is called whenever the option is invoked, passing in the option itself, the option name, the string to parse, and a reference to a return value. If the string to parse is not well-formed, the parser should output an error message and return true. Otherwise it should return false and set ‘Val’ to the parsed value. In our example, we implement `parse` as:

```
bool FileSizeParser::parse(cl::Option &O, StringRef ArgName,
                          const std::string &Arg, unsigned &Val) {
    const char *ArgStart = Arg.c_str();
    char *End;

    // Parse integer part, leaving 'End' pointing to the first non-integer char
    Val = (unsigned) strtol(ArgStart, &End, 0);

    while (1) {
        switch (*End++) {
            case 0: return false; // No error
            case 'i': // Ignore the 'i' in KiB if people use that
            case 'b': case 'B': // Ignore B suffix
                break;

            case 'g': case 'G': Val *= 1024*1024*1024; break;
            case 'm': case 'M': Val *= 1024*1024; break;
            case 'k': case 'K': Val *= 1024; break;

            default:
                // Print an error message if unrecognized character!
                return O.error("'" + Arg + "' value invalid for file size argument!");
        }
    }
}
```

This function implements a very simple parser for the kinds of strings we are interested in. Although it has some holes (it allows “123KKK” for example), it is good enough for this example. Note that we use the option itself to print out the error message (the `error` method always returns true) in order to get a nice error message (shown below). Now that we have our parser class, we can use it like this:

```
static cl::opt<unsigned, false, FileSizeParser>
MFS("max-file-size", cl::desc("Maximum file size to accept"),
    cl::value_desc("size"));
```

Which adds this to the output of our program:

```
OPTIONS:
  -help                - display available options (-help-hidden for more)
  ...
  -max-file-size=<size> - Maximum file size to accept
```

And we can test that our parse works correctly now (the test program just prints out the max-file-size argument value):

```
$ ./test
MFS: 0
$ ./test -max-file-size=123MB
MFS: 128974848
$ ./test -max-file-size=3G
MFS: 3221225472
$ ./test -max-file-size=dog
-max-file-size option: 'dog' value invalid for file size argument!
```

It looks like it works. The error message that we get is nice and helpful, and we seem to accept reasonable file sizes. This wraps up the “custom parser” tutorial.

Exploiting external storage

Several of the LLVM libraries define static `cl::opt` instances that will automatically be included in any program that links with that library. This is a feature. However, sometimes it is necessary to know the value of the command line option outside of the library. In these cases the library does or should provide an external storage location that is accessible to users of the library. Examples of this include the `llvm::DebugFlag` exported by the `lib/Support/Debug.cpp` file and the `llvm::TimePassesIsEnabled` flag exported by the `lib/VMCore/PassManager.cpp` file.

Dynamically adding command line options

3.4 Architecture & Platform Information for Compiler Writers

- *Hardware*
 - *AArch64 & ARM*
 - *Itanium (ia64)*
 - *Lanai*
 - *MIPS*
 - *PowerPC*
 - * *IBM - Official manuals and docs*
 - * *Other documents, collections, notes*

- *AMDGPU*
 - *RISC-V*
 - *SPARC*
 - *SystemZ*
 - *X86*
 - *XCore*
 - *Hexagon*
 - *Other relevant lists*
- *ABI*
 - *Linux*
 - *OS X*
 - *Windows*
- *NVPTX*
- *Miscellaneous Resources*

Note: This document is a work-in-progress. Additions and clarifications are welcome.

3.4.1 Hardware

AArch64 & ARM

- [ARMv8-A Architecture Reference Manual](#) (authentication required, free sign-up). This document covers both AArch64 and ARM instructions
- [ARMv7-M Architecture Reference Manual](#) (authentication required, free sign-up). This covers the Thumb2-only microcontrollers
- [ARMv6-M Architecture Reference Manual](#) (authentication required, free sign-up). This covers the Thumb1-only microcontrollers
- [ARM C Language Extensions](#)
- [AArch32 ABI Addenda and Errata](#)

Itanium (ia64)

- [Itanium documentation](#)

Lanai

- [Lanai Instruction Set Architecture](#)

MIPS

- [MIPS Processor Architecture](#)
- [MIPS 64-bit ELF Object File Specification](#)

PowerPC

IBM - Official manuals and docs

- [Power Instruction Set Architecture, Versions 2.03 through 2.06](#) (authentication required, free sign-up)
- [PowerPC Compiler Writer's Guide](#)
- [Intro to PowerPC Architecture](#)
- [PowerPC Processor Manuals](#) (embedded)
- [Various IBM specifications and white papers](#)
- [IBM AIX/5L for POWER Assembly Reference](#)

Other documents, collections, notes

- [PowerPC ABI documents](#)
- [PowerPC64 alignment of long doubles](#) (from GCC)
- [Long branch stubs for powerpc64-linux](#) (from binutils)

AMDGPU

Refer to *User Guide for AMDGPU Backend* for additional documentation.

RISC-V

- [RISC-V User-Level ISA Specification](#)

SPARC

- [SPARC standards](#)
- [SPARC V9 ABI](#)
- [SPARC V8 ABI](#)

SystemZ

- [z/Architecture Principles of Operation](#) (registration required, free sign-up)

X86

- AMD processor manuals
- Intel 64 and IA-32 manuals
- Intel Itanium documentation
- X86 and X86-64 SysV psABI
- Calling conventions for different C++ compilers and operating systems

XCore

- The XMOS XS1 Architecture (ISA)
- Tools Development Guide (includes ABI)

Hexagon

- Hexagon Programmer's Reference Manuals and Hexagon ABI Specification (registration required, free sign-up)

Other relevant lists

- GCC reading list

3.4.2 ABI

- System V Application Binary Interface
- Itanium C++ ABI

Linux

- Linux extensions to gabi
- PowerPC 64-bit ELF ABI Supplement
- Procedure Call Standard for the AArch64 Architecture
- ELF for the ARM Architecture
- ELF for the ARM 64-bit Architecture (AArch64)
- System z ELF ABI Supplement

OS X

- Mach-O Runtime Architecture
- Notes on Mach-O ABI

Windows

- Microsoft PE/COFF Specification

3.4.3 NVPTX

- [CUDA Documentation](#) includes the PTX ISA and Driver API documentation

3.4.4 Miscellaneous Resources

- [Executable File Format library](#)
- [GCC prefetch project](#) page has a good survey of the prefetching capabilities of a variety of modern processors.

3.5 Extending LLVM: Adding instructions, intrinsics, types, etc.

3.5.1 Introduction and Warning

During the course of using LLVM, you may wish to customize it for your research project or for experimentation. At this point, you may realize that you need to add something to LLVM, whether it be a new fundamental type, a new intrinsic function, or a whole new instruction.

When you come to this realization, stop and think. Do you really need to extend LLVM? Is it a new fundamental capability that LLVM does not support at its current incarnation or can it be synthesized from already pre-existing LLVM elements? If you are not sure, ask on the [LLVM-dev](#) list. The reason is that extending LLVM will get involved as you need to update all the different passes that you intend to use with your extension, and there are many LLVM analyses and transformations, so it may be quite a bit of work.

Adding an *intrinsic function* is far easier than adding an instruction, and is transparent to optimization passes. If your added functionality can be expressed as a function call, an intrinsic function is the method of choice for LLVM extension.

Before you invest a significant amount of effort into a non-trivial extension, **ask on the list** if what you are looking to do can be done with already-existing infrastructure, or if maybe someone else is already working on it. You will save yourself a lot of time and effort by doing so.

3.5.2 Adding a new intrinsic function

Adding a new intrinsic function to LLVM is much easier than adding a new instruction. Almost all extensions to LLVM should start as an intrinsic function and then be turned into an instruction if warranted.

1. `llvm/docs/LangRef.html`:

Document the intrinsic. Decide whether it is code generator specific and what the restrictions are. Talk to other people about it so that you are sure it's a good idea.

2. `llvm/include/llvm/IR/Intrinsics*.td`:

Add an entry for your intrinsic. Describe its memory access characteristics for optimization (this controls whether it will be DCE'd, CSE'd, etc). Note that any intrinsic using one of the `llvm_any*_ty` types for an argument or return type will be deemed by `tblgen` as overloaded and the corresponding suffix will be required on the intrinsic's name.

3. `llvm/lib/Analysis/ConstantFolding.cpp`:

If it is possible to constant fold your intrinsic, add support to it in the `canConstantFoldCallTo` and `ConstantFoldCall` functions.

4. `llvm/test/*:`

Add test cases for your test cases to the test suite

Once the intrinsic has been added to the system, you must add code generator support for it. Generally you must do the following steps:

Add support to the `.td` file for the target(s) of your choice in `lib/Target/*/*.td`.

This is usually a matter of adding a pattern to the `.td` file that matches the intrinsic, though it may obviously require adding the instructions you want to generate as well. There are lots of examples in the PowerPC and X86 backend to follow.

3.5.3 Adding a new SelectionDAG node

As with intrinsics, adding a new SelectionDAG node to LLVM is much easier than adding a new instruction. New nodes are often added to help represent instructions common to many targets. These nodes often map to an LLVM instruction (`add`, `sub`) or intrinsic (`byteswap`, `population count`). In other cases, new nodes have been added to allow many targets to perform a common task (converting between floating point and integer representation) or capture more complicated behavior in a single node (`rotate`).

1. `include/llvm/CodeGen/ISDOpcodes.h:`

Add an enum value for the new SelectionDAG node.

2. `lib/CodeGen/SelectionDAG/SelectionDAG.cpp:`

Add code to print the node to `getOperationName`. If your new node can be evaluated at compile time when given constant arguments (such as an `add` of a constant with another constant), find the `getNode` method that takes the appropriate number of arguments, and add a case for your node to the switch statement that performs constant folding for nodes that take the same number of arguments as your new node.

3. `lib/CodeGen/SelectionDAG/LegalizedAG.cpp:`

Add code to **legalize, promote, and expand** the node as necessary. At a minimum, you will need to add a case statement for your node in `LegalizeOp` which calls `LegalizeOp` on the node's operands, and returns a new node if any of the operands changed as a result of being legalized. It is likely that not all targets supported by the SelectionDAG framework will natively support the new node. In this case, you must also add code in your node's case statement in `LegalizeOp` to **Expand** your node into simpler, legal operations. The case for `ISD::UREM` for expanding a remainder into a divide, multiply, and a subtract is a good example.

4. `lib/CodeGen/SelectionDAG/LegalizedAG.cpp:`

If targets may support the new node being added only at certain sizes, you will also need to add code to your node's case statement in `LegalizeOp` to **Promote** your node's operands to a larger size, and perform the correct operation. You will also need to add code to `PromoteOp` to do this as well. For a good example, see `ISD::BSWAP`, which promotes its operand to a wider size, performs the byteswap, and then shifts the correct bytes right to emulate the narrower byteswap in the wider type.

5. `lib/CodeGen/SelectionDAG/LegalizedAG.cpp:`

Add a case for your node in `ExpandOp` to teach the legalizer how to perform the action represented by the new node on a value that has been split into high and low halves. This case will be used to support your node with a 64 bit operand on a 32 bit target.

6. `lib/CodeGen/SelectionDAG/DAGCombiner.cpp:`

If your node can be combined with itself, or other existing nodes in a peephole-like fashion, add a visit function for it, and call that function from. There are several good examples for simple combines you can do; `visitFABS` and `visitSRL` are good starting places.

7. `lib/Target/PowerPC/PPCISelLowering.cpp`:

Each target has an implementation of the `TargetLowering` class, usually in its own file (although some targets include it in the same file as the `DAGToDAGISel`). The default behavior for a target is to assume that your new node is legal for all types that are legal for that target. If this target does not natively support your node, then tell the target to either Promote it (if it is supported at a larger type) or Expand it. This will cause the code you wrote in `LegalizeOp` above to decompose your new node into other legal nodes for this target.

8. `lib/Target/TargetSelectionDAG.td`:

Most current targets supported by LLVM generate code using the DAGToDAG method, where `SelectionDAG` nodes are pattern matched to target-specific nodes, which represent individual instructions. In order for the targets to match an instruction to your new node, you must add a def for that node to the list in this file, with the appropriate type constraints. Look at `add`, `bswap`, and `fadd` for examples.

9. `lib/Target/PowerPC/PPCInstrInfo.td`:

Each target has a `tablegen` file that describes the target's instruction set. For targets that use the DAGToDAG instruction selection framework, add a pattern for your new node that uses one or more target nodes. Documentation for this is a bit sparse right now, but there are several decent examples. See the patterns for `rotl` in `PPCInstrInfo.td`.

10. TODO: document complex patterns.

11. `llvm/test/CodeGen/*`:

Add test cases for your new node to the test suite. `llvm/test/CodeGen/X86/bswap.ll` is a good example.

3.5.4 Adding a new instruction

Warning: Adding instructions changes the bitcode format, and it will take some effort to maintain compatibility with the previous version. Only add an instruction if it is absolutely necessary.

1. `llvm/include/llvm/IR/Instruction.def`:

add a number for your instruction and an enum name

2. `llvm/include/llvm/IR/Instructions.h`:

add a definition for the class that will represent your instruction

3. `llvm/include/llvm/IR/InstVisitor.h`:

add a prototype for a visitor to your new instruction type

4. `llvm/lib/AsmParser/LLLexer.cpp`:

add a new token to parse your instruction from assembly text file

5. `llvm/lib/AsmParser/LLParser.cpp`:

add the grammar on how your instruction can be read and what it will construct as a result

6. `llvm/lib/Bitcode/Reader/BitcodeReader.cpp`:

add a case for your instruction and how it will be parsed from bitcode

7. `llvm/lib/Bitcode/Writer/BitcodeWriter.cpp`:

add a case for your instruction and how it will be parsed from bitcode

8. `llvm/lib/IR/Instruction.cpp`:
add a case for how your instruction will be printed out to assembly
9. `llvm/lib/IR/Instructions.cpp`:
implement the class you defined in `llvm/include/llvm/Instructions.h`
10. Test your instruction
11. `llvm/lib/Target/*`:
add support for your instruction to code generators, or add a lowering pass.
12. `llvm/test/*`:
add your test cases to the test suite.

Also, you need to implement (or modify) any analyses or passes that you want to understand this new instruction.

3.5.5 Adding a new type

Warning: Adding new types changes the bitcode format, and will break compatibility with currently-existing LLVM installations. Only add new types if it is absolutely necessary.

Adding a fundamental type

1. `llvm/include/llvm/IR/Type.h`:
add enum for the new type; add static `Type*` for this type
2. `llvm/lib/IR/Type.cpp` and `llvm/lib/IR/ValueTypes.cpp`:
add mapping from `TypeID => Type*`; initialize the static `Type*`
3. `llvm/llvm-llvm-c/Core.cpp`:
add enum `LLVMTypeKind` and modify `LLVMTypeKind LLVMGetTypeKind(LLVMTypeRef Ty)` for the new type
4. `llvm/include/llvm/IR/TypeBuilder.h`:
add new class to represent new type in the hierarchy
5. `llvm/lib/AsmParser/LLLexer.cpp`:
add ability to parse in the type from text assembly
6. `llvm/lib/AsmParser/LLParser.cpp`:
add a token for that type
7. `llvm/lib/Bitcode/Writer/BitcodeWriter.cpp`:
modify `static void WriteTypeTable(const ValueEnumerator &VE, BitstreamWriter &Stream)` to serialize your type
8. `llvm/lib/Bitcode/Reader/BitcodeReader.cpp`:
modify `bool BitcodeReader::ParseTypeType()` to read your data type
9. `include/llvm/Bitcode/LLVMBitCodes.h`:
add enum `TypeCodes` for the new type

Adding a derived type

1. `llvm/include/llvm/IR/Type.h`:
add enum for the new type; add a forward declaration of the type also
2. `llvm/include/llvm/IR/DerivedTypes.h`:
add new class to represent new class in the hierarchy; add forward declaration to the `TypeMap` value type
3. `llvm/lib/IR/Type.cpp` and `llvm/lib/IR/ValueTypes.cpp`:
add support for derived type, notably *enum TypeID* and *is, get* methods.
4. `llvm/llvm/llvm-c/Core.cpp`:
add enum `LLVMTypeKind` and modify *LLVMTypeKind LLVMGetTypeKind(LLVMTypeRefTy)* for the new type
5. `llvm/include/llvm/IR/TypeBuilder.h`:
add new class to represent new class in the hierarchy
6. `llvm/lib/AsmParser/LLLexer.cpp`:
modify `lltok::Kind LLLexer::LexIdentifier()` to add ability to parse in the type from text assembly
7. `llvm/lib/Bitcode/Writer/BitcodeWriter.cpp`:
modify `static void WriteTypeTable(const ValueEnumerator &VE, BitstreamWriter &Stream)` to serialize your type
8. `llvm/lib/Bitcode/Reader/BitcodeReader.cpp`:
modify `bool BitcodeReader::ParseTypeType()` to read your data type
9. `include/llvm/Bitcode/LLVMBitCodes.h`:
add enum `TypeCodes` for the new type
10. `llvm/lib/IR/AsmWriter.cpp`:
modify `void TypePrinting::print(Type *Ty, raw_ostream &OS)` to output the new derived type

3.6 How to set up LLVM-style RTTI for your class hierarchy

Contents

- *How to set up LLVM-style RTTI for your class hierarchy*
 - *Background*
 - *Basic Setup*
 - *Concrete Bases and Deeper Hierarchies*
 - * *A Bug to be Aware Of*
 - * *The Contract of `classof`*
 - *Rules of Thumb*

3.6.1 Background

LLVM avoids using C++’s built in RTTI. Instead, it pervasively uses its own hand-rolled form of RTTI which is much more efficient and flexible, although it requires a bit more work from you as a class author.

A description of how to use LLVM-style RTTI from a client’s perspective is given in the [Programmer’s Manual](#). This document, in contrast, discusses the steps you need to take as a class hierarchy author to make LLVM-style RTTI available to your clients.

Before diving in, make sure that you are familiar with the Object Oriented Programming concept of “is-a”.

3.6.2 Basic Setup

This section describes how to set up the most basic form of LLVM-style RTTI (which is sufficient for 99.9% of the cases). We will set up LLVM-style RTTI for this class hierarchy:

```
class Shape {
public:
    Shape() {}
    virtual double computeArea() = 0;
};

class Square : public Shape {
    double SideLength;
public:
    Square(double S) : SideLength(S) {}
    double computeArea() override;
};

class Circle : public Shape {
    double Radius;
public:
    Circle(double R) : Radius(R) {}
    double computeArea() override;
};
```

The most basic working setup for LLVM-style RTTI requires the following steps:

1. In the header where you declare Shape, you will want to `#include "llvm/Support/Casting.h"`, which declares LLVM’s RTTI templates. That way your clients don’t even have to think about it.

```
#include "llvm/Support/Casting.h"
```

2. In the base class, introduce an enum which discriminates all of the different concrete classes in the hierarchy, and stash the enum value somewhere in the base class.

Here is the code after introducing this change:

```
class Shape {
public:
+   /// Discriminator for LLVM-style RTTI (dyn_cast<> et al.)
+   enum ShapeKind {
+       SK_Square,
+       SK_Circle
+   };
+private:
+   const ShapeKind Kind;
+public:
```

```
+ ShapeKind getKind() const { return Kind; }
+
  Shape() {}
  virtual double computeArea() = 0;
};
```

You will usually want to keep the `Kind` member encapsulated and private, but let the enum `ShapeKind` be public along with providing a `getKind()` method. This is convenient for clients so that they can do a switch over the enum.

A common naming convention is that these enums are “kind”s, to avoid ambiguity with the words “type” or “class” which have overloaded meanings in many contexts within LLVM. Sometimes there will be a natural name for it, like “opcode”. Don’t bikeshed over this; when in doubt use `Kind`.

You might wonder why the `Kind` enum doesn’t have an entry for `Shape`. The reason for this is that since `Shape` is abstract (`computeArea() = 0;`), you will never actually have non-derived instances of exactly that class (only subclasses). See *Concrete Bases and Deeper Hierarchies* for information on how to deal with non-abstract bases. It’s worth mentioning here that unlike `dynamic_cast<>`, LLVM-style RTTI can be used (and is often used) for classes that don’t have v-tables.

3. Next, you need to make sure that the `Kind` gets initialized to the value corresponding to the dynamic type of the class. Typically, you will want to have it be an argument to the constructor of the base class, and then pass in the respective `XXXXKind` from subclass constructors.

Here is the code after that change:

```
class Shape {
public:
  /// Discriminator for LLVM-style RTTI (dyn_cast<> et al.)
  enum ShapeKind {
    SK_Square,
    SK_Circle
  };
private:
  const ShapeKind Kind;
public:
  ShapeKind getKind() const { return Kind; }

- Shape() {}
+ Shape(ShapeKind K) : Kind(K) {}
  virtual double computeArea() = 0;
};

class Square : public Shape {
  double SideLength;
public:
- Square(double S) : SideLength(S) {}
+ Square(double S) : Shape(SK_Square), SideLength(S) {}
  double computeArea() override;
};

class Circle : public Shape {
  double Radius;
public:
- Circle(double R) : Radius(R) {}
+ Circle(double R) : Shape(SK_Circle), Radius(R) {}
  double computeArea() override;
};
```

4. Finally, you need to inform LLVM’s RTTI templates how to dynamically determine the type of a class (i.e. whether the `isa<>/dyn_cast<>` should succeed). The default “99.9% of use cases” way to accomplish this is through a small static member function `classof`. In order to have proper context for an explanation, we will display this code first, and then below describe each part:

```
class Shape {
public:
    /// Discriminator for LLVM-style RTTI (dyn_cast<> et al.)
    enum ShapeKind {
        SK_Square,
        SK_Circle
    };
private:
    const ShapeKind Kind;
public:
    ShapeKind getKind() const { return Kind; }

    Shape(ShapeKind K) : Kind(K) {}
    virtual double computeArea() = 0;
};

class Square : public Shape {
    double SideLength;
public:
    Square(double S) : Shape(SK_Square), SideLength(S) {}
    double computeArea() override;
+
+ static bool classof(const Shape *S) {
+     return S->getKind() == SK_Square;
+ }
};

class Circle : public Shape {
    double Radius;
public:
    Circle(double R) : Shape(SK_Circle), Radius(R) {}
    double computeArea() override;
+
+ static bool classof(const Shape *S) {
+     return S->getKind() == SK_Circle;
+ }
};
```

The job of `classof` is to dynamically determine whether an object of a base class is in fact of a particular derived class. In order to downcast a type `Base` to a type `Derived`, there needs to be a `classof` in `Derived` which will accept an object of type `Base`.

To be concrete, consider the following code:

```
Shape *S = ...;
if (isa<Circle>(S)) {
    /* do something ... */
}
```

The code of the `isa<>` test in this code will eventually boil down—after template instantiation and some other machinery—to a check roughly like `Circle::classof(S)`. For more information, see [The Contract of `classof`](#).

The argument to `classof` should always be an *ancestor* class because the implementation has logic to allow

and optimize away upcasts/up-isa<>’s automatically. It is as though every class `Foo` automatically has a `classof` like:

```
class Foo {
    [...]
    template <class T>
    static bool classof(const T *,
                       ::std::enable_if<
                           ::std::is_base_of<Foo, T>::value
                       >::type* = 0) { return true; }
    [...]
};
```

Note that this is the reason that we did not need to introduce a `classof` into `Shape`: all relevant classes derive from `Shape`, and `Shape` itself is abstract (has no entry in the `Kind` enum), so this notional inferred `classof` is all we need. See *Concrete Bases and Deeper Hierarchies* for more information about how to extend this example to more general hierarchies.

Although for this small example setting up LLVM-style RTTI seems like a lot of “boilerplate”, if your classes are doing anything interesting then this will end up being a tiny fraction of the code.

3.6.3 Concrete Bases and Deeper Hierarchies

For concrete bases (i.e. non-abstract interior nodes of the inheritance tree), the `Kind` check inside `classof` needs to be a bit more complicated. The situation differs from the example above in that

- Since the class is concrete, it must itself have an entry in the `Kind` enum because it is possible to have objects with this class as a dynamic type.
- Since the class has children, the check inside `classof` must take them into account.

Say that `SpecialSquare` and `OtherSpecialSquare` derive from `Square`, and so `ShapeKind` becomes:

```
enum ShapeKind {
    SK_Square,
+   SK_SpecialSquare,
+   SK_OtherSpecialSquare,
    SK_Circle
}
```

Then in `Square`, we would need to modify the `classof` like so:

```
- static bool classof(const Shape *S) {
-     return S->getKind() == SK_Square;
- }
+ static bool classof(const Shape *S) {
+     return S->getKind() >= SK_Square &&
+           S->getKind() <= SK_OtherSpecialSquare;
+ }
```

The reason that we need to test a range like this instead of just equality is that both `SpecialSquare` and `OtherSpecialSquare` “is-a” `Square`, and so `classof` needs to return `true` for them.

This approach can be made to scale to arbitrarily deep hierarchies. The trick is that you arrange the enum values so that they correspond to a preorder traversal of the class hierarchy tree. With that arrangement, all subclass tests can be done with two comparisons as shown above. If you just list the class hierarchy like a list of bullet points, you’ll get the ordering right:


```
| Shape
|   Square
|     SpecialSquare
|     OtherSpecialSquare
|   Circle
```

A Bug to be Aware Of

The example just given opens the door to bugs where the `classofs` are not updated to match the `Kind` enum when adding (or removing) classes to (from) the hierarchy.

Continuing the example above, suppose we add a `SomewhatSpecialSquare` as a subclass of `Square`, and update the `ShapeKind` enum like so:

```
enum ShapeKind {
    SK_Square,
    SK_SpecialSquare,
    SK_OtherSpecialSquare,
+   SK_SomewhatSpecialSquare,
    SK_Circle
}
```

Now, suppose that we forget to update `Square::classof()`, so it still looks like:

```
static bool classof(const Shape *S) {
    // BUG: Returns false when S->getKind() == SK_SomewhatSpecialSquare,
    // even though SomewhatSpecialSquare "is a" Square.
    return S->getKind() >= SK_Square &&
           S->getKind() <= SK_OtherSpecialSquare;
}
```

As the comment indicates, this code contains a bug. A straightforward and non-clever way to avoid this is to introduce an explicit `SK_LastSquare` entry in the enum when adding the first subclass(es). For example, we could rewrite the example at the beginning of *Concrete Bases and Deeper Hierarchies* as:

```
enum ShapeKind {
    SK_Square,
+   SK_SpecialSquare,
+   SK_OtherSpecialSquare,
+   SK_LastSquare,
    SK_Circle
}
...
// Square::classof()
- static bool classof(const Shape *S) {
-     return S->getKind() == SK_Square;
- }
+ static bool classof(const Shape *S) {
+     return S->getKind() >= SK_Square &&
+           S->getKind() <= SK_LastSquare;
+ }
```

Then, adding new subclasses is easy:

```
enum ShapeKind {
    SK_Square,
```

```
    SK_SpecialSquare,  
    SK_OtherSpecialSquare,  
+   SK_SomewhatSpecialSquare,  
    SK_LastSquare,  
    SK_Circle  
}
```

Notice that `Square::classof` does not need to be changed.

The Contract of `classof`

To be more precise, let `classof` be inside a class `C`. Then the contract for `classof` is “return `true` if the dynamic type of the argument is-a `C`”. As long as your implementation fulfills this contract, you can tweak and optimize it as much as you want.

For example, LLVM-style RTTI can work fine in the presence of multiple-inheritance by defining an appropriate `classof`. An example of this in practice is `Decl` vs. `DeclContext` inside Clang. The `Decl` hierarchy is done very similarly to the example setup demonstrated in this tutorial. The key part is how to then incorporate `DeclContext`: all that is needed is in `bool DeclContext::classof(const Decl *)`, which asks the question “Given a `Decl`, how can I determine if it is-a `DeclContext`?”. It answers this with a simple switch over the set of `Decl` “kinds”, and returning `true` for ones that are known to be `DeclContext`’s.

3.6.4 Rules of Thumb

1. The `Kind` enum should have one entry per concrete class, ordered according to a preorder traversal of the inheritance tree.
2. The argument to `classof` should be a `const Base *`, where `Base` is some ancestor in the inheritance hierarchy. The argument should *never* be a derived class or the class itself: the template machinery for `isa<>` already handles this case and optimizes it.
3. For each class in the hierarchy that has no children, implement a `classof` that checks only against its `Kind`.
4. For each class in the hierarchy that has children, implement a `classof` that checks a range of the first child’s `Kind` and the last child’s `Kind`.

3.7 LLVM Programmer’s Manual

- *Introduction*
- *General Information*
 - *The C++ Standard Template Library*
 - *Other useful references*
- *Important and useful LLVM APIs*
 - *The `isa<>`, `cast<>` and `dyn_cast<>` templates*
 - *Passing strings (the `StringRef` and `Twine` classes)*
 - * *The `StringRef` class*

- * *The Twine class*
- *Formatting strings (the `formatv` function)*
 - * *Simple formatting*
 - * *Custom formatting*
 - * *`formatv` Examples*
- *Error handling*
 - * *Programmatic Errors*
 - * *Recoverable Errors*
 - *StringError*
 - *Interoperability with `std::error_code` and `ErrorOr`*
 - *Returning Errors from error handlers*
 - *Using `ExitOnError` to simplify tool code*
 - *Using `cantFail` to simplify safe callsites*
 - *Fallible constructors*
 - *Propagating and consuming errors based on types*
 - *Concatenating Errors with `joinErrors`*
 - *Building fallible iterators and iterator ranges*
- *Passing functions and other callable objects*
 - * *Function template*
 - * *The `function_ref` class template*
- *The `DEBUG()` macro and `-debug` option*
 - * *Fine grained debug info with `DEBUG_TYPE` and the `-debug-only` option*
- *The `Statistic` class & `-stats` option*
- *Adding debug counters to aid in debugging your code*
- *Viewing graphs while debugging code*
- *Picking the Right Data Structure for a Task*
 - *Sequential Containers (`std::vector`, `std::list`, etc)*
 - * *`llvm/ADT/ArrayRef.h`*
 - * *Fixed Size Arrays*
 - * *Heap Allocated Arrays*
 - * *`llvm/ADT/TinyPtrVector.h`*
 - * *`llvm/ADT/SmallVector.h`*
 - * *`<vector>`*
 - * *`<deque>`*
 - * *`<list>`*

- * *llvm/ADT/ilist.h*
- * *llvm/ADT/PackedVector.h*
- * *ilist_traits*
- * *iplist*
- * *llvm/ADT/ilist_node.h*
- * *Sentinels*
- * *Other Sequential Container options*
- *String-like containers*
 - * *llvm/ADT/StringRef.h*
 - * *llvm/ADT/Twine.h*
 - * *llvm/ADT/SmallString.h*
 - * *std::string*
- *Set-Like Containers (std::set, SmallSet, SetVector, etc)*
 - * *A sorted ‘vector’*
 - * *llvm/ADT/SmallSet.h*
 - * *llvm/ADT/SmallPtrSet.h*
 - * *llvm/ADT/StringSet.h*
 - * *llvm/ADT/DenseSet.h*
 - * *llvm/ADT/SparseSet.h*
 - * *llvm/ADT/SparseMultiSet.h*
 - * *llvm/ADT/FoldingSet.h*
 - * *<set>*
 - * *llvm/ADT/SetVector.h*
 - * *llvm/ADT/UniqueVector.h*
 - * *llvm/ADT/ImmutableSet.h*
 - * *Other Set-Like Container Options*
- *Map-Like Containers (std::map, DenseMap, etc)*
 - * *A sorted ‘vector’*
 - * *llvm/ADT/StringMap.h*
 - * *llvm/ADT/IndexedMap.h*
 - * *llvm/ADT/DenseMap.h*
 - * *llvm/IR/ValueMap.h*
 - * *llvm/ADT/IntervalMap.h*
 - * *<map>*
 - * *llvm/ADT/MapVector.h*

- * *llvm/ADT/IntEqClasses.h*
- * *llvm/ADT/ImmutableMap.h*
- * *Other Map-Like Container Options*
- *Bit storage containers (BitVector, SparseBitVector)*
 - * *BitVector*
 - * *SmallBitVector*
 - * *SparseBitVector*
- *Debugging*
- *Helpful Hints for Common Operations*
 - *Basic Inspection and Traversal Routines*
 - * *Iterating over the BasicBlock in a Function*
 - * *Iterating over the Instruction in a BasicBlock*
 - * *Iterating over the Instruction in a Function*
 - * *Turning an iterator into a class pointer (and vice-versa)*
 - * *Finding call sites: a slightly more complex example*
 - * *Treating calls and invokes the same way*
 - * *Iterating over def-use & use-def chains*
 - * *Iterating over predecessors & successors of blocks*
 - *Making simple changes*
 - * *Creating and inserting new Instructions*
 - * *Deleting Instructions*
 - * *Replacing an Instruction with another Value*
 - *Replacing individual instructions*
 - *Deleting Instructions*
 - *Replacing multiple uses of Users and Values*
 - * *Deleting GlobalVariables*
 - *How to Create Types*
- *Threads and LLVM*
 - *Ending Execution with `llvm_shutdown()`*
 - *Lazy Initialization with `ManagedStatic`*
 - *Achieving Isolation with `LLVMContext`*
 - *Threads and the JIT*
- *Advanced Topics*
 - *The `ValueSymbolTable` class*
 - *The User and owned Use classes' memory layout*

- * *Interaction and relationship between User and Use objects*
- * *The waymarking algorithm*
- * *Reference implementation*
- * *Tagging considerations*
- *Designing Type Hierarchies and Polymorphic Interfaces*
- *ABI Breaking Checks*
- *The Core LLVM Class Hierarchy Reference*
 - *The Type class and Derived Types*
 - * *Important Public Methods*
 - * *Important Derived Types*
 - *The Module class*
 - * *Important Public Members of the Module class*
 - *The Value class*
 - * *Important Public Members of the Value class*
 - *The User class*
 - * *Important Public Members of the User class*
 - *The Instruction class*
 - * *Important Subclasses of the Instruction class*
 - * *Important Public Members of the Instruction class*
 - *The Constant class and subclasses*
 - * *Important Subclasses of Constant*
 - *The GlobalValue class*
 - * *Important Public Members of the GlobalValue class*
 - *The Function class*
 - * *Important Public Members of the Function*
 - *The GlobalVariable class*
 - * *Important Public Members of the GlobalVariable class*
 - *The BasicBlock class*
 - * *Important Public Members of the BasicBlock class*
 - *The Argument class*

Warning: This is always a work in progress.

3.7.1 Introduction

This document is meant to highlight some of the important classes and interfaces available in the LLVM source-base. This manual is not intended to explain what LLVM is, how it works, and what LLVM code looks like. It assumes that you know the basics of LLVM and are interested in writing transformations or otherwise analyzing or manipulating the code.

This document should get you oriented so that you can find your way in the continuously growing source code that makes up the LLVM infrastructure. Note that this manual is not intended to serve as a replacement for reading the source code, so if you think there should be a method in one of these classes to do something, but it's not listed, check the source. Links to the [doxygen](#) sources are provided to make this as easy as possible.

The first section of this document describes general information that is useful to know when working in the LLVM infrastructure, and the second describes the Core LLVM classes. In the future this manual will be extended with information describing how to use extension libraries, such as dominator information, CFG traversal routines, and useful utilities like the `InstVisitor` ([doxygen](#)) template.

3.7.2 General Information

This section contains general information that is useful if you are working in the LLVM source-base, but that isn't specific to any particular API.

The C++ Standard Template Library

LLVM makes heavy use of the C++ Standard Template Library (STL), perhaps much more than you are used to, or have seen before. Because of this, you might want to do a little background reading in the techniques used and capabilities of the library. There are many good pages that discuss the STL, and several books on the subject that you can get, so it will not be discussed in this document.

Here are some useful links:

1. [cppreference.com](#) - an excellent reference for the STL and other parts of the standard C++ library.
2. [C++ In a Nutshell](#) - This is an O'Reilly book in the making. It has a decent Standard Library Reference that rivals Dinkumware's, and is unfortunately no longer free since the book has been published.
3. [C++ Frequently Asked Questions](#).
4. [SGI's STL Programmer's Guide](#) - Contains a useful [Introduction to the STL](#).
5. [Bjarne Stroustrup's C++ Page](#).
6. [Bruce Eckel's Thinking in C++, 2nd ed. Volume 2 Revision 4.0](#) (even better, [get the book](#)).

You are also encouraged to take a look at the [LLVM Coding Standards](#) guide which focuses on how to write maintainable code more than where to put your curly braces.

Other useful references

1. [Using static and shared libraries across platforms](#)

3.7.3 Important and useful LLVM APIs

Here we highlight some LLVM APIs that are generally useful and good to know about when writing transformations.

The `isa<>`, `cast<>` and `dyn_cast<>` templates

The LLVM source-base makes extensive use of a custom form of RTTI. These templates have many similarities to the C++ `dynamic_cast<>` operator, but they don't have some drawbacks (primarily stemming from the fact that `dynamic_cast<>` only works on classes that have a v-table). Because they are used so often, you must know what they do and how they work. All of these templates are defined in the `llvm/Support/Casting.h` ([doxygen](#)) file (note that you very rarely have to include this file directly).

`isa<>`: The `isa<>` operator works exactly like the Java “instanceof” operator. It returns true or false depending on whether a reference or pointer points to an instance of the specified class. This can be very useful for constraint checking of various sorts (example below).

`cast<>`: The `cast<>` operator is a “checked cast” operation. It converts a pointer or reference from a base class to a derived class, causing an assertion failure if it is not really an instance of the right type. This should be used in cases where you have some information that makes you believe that something is of the right type. An example of the `isa<>` and `cast<>` template is:

```
static bool isLoopInvariant(const Value *V, const Loop *L) {
    if (isa<Constant>(V) || isa<Argument>(V) || isa<GlobalValue>(V))
        return true;

    // Otherwise, it must be an instruction...
    return !L->contains(cast<Instruction>(V)->getParent());
}
```

Note that you should **not** use an `isa<>` test followed by a `cast<>`, for that use the `dyn_cast<>` operator.

`dyn_cast<>`: The `dyn_cast<>` operator is a “checking cast” operation. It checks to see if the operand is of the specified type, and if so, returns a pointer to it (this operator does not work with references). If the operand is not of the correct type, a null pointer is returned. Thus, this works very much like the `dynamic_cast<>` operator in C++, and should be used in the same circumstances. Typically, the `dyn_cast<>` operator is used in an `if` statement or some other flow control statement like this:

```
if (auto *AI = dyn_cast<AllocationInst>(Val)) {
    // ...
}
```

This form of the `if` statement effectively combines together a call to `isa<>` and a call to `cast<>` into one statement, which is very convenient.

Note that the `dyn_cast<>` operator, like C++’s `dynamic_cast<>` or Java’s `instanceof` operator, can be abused. In particular, you should not use big chained `if/then/else` blocks to check for lots of different variants of classes. If you find yourself wanting to do this, it is much cleaner and more efficient to use the `InstVisitor` class to dispatch over the instruction type directly.

`cast_or_null<>`: The `cast_or_null<>` operator works just like the `cast<>` operator, except that it allows for a null pointer as an argument (which it then propagates). This can sometimes be useful, allowing you to combine several null checks into one.

`dyn_cast_or_null<>`: The `dyn_cast_or_null<>` operator works just like the `dyn_cast<>` operator, except that it allows for a null pointer as an argument (which it then propagates). This can sometimes be useful, allowing you to combine several null checks into one.

These five templates can be used with any classes, whether they have a v-table or not. If you want to add support for these templates, see the document [How to set up LLVM-style RTTI for your class hierarchy](#)

Passing strings (the `StringRef` and `Twine` classes)

Although LLVM generally does not do much string manipulation, we do have several important APIs which take strings. Two important examples are the `Value` class – which has names for instructions, functions, etc. – and the `StringMap` class which is used extensively in LLVM and Clang.

These are generic classes, and they need to be able to accept strings which may have embedded null characters. Therefore, they cannot simply take a `const char *`, and taking a `const std::string&` requires clients to perform a heap allocation which is usually unnecessary. Instead, many LLVM APIs use a `StringRef` or a `const Twine&` for passing strings efficiently.

The `StringRef` class

The `StringRef` data type represents a reference to a constant string (a character array and a length) and supports the common operations available on `std::string`, but does not require heap allocation.

It can be implicitly constructed using a C style null-terminated string, an `std::string`, or explicitly with a character pointer and length. For example, the `StringRef` `find` function is declared as:

```
iterator find(StringRef Key);
```

and clients can call it using any one of:

```
Map.find("foo");           // Lookup "foo"
Map.find(std::string("bar")); // Lookup "bar"
Map.find(StringRef("\0baz", 4)); // Lookup "\0baz"
```

Similarly, APIs which need to return a string may return a `StringRef` instance, which can be used directly or converted to an `std::string` using the `str` member function. See `llvm/ADT/StringRef.h` ([doxygen](#)) for more information.

You should rarely use the `StringRef` class directly, because it contains pointers to external memory it is not generally safe to store an instance of the class (unless you know that the external storage will not be freed). `StringRef` is small and pervasive enough in LLVM that it should always be passed by value.

The `Twine` class

The `Twine` ([doxygen](#)) class is an efficient way for APIs to accept concatenated strings. For example, a common LLVM paradigm is to name one instruction based on the name of another instruction with a suffix, for example:

```
New = CmpInst::Create(..., SO->getName() + ".cmp");
```

The `Twine` class is effectively a lightweight [rope](#) which points to temporary (stack allocated) objects. Twines can be implicitly constructed as the result of the plus operator applied to strings (i.e., a C strings, an `std::string`, or a `StringRef`). The `twine` delays the actual concatenation of strings until it is actually required, at which point it can be efficiently rendered directly into a character array. This avoids unnecessary heap allocation involved in constructing the temporary results of string concatenation. See `llvm/ADT/Twine.h` ([doxygen](#)) and [here](#) for more information.

As with a `StringRef`, `Twine` objects point to external memory and should almost never be stored or mentioned directly. They are intended solely for use when defining a function which should be able to efficiently accept concatenated strings.

Formatting strings (the `formatv` function)

While LLVM doesn't necessarily do a lot of string manipulation and parsing, it does do a lot of string formatting. From diagnostic messages, to `llvm` tool outputs such as `llvm-readobj` to printing verbose disassembly listings and LLDB runtime logging, the need for string formatting is pervasive.

The `formatv` is similar in spirit to `printf`, but uses a different syntax which borrows heavily from Python and C#. Unlike `printf` it deduces the type to be formatted at compile time, so it does not need a format specifier such as `%d`. This reduces the mental overhead of trying to construct portable format strings, especially for platform-specific types like `size_t` or pointer types. Unlike both `printf` and Python, it additionally fails to compile if LLVM does not know how to format the type. These two properties ensure that the function is both safer and simpler to use than traditional formatting methods such as the `printf` family of functions.

Simple formatting

A call to `formatv` involves a single **format string** consisting of 0 or more **replacement sequences**, followed by a variable length list of **replacement values**. A replacement sequence is a string of the form `{N[, align]:style}}`.

`N` refers to the 0-based index of the argument from the list of replacement values. Note that this means it is possible to reference the same parameter multiple times, possibly with different style and/or alignment options, in any order.

`align` is an optional string specifying the width of the field to format the value into, and the alignment of the value within the field. It is specified as an optional **alignment style** followed by a positive integral **field width**. The alignment style can be one of the characters `-` (left align), `=` (center align), or `+` (right align). The default is right aligned.

`style` is an optional string consisting of a type specific that controls the formatting of the value. For example, to format a floating point value as a percentage, you can use the style option `P`.

Custom formatting

There are two ways to customize the formatting behavior for a type.

1. Provide a template specialization of `llvm::format_provider<T>` for your type `T` with the appropriate static format method.

```
namespace llvm {
  template<>
  struct format_provider<MyFooBar> {
    static void format(const MyFooBar &V, raw_ostream &Stream,StringRef_
↪Style) {
        // Do whatever is necessary to format `V` into `Stream`
    }
  };
  void foo() {
    MyFooBar X;
    std::string S = formatv("{0}", X);
  }
}
```

This is a useful extensibility mechanism for adding support for formatting your own custom types with your own custom Style options. But it does not help when you want to extend the mechanism for formatting a type that the library already knows how to format. For that, we need something else.

2. Provide a **format adapter** inheriting from `llvm::FormatAdapter<T>`.

```

namespace anything {
    struct format_int_custom : public llvm::FormatAdapter<int> {
        explicit format_int_custom(int N) : llvm::FormatAdapter<int>(N) {}
        void format(llvm::raw_ostream &Stream,StringRef Style) override {
            // Do whatever is necessary to format ``this->Item`` into ``Stream``
        }
    };
}

namespace llvm {
    void foo() {
        std::string S = formatv("{0}", anything::format_int_custom(42));
    }
}

```

If the type is detected to be derived from `FormatAdapter<T>`, `formatv` will call the `format` method on the argument passing in the specified style. This allows one to provide custom formatting of any type, including one which already has a builtin format provider.

formatv Examples

Below is intended to provide an incomplete set of examples demonstrating the usage of `formatv`. More information can be found by reading the doxygen documentation or by looking at the unit test suite.

```

std::string S;
// Simple formatting of basic types and implicit string conversion.
S = formatv("{0} ({1:P})", 7, 0.35); // S == "7 (35.00%)"

// Out-of-order referencing and multi-referencing
outs() << formatv("{0} {2} {1} {0}", 1, "test", 3); // prints "1 3 test 1"

// Left, right, and center alignment
S = formatv("{0,7}", 'a'); // S == "      a";
S = formatv("{0,-7}", 'a'); // S == "a      ";
S = formatv("{0,=7}", 'a'); // S == " a      ";
S = formatv("{0,+7}", 'a'); // S == "      a";

// Custom styles
S = formatv("{0:N} - {0:x} - {1:E}", 12345, 123908342); // S == "12,345 - 0x3039 - 1.
↪24E8"

// Adapters
S = formatv("{0}", fmt_align(42, AlignStyle::Center, 7)); // S == "  42  "
S = formatv("{0}", fmt_repeat("hi", 3)); // S == "hihihi"
S = formatv("{0}", fmt_pad("hi", 2, 6)); // S == "  hi      "

// Ranges
std::vector<int> V = {8, 9, 10};
S = formatv("{0}", make_range(V.begin(), V.end())); // S == "8, 9, 10"
S = formatv("{0:$[+]}", make_range(V.begin(), V.end())); // S == "8+9+10"
S = formatv("{0:$[ + ]@[x]}", make_range(V.begin(), V.end())); // S == "0x8 + 0x9 +
↪0xA"

```

Error handling

Proper error handling helps us identify bugs in our code, and helps end-users understand errors in their tool usage. Errors fall into two broad categories: *programmatic* and *recoverable*, with different strategies for handling and reporting.

Programmatic Errors

Programmatic errors are violations of program invariants or API contracts, and represent bugs within the program itself. Our aim is to document invariants, and to abort quickly at the point of failure (providing some basic diagnostic) when invariants are broken at runtime.

The fundamental tools for handling programmatic errors are assertions and the `llvm_unreachable` function. Assertions are used to express invariant conditions, and should include a message describing the invariant:

```
assert(isPhysReg(R) && "All virt regs should have been allocated already.");
```

The `llvm_unreachable` function can be used to document areas of control flow that should never be entered if the program invariants hold:

```
enum { Foo, Bar, Baz } X = foo();

switch (X) {
  case Foo: /* Handle Foo */; break;
  case Bar: /* Handle Bar */; break;
  default:
    llvm_unreachable("X should be Foo or Bar here");
}
```

Recoverable Errors

Recoverable errors represent an error in the program's environment, for example a resource failure (a missing file, a dropped network connection, etc.), or malformed input. These errors should be detected and communicated to a level of the program where they can be handled appropriately. Handling the error may be as simple as reporting the issue to the user, or it may involve attempts at recovery.

Recoverable errors are modeled using LLVM's `Error` scheme. This scheme represents errors using function return values, similar to classic C integer error codes, or C++'s `std::error_code`. However, the `Error` class is actually a lightweight wrapper for user-defined error types, allowing arbitrary information to be attached to describe the error. This is similar to the way C++ exceptions allow throwing of user-defined types.

Success values are created by calling `Error::success()`, E.g.:

```
Error foo() {
  // Do something.
  // Return success.
  return Error::success();
}
```

Success values are very cheap to construct and return - they have minimal impact on program performance.

Failure values are constructed using `make_error<T>`, where `T` is any class that inherits from the `ErrorInfo` utility, E.g.:

```
class BadFileFormat : public ErrorInfo<BadFileFormat> {
public:
```

```

static char ID;
std::string Path;

BadFileFormat(StringRef Path) : Path(Path.str()) {}

void log(raw_ostream &OS) const override {
    OS << Path << " is malformed";
}

std::error_code convertToErrorCode() const override {
    return make_error_code(object_error::parse_failed);
}
};

char BadFileFormat::ID; // This should be declared in the C++ file.

Error printFormattedFile(StringRef Path) {
    if (<check for valid format>)
        return make_error<InvalidObjectFile>(Path);
    // print file contents.
    return Error::success();
}

```

Error values can be implicitly converted to bool: true for error, false for success, enabling the following idiom:

```

Error mayFail();

Error foo() {
    if (auto Err = mayFail())
        return Err;
    // Success! We can proceed.
    ...
}

```

For functions that can fail but need to return a value the `Expected<T>` utility can be used. Values of this type can be constructed with either a `T`, or an `Error`. `Expected<T>` values are also implicitly convertible to boolean, but with the opposite convention to `Error`: true for success, false for error. If success, the `T` value can be accessed via the dereference operator. If failure, the `Error` value can be extracted using the `takeError()` method. Idiomatic usage looks like:

```

Expected<FormattedFile> openFormattedFile(StringRef Path) {
    // If badly formatted, return an error.
    if (auto Err = checkFormat(Path))
        return std::move(Err);
    // Otherwise return a FormattedFile instance.
    return FormattedFile(Path);
}

Error processFormattedFile(StringRef Path) {
    // Try to open a formatted file
    if (auto FileOrErr = openFormattedFile(Path)) {
        // On success, grab a reference to the file and continue.
        auto &File = *FileOrErr;
        ...
    } else
        // On error, extract the Error value and return it.
        return FileOrErr.takeError();
}

```

If an `Expected<T>` value is in success mode then the `takeError()` method will return a success value. Using this fact, the above function can be rewritten as:

```
Error processFormattedFile(StringRef Path) {  
    // Try to open a formatted file  
    auto FileOrErr = openFormattedFile(Path);  
    if (auto Err = FileOrErr.takeError())  
        // On error, extract the Error value and return it.  
        return Err;  
    // On success, grab a reference to the file and continue.  
    auto &File = *FileOrErr;  
    ...  
}
```

This second form is often more readable for functions that involve multiple `Expected<T>` values as it limits the indentation required.

All `Error` instances, whether success or failure, must be either checked or moved from (via `std::move` or a `return`) before they are destructed. Accidentally discarding an unchecked error will cause a program abort at the point where the unchecked value's destructor is run, making it easy to identify and fix violations of this rule.

Success values are considered checked once they have been tested (by invoking the boolean conversion operator):

```
if (auto Err = mayFail(...))  
    return Err; // Failure value - move error to caller.  
  
// Safe to continue: Err was checked.
```

In contrast, the following code will always cause an abort, even if `mayFail` returns a success value:

```
mayFail();  
// Program will always abort here, even if mayFail() returns Success, since  
// the value is not checked.
```

Failure values are considered checked once a handler for the error type has been activated:

```
handleErrors(  
    processFormattedFile(...),  
    [] (const BadFileFormat &BFF) {  
        report("Unable to process " + BFF.Path + ": bad format");  
    },  
    [] (const FileNotFound &FNF) {  
        report("File not found " + FNF.Path);  
    });
```

The `handleErrors` function takes an error as its first argument, followed by a variadic list of “handlers”, each of which must be a callable type (a function, lambda, or class with a call operator) with one argument. The `handleErrors` function will visit each handler in the sequence and check its argument type against the dynamic type of the error, running the first handler that matches. This is the same decision process that is used to decide which catch clause to run for a C++ exception.

Since the list of handlers passed to `handleErrors` may not cover every error type that can occur, the `handleErrors` function also returns an `Error` value that must be checked or propagated. If the error value that is passed to `handleErrors` does not match any of the handlers it will be returned from `handleErrors`. Idiomatic use of `handleErrors` thus looks like:

```
if (auto Err =  
    handleErrors(  
        processFormattedFile(...),
```

```

    [] (const BadFileFormat &BFF) {
        report("Unable to process " + BFF.Path + ": bad format");
    },
    [] (const FileNotFound &FNF) {
        report("File not found " + FNF.Path);
    })
    return Err;

```

In cases where you truly know that the handler list is exhaustive the `handleAllErrors` function can be used instead. This is identical to `handleErrors` except that it will terminate the program if an unhandled error is passed in, and can therefore return `void`. The `handleAllErrors` function should generally be avoided: the introduction of a new error type elsewhere in the program can easily turn a formerly exhaustive list of errors into a non-exhaustive list, risking unexpected program termination. Where possible, use `handleErrors` and propagate unknown errors up the stack instead.

For tool code, where errors can be handled by printing an error message then exiting with an error code, the *ExitOnError* utility may be a better choice than `handleErrors`, as it simplifies control flow when calling fallible functions.

In situations where it is known that a particular call to a fallible function will always succeed (for example, a call to a function that can only fail on a subset of inputs with an input that is known to be safe) the *cantFail* functions can be used to remove the error type, simplifying control flow.

StringError

Many kinds of errors have no recovery strategy, the only action that can be taken is to report them to the user so that the user can attempt to fix the environment. In this case representing the error as a string makes perfect sense. LLVM provides the `StringError` class for this purpose. It takes two arguments: A string error message, and an equivalent `std::error_code` for interoperability:

```

make_error<StringError>("Bad executable",
                        make_error_code(errc::executable_format_error));

```

If you're certain that the error you're building will never need to be converted to a `std::error_code` you can use the `inconvertibleErrorCode()` function:

```

make_error<StringError>("Bad executable", inconvertibleErrorCode());

```

This should be done only after careful consideration. If any attempt is made to convert this error to a `std::error_code` it will trigger immediate program termination. Unless you are certain that your errors will not need interoperability you should look for an existing `std::error_code` that you can convert to, and even (as painful as it is) consider introducing a new one as a stopgap measure.

Interoperability with `std::error_code` and `ErrorOr`

Many existing LLVM APIs use `std::error_code` and its partner `ErrorOr<T>` (which plays the same role as `Expected<T>`, but wraps a `std::error_code` rather than an `Error`). The infectious nature of error types means that an attempt to change one of these functions to return `Error` or `Expected<T>` instead often results in an avalanche of changes to callers, callers of callers, and so on. (The first such attempt, returning an `Error` from `MachObjectFile`'s constructor, was abandoned after the diff reached 3000 lines, impacted half a dozen libraries, and was still growing).

To solve this problem, the `Error/std::error_code` interoperability requirement was introduced. Two pairs of functions allow any `Error` value to be converted to a `std::error_code`, any `Expected<T>` to be converted to an `ErrorOr<T>`, and vice versa:

```
std::error_code errorToErrorCode(Error Err);
Error errorCodeToError(std::error_code EC);

template <typename T> ErrorOr<T> expectedToErrorOr(Expected<T> TOrErr);
template <typename T> Expected<T> errorOrToExpected(ErrorOr<T> TOrEC);
```

Using these APIs it is easy to make surgical patches that update individual functions from `std::error_code` to `Error`, and from `ErrorOr<T>` to `Expected<T>`.

Returning Errors from error handlers

Error recovery attempts may themselves fail. For that reason, `handleErrors` actually recognises three different forms of handler signature:

```
// Error must be handled, no new errors produced:
void(UserDefinedError &E);

// Error must be handled, new errors can be produced:
Error(UserDefinedError &E);

// Original error can be inspected, then re-wrapped and returned (or a new
// error can be produced):
Error(std::unique_ptr<UserDefinedError> E);
```

Any error returned from a handler will be returned from the `handleErrors` function so that it can be handled itself, or propagated up the stack.

Using `ExitOnError` to simplify tool code

Library code should never call `exit` for a recoverable error, however in tool code (especially command line tools) this can be a reasonable approach. Calling `exit` upon encountering an error dramatically simplifies control flow as the error no longer needs to be propagated up the stack. This allows code to be written in straight-line style, as long as each fallible call is wrapped in a check and call to `exit`. The `ExitOnError` class supports this pattern by providing call operators that inspect `Error` values, stripping the error away in the success case and logging to `stderr` then exiting in the failure case.

To use this class, declare a global `ExitOnError` variable in your program:

```
ExitOnError ExitOnErr;
```

Calls to fallible functions can then be wrapped with a call to `ExitOnErr`, turning them into non-failing calls:

```
Error mayFail();
Expected<int> mayFail2();

void foo() {
    ExitOnErr(mayFail());
    int X = ExitOnErr(mayFail2());
}
```

On failure, the error's log message will be written to `stderr`, optionally preceded by a string “banner” that can be set by calling the `setBanner` method. A mapping can also be supplied from `Error` values to exit codes using the `setExitCodeMapper` method:


```
int main(int argc, char *argv[]) {
    ExitOnError.setBanner(std::string(argv[0]) + " error:");
    ExitOnError.setExitCodeMapper(
        [] (const Error &Err) {
            if (Err.isA<BadFileFormat>())
                return 2;
            return 1;
        });
}
```

Use `ExitOnError` in your tool code where possible as it can greatly improve readability.

Using `cantFail` to simplify safe callsites

Some functions may only fail for a subset of their inputs, so calls using known safe inputs can be assumed to succeed.

The `cantFail` functions encapsulate this by wrapping an assertion that their argument is a success value and, in the case of `Expected<T>`, unwrapping the `T` value:

```
Error onlyFailsForSomeXValues(int X);
Expected<int> onlyFailsForSomeXValues2(int X);

void foo() {
    cantFail(onlyFailsForSomeXValues(KnownSafeValue));
    int Y = cantFail(onlyFailsForSomeXValues2(KnownSafeValue));
    ...
}
```

Like the `ExitOnError` utility, `cantFail` simplifies control flow. Their treatment of error cases is very different however: Where `ExitOnError` is guaranteed to terminate the program on an error input, `cantFail` simply asserts that the result is success. In debug builds this will result in an assertion failure if an error is encountered. In release builds the behavior of `cantFail` for failure values is undefined. As such, care must be taken in the use of `cantFail`: clients must be certain that a `cantFail` wrapped call really can not fail with the given arguments.

Use of the `cantFail` functions should be rare in library code, but they are likely to be of more use in tool and unit-test code where inputs and/or mocked-up classes or functions may be known to be safe.

Fallible constructors

Some classes require resource acquisition or other complex initialization that can fail during construction. Unfortunately constructors can't return errors, and having clients test objects after they're constructed to ensure that they're valid is error prone as it's all too easy to forget the test. To work around this, use the named constructor idiom and return an `Expected<T>`:

```
class Foo {
public:

    static Expected<Foo> Create(Resource R1, Resource R2) {
        Error Err;
        Foo F(R1, R2, Err);
        if (Err)
            return std::move(Err);
        return std::move(F);
    }

private:
}
```

```
Foo(Resource R1, Resource R2, Error &Err) {
    ErrorAsOutParameter EAO(&Err);
    if (auto Err2 = R1.acquire()) {
        Err = std::move(Err2);
        return;
    }
    Err = R2.acquire();
}
};
```

Here, the named constructor passes an `Error` by reference into the actual constructor, which the constructor can then use to return errors. The `ErrorAsOutParameter` utility sets the `Error` value's checked flag on entry to the constructor so that the error can be assigned to, then resets it on exit to force the client (the named constructor) to check the error.

By using this idiom, clients attempting to construct a `Foo` receive either a well-formed `Foo` or an `Error`, never an object in an invalid state.

Propagating and consuming errors based on types

In some contexts, certain types of error are known to be benign. For example, when walking an archive, some clients may be happy to skip over badly formatted object files rather than terminating the walk immediately. Skipping badly formatted objects could be achieved using an elaborate handler method, but the `Error.h` header provides two utilities that make this idiom much cleaner: the type inspection method, `isA`, and the `consumeError` function:

```
Error walkArchive(Archive A) {
    for (unsigned I = 0; I != A.numMembers(); ++I) {
        auto ChildOrErr = A.getMember(I);
        if (auto Err = ChildOrErr.takeError()) {
            if (Err.isA<BadFileFormat>())
                consumeError(std::move(Err))
            else
                return Err;
        }
        auto &Child = *ChildOrErr;
        // Use Child
        ...
    }
    return Error::success();
}
```

Concatenating Errors with joinErrors

In the archive walking example above `BadFileFormat` errors are simply consumed and ignored. If the client had wanted report these errors after completing the walk over the archive they could use the `joinErrors` utility:

```
Error walkArchive(Archive A) {
    Error DeferredErrs = Error::success();
    for (unsigned I = 0; I != A.numMembers(); ++I) {
        auto ChildOrErr = A.getMember(I);
        if (auto Err = ChildOrErr.takeError())
            if (Err.isA<BadFileFormat>())
                DeferredErrs = joinErrors(std::move(DeferredErrs), std::move(Err));
    }
}
```

```

        else
            return Err;
        auto &Child = *ChildOrErr;
        // Use Child
        ...
    }
    return DeferredErrs;
}

```

The `joinErrors` routine builds a special error type called `ErrorList`, which holds a list of user defined errors. The `handleErrors` routine recognizes this type and will attempt to handle each of the contained errors in order. If all contained errors can be handled, `handleErrors` will return `Error::success()`, otherwise `handleErrors` will concatenate the remaining errors and return the resulting `ErrorList`.

Building fallible iterators and iterator ranges

The archive walking examples above retrieve archive members by index, however this requires considerable boilerplate for iteration and error checking. We can clean this up by using `Error` with the “fallible iterator” pattern. The usual C++ iterator patterns do not allow for failure on increment, but we can incorporate support for it by having iterators hold an `Error` reference through which they can report failure. In this pattern, if an increment operation fails the failure is recorded via the `Error` reference and the iterator value is set to the end of the range in order to terminate the loop. This ensures that the dereference operation is safe anywhere that an ordinary iterator dereference would be safe (i.e. when the iterator is not equal to end). Where this pattern is followed (as in the `llvm::object::Archive` class) the result is much cleaner iteration idiom:

```

Error Err;
for (auto &Child : Ar->children(Err)) {
    // Use Child - we only enter the loop when it's valid
    ...
}
// Check Err after the loop to ensure it didn't break due to an error.
if (Err)
    return Err;

```

More information on `Error` and its related utilities can be found in the `Error.h` header file.

Passing functions and other callable objects

Sometimes you may want a function to be passed a callback object. In order to support lambda expressions and other function objects, you should not use the traditional C approach of taking a function pointer and an opaque cookie:

```
void takeCallback(bool (*Callback)(Function *, void *), void *Cookie);
```

Instead, use one of the following approaches:

Function template

If you don’t mind putting the definition of your function into a header file, make it a function template that is templated on the callable type.

```

template<typename Callable>
void takeCallback(Callable Callback) {

```

```
    Callback(1, 2, 3);  
}
```

The `function_ref` class template

The `function_ref` (doxygen) class template represents a reference to a callable object, templated over the type of the callable. This is a good choice for passing a callback to a function, if you don't need to hold onto the callback after the function returns. In this way, `function_ref` is to `std::function` as `StringRef` is to `std::string`.

`function_ref<Ret (Param1, Param2, ...)>` can be implicitly constructed from any callable object that can be called with arguments of type `Param1, Param2, ...`, and returns a value that can be converted to type `Ret`. For example:

```
void visitBasicBlocks(Function *F, function_ref<bool (BasicBlock*)> Callback) {  
    for (BasicBlock &BB : *F)  
        if (Callback(&BB))  
            return;  
}
```

can be called using:

```
visitBasicBlocks(F, [&](BasicBlock *BB) {  
    if (process(BB))  
        return isEmpty(BB);  
    return false;  
});
```

Note that a `function_ref` object contains pointers to external memory, so it is not generally safe to store an instance of the class (unless you know that the external storage will not be freed). If you need this ability, consider using `std::function`. `function_ref` is small enough that it should always be passed by value.

The `DEBUG()` macro and `-debug` option

Often when working on your pass you will put a bunch of debugging printouts and other code into your pass. After you get it working, you want to remove it, but you may need it again in the future (to work out new bugs that you run across).

Naturally, because of this, you don't want to delete the debug printouts, but you don't want them to always be noisy. A standard compromise is to comment them out, allowing you to enable them if you need them in the future.

The `llvm/Support/Debug.h` (doxygen) file provides a macro named `DEBUG()` that is a much nicer solution to this problem. Basically, you can put arbitrary code into the argument of the `DEBUG` macro, and it is only executed if 'opt' (or any other tool) is run with the `'-debug'` command line argument:

```
DEBUG(errs() << "I am here!\n");
```

Then you can run your pass like this:

```
$ opt < a.bc > /dev/null -mypass  
<no output>  
$ opt < a.bc > /dev/null -mypass -debug  
I am here!
```

Using the `DEBUG()` macro instead of a home-brewed solution allows you to not have to create “yet another” command line option for the debug output for your pass. Note that `DEBUG()` macros are disabled for non-asserts builds, so they do not cause a performance impact at all (for the same reason, they should also not contain side-effects!).

One additional nice thing about the `DEBUG()` macro is that you can enable or disable it directly in `gdb`. Just use “set `DebugFlag=0`” or “set `DebugFlag=1`” from the `gdb` if the program is running. If the program hasn’t been started yet, you can always just run it with `-debug`.

Fine grained debug info with `DEBUG_TYPE` and the `-debug-only` option

Sometimes you may find yourself in a situation where enabling `-debug` just turns on **too much** information (such as when working on the code generator). If you want to enable debug information with more fine-grained control, you should define the `DEBUG_TYPE` macro and use the `-debug-only` option as follows:

```
#define DEBUG_TYPE "foo"
DEBUG(errs() << "'foo' debug type\n");
#undef DEBUG_TYPE
#define DEBUG_TYPE "bar"
DEBUG(errs() << "'bar' debug type\n"));
#undef DEBUG_TYPE
```

Then you can run your pass like this:

```
$ opt < a.bc > /dev/null -mypass
<no output>
$ opt < a.bc > /dev/null -mypass -debug
'foo' debug type
'bar' debug type
$ opt < a.bc > /dev/null -mypass -debug-only=foo
'foo' debug type
$ opt < a.bc > /dev/null -mypass -debug-only=bar
'bar' debug type
$ opt < a.bc > /dev/null -mypass -debug-only=foo,bar
'foo' debug type
'bar' debug type
```

Of course, in practice, you should only set `DEBUG_TYPE` at the top of a file, to specify the debug type for the entire module. Be careful that you only do this after including `Debug.h` and not around any `#include` of headers. Also, you should use names more meaningful than “foo” and “bar”, because there is no system in place to ensure that names do not conflict. If two different modules use the same string, they will all be turned on when the name is specified. This allows, for example, all debug information for instruction scheduling to be enabled with `-debug-only=InstrSched`, even if the source lives in multiple files. The name must not include a comma (,) as that is used to separate the arguments of the `-debug-only` option.

For performance reasons, `-debug-only` is not available in optimized build (`--enable-optimized`) of LLVM.

The `DEBUG_WITH_TYPE` macro is also available for situations where you would like to set `DEBUG_TYPE`, but only for one specific `DEBUG` statement. It takes an additional first parameter, which is the type to use. For example, the preceding example could be written as:

```
DEBUG_WITH_TYPE("foo", errs() << "'foo' debug type\n");
DEBUG_WITH_TYPE("bar", errs() << "'bar' debug type\n");
```

The `Statistic` class & `-stats` option

The `llvm/ADT/Statistic.h` (doxygen) file provides a class named `Statistic` that is used as a unified way to keep track of what the LLVM compiler is doing and how effective various optimizations are. It is useful to see what optimizations are contributing to making a particular program run faster.

Often you may run your pass on some big program, and you're interested to see how many times it makes a certain transformation. Although you can do this with hand inspection, or some ad-hoc method, this is a real pain and not very useful for big programs. Using the `Statistic` class makes it very easy to keep track of this information, and the calculated information is presented in a uniform manner with the rest of the passes being executed.

There are many examples of `Statistic` uses, but the basics of using it are as follows:

Define your statistic like this:

```
#define DEBUG_TYPE "mypassname" // This goes before any #includes.
STATISTIC(NumXForms, "The # of times I did stuff");
```

The `STATISTIC` macro defines a static variable, whose name is specified by the first argument. The pass name is taken from the `DEBUG_TYPE` macro, and the description is taken from the second argument. The variable defined (“NumXForms” in this case) acts like an unsigned integer.

Whenever you make a transformation, bump the counter:

```
++NumXForms; // I did stuff!
```

That's all you have to do. To get ‘opt’ to print out the statistics gathered, use the ‘`-stats`’ option:

```
$ opt -stats -mypassname < program.bc > /dev/null
... statistics output ...
```

Note that in order to use the ‘`-stats`’ option, LLVM must be compiled with assertions enabled.

When running `opt` on a C file from the SPEC benchmark suite, it gives a report that looks like this:

```
7646 bitcodewriter - Number of normal instructions
725 bitcodewriter - Number of oversized instructions
129996 bitcodewriter - Number of bitcode bytes written
2817 raise - Number of insts DCEd or constprop'd
3213 raise - Number of cast-of-self removed
5046 raise - Number of expression trees converted
75 raise - Number of other getelementptr's formed
138 raise - Number of load/store peepholes
42 deadtypeelim - Number of unused typenames removed from symtab
392 funcresolve - Number of varargs functions resolved
27 globaldce - Number of global variables removed
2 adce - Number of basic blocks removed
134 cee - Number of branches revectorized
49 cee - Number of setcc instruction eliminated
532 gcse - Number of loads removed
2919 gcse - Number of instructions removed
86 indvars - Number of canonical indvars added
87 indvars - Number of aux indvars removed
25 instcombine - Number of dead inst eliminate
434 instcombine - Number of insts combined
248 licm - Number of load insts hoisted
1298 licm - Number of insts hoisted to a loop pre-header
3 licm - Number of insts hoisted to multiple loop preds (bad, no loop_
↳pre-header)
```

75 mem2reg	- Number of alloca's promoted
1444 cfgsimplify	- Number of blocks simplified

Obviously, with so many optimizations, having a unified framework for this stuff is very nice. Making your pass fit well into the framework makes it more maintainable and useful.

Adding debug counters to aid in debugging your code

Sometimes, when writing new passes, or trying to track down bugs, it is useful to be able to control whether certain things in your pass happen or not. For example, there are times the minimization tooling can only easily give you large testcases. You would like to narrow your bug down to a specific transformation happening or not happening, automatically, using bisection. This is where debug counters help. They provide a framework for making parts of your code only execute a certain number of times.

The `llvm/Support/DebugCounter.h` (doxygen) file provides a class named `DebugCounter` that can be used to create command line counter options that control execution of parts of your code.

Define your `DebugCounter` like this:

```
DEBUG_COUNTER(DeleteAnInstruction, "passname-delete-instruction",
              "Controls which instructions get delete").
```

The `DEBUG_COUNTER` macro defines a static variable, whose name is specified by the first argument. The name of the counter (which is used on the command line) is specified by the second argument, and the description used in the help is specified by the third argument.

Whatever code you want that control, use `DebugCounter::shouldExecute` to control it.

```
if (DebugCounter::shouldExecute(DeleteAnInstruction))
    I->eraseFromParent();
```

That's all you have to do. Now, using `opt`, you can control when this code triggers using the `'--debug-counter'` option. There are two counters provided, `skip` and `count`. `skip` is the number of times to skip execution of the codepath. `count` is the number of times, once we are done skipping, to execute the codepath.

```
$ opt --debug-counter=passname-delete-instruction-skip=1,passname-delete-instruction-
↪count=2 -passname
```

This will skip the above code the first time we hit it, then execute it twice, then skip the rest of the executions.

So if executed on the following code:

```
%1 = add i32 %a, %b
%2 = add i32 %a, %b
%3 = add i32 %a, %b
%4 = add i32 %a, %b
```

It would delete number `%2` and `%3`.

A utility is provided in `utils/bisect-skip-count` to binary search skip and count arguments. It can be used to automatically minimize the skip and count for a debug-counter variable.

Viewing graphs while debugging code

Several of the important data structures in LLVM are graphs: for example CFGs made out of LLVM *BasicBlocks*, CFGs made out of LLVM *MachineBasicBlocks*, and *Instruction Selection DAGs*. In many cases, while debugging various parts of the compiler, it is nice to instantly visualize these graphs.

LLVM provides several callbacks that are available in a debug build to do exactly that. If you call the `Function::viewCFG()` method, for example, the current LLVM tool will pop up a window containing the CFG for the function where each basic block is a node in the graph, and each node contains the instructions in the block. Similarly, there also exists `Function::viewCFGOnly()` (does not include the instructions), the `MachineFunction::viewCFG()` and `MachineFunction::viewCFGOnly()`, and the `SelectionDAG::viewGraph()` methods. Within GDB, for example, you can usually use something like `call DAG.viewGraph()` to pop up a window. Alternatively, you can sprinkle calls to these functions in your code in places you want to debug.

Getting this to work requires a small amount of setup. On Unix systems with X11, install the [graphviz](#) toolkit, and make sure ‘dot’ and ‘gv’ are in your path. If you are running on Mac OS X, download and install the [Mac OS X Graphviz program](#) and add `/Applications/Graphviz.app/Contents/MacOS/` (or wherever you install it) to your path. The programs need not be present when configuring, building or running LLVM and can simply be installed when needed during an active debug session.

`SelectionDAG` has been extended to make it easier to locate *interesting* nodes in large complex graphs. From `gdb`, if you call `DAG.setGraphColor(node, "color")`, then the next call `DAG.viewGraph()` would highlight the node in the specified color (choices of colors can be found at [colors](#).) More complex node attributes can be provided with call `DAG.setGraphAttrs(node, "attributes")` (choices can be found at [Graph attributes](#).) If you want to restart and clear all the current graph attributes, then you can call `DAG.clearGraphAttrs()`.

Note that graph visualization features are compiled out of Release builds to reduce file size. This means that you need a Debug+Asserts or Release+Asserts build to use these features.

3.7.4 Picking the Right Data Structure for a Task

LLVM has a plethora of data structures in the `llvm/ADT/` directory, and we commonly use STL data structures. This section describes the trade-offs you should consider when you pick one.

The first step is a choose your own adventure: do you want a sequential container, a set-like container, or a map-like container? The most important thing when choosing a container is the algorithmic properties of how you plan to access the container. Based on that, you should use:

- a *map-like* container if you need efficient look-up of a value based on another value. Map-like containers also support efficient queries for containment (whether a key is in the map). Map-like containers generally do not support efficient reverse mapping (values to keys). If you need that, use two maps. Some map-like containers also support efficient iteration through the keys in sorted order. Map-like containers are the most expensive sort, only use them if you need one of these capabilities.
- a *set-like* container if you need to put a bunch of stuff into a container that automatically eliminates duplicates. Some set-like containers support efficient iteration through the elements in sorted order. Set-like containers are more expensive than sequential containers.
- a *sequential* container provides the most efficient way to add elements and keeps track of the order they are added to the collection. They permit duplicates and support efficient iteration, but do not support efficient look-up based on a key.
- a *string* container is a specialized sequential container or reference structure that is used for character or byte arrays.
- a *bit* container provides an efficient way to store and perform set operations on sets of numeric id’s, while automatically eliminating duplicates. Bit containers require a maximum of 1 bit for each identifier you want to store.

Once the proper category of container is determined, you can fine tune the memory use, constant factors, and cache behaviors of access by intelligently picking a member of the category. Note that constant factors and cache behavior can be a big deal. If you have a vector that usually only contains a few elements (but could contain many), for example,

it's much better to use *SmallVector* than *vector*. Doing so avoids (relatively) expensive malloc/free calls, which dwarf the cost of adding the elements to the container.

Sequential Containers (`std::vector`, `std::list`, etc)

There are a variety of sequential containers available for you, based on your needs. Pick the first in this section that will do what you want.

llvm/ADT/ArrayRef.h

The `llvm::ArrayRef` class is the preferred class to use in an interface that accepts a sequential list of elements in memory and just reads from them. By taking an `ArrayRef`, the API can be passed a fixed size array, an `std::vector`, an `llvm::SmallVector` and anything else that is contiguous in memory.

Fixed Size Arrays

Fixed size arrays are very simple and very fast. They are good if you know exactly how many elements you have, or you have a (low) upper bound on how many you have.

Heap Allocated Arrays

Heap allocated arrays (`new[] + delete[]`) are also simple. They are good if the number of elements is variable, if you know how many elements you will need before the array is allocated, and if the array is usually large (if not, consider a *SmallVector*). The cost of a heap allocated array is the cost of the new/delete (aka malloc/free). Also note that if you are allocating an array of a type with a constructor, the constructor and destructors will be run for every element in the array (re-sizable vectors only construct those elements actually used).

llvm/ADT/TinyPtrVector.h

`TinyPtrVector<Type>` is a highly specialized collection class that is optimized to avoid allocation in the case when a vector has zero or one elements. It has two major restrictions: 1) it can only hold values of pointer type, and 2) it cannot hold a null pointer.

Since this container is highly specialized, it is rarely used.

llvm/ADT/SmallVector.h

`SmallVector<Type, N>` is a simple class that looks and smells just like `vector<Type>`: it supports efficient iteration, lays out elements in memory order (so you can do pointer arithmetic between elements), supports efficient `push_back/pop_back` operations, supports efficient random access to its elements, etc.

The advantage of `SmallVector` is that it allocates space for some number of elements (N) **in the object itself**. Because of this, if the `SmallVector` is dynamically smaller than N, no malloc is performed. This can be a big win in cases where the malloc/free call is far more expensive than the code that fiddles around with the elements.

This is good for vectors that are “usually small” (e.g. the number of predecessors/successors of a block is usually less than 8). On the other hand, this makes the size of the `SmallVector` itself large, so you don't want to allocate lots of them (doing so will waste a lot of space). As such, `SmallVectors` are most useful when on the stack.

`SmallVector` also provides a nice portable and efficient replacement for `alloca`.

Note: Prefer to use `SmallVectorImpl<T>` as a parameter type.

In APIs that don't care about the "small size" (most?), prefer to use the `SmallVectorImpl<T>` class, which is basically just the "vector header" (and methods) without the elements allocated after it. Note that `SmallVector<T, N>` inherits from `SmallVectorImpl<T>` so the conversion is implicit and costs nothing. E.g.

```
// BAD: Clients cannot pass e.g. SmallVector<Foo, 4>.
hardcodedSmallSize(SmallVector<Foo, 2> &Out);
// GOOD: Clients can pass any SmallVector<Foo, N>.
allowsAnySmallSize(SmallVectorImpl<Foo> &Out);

void someFunc() {
    SmallVector<Foo, 8> Vec;
    hardcodedSmallSize(Vec); // Error.
    allowsAnySmallSize(Vec); // Works.
}
```

Even though it has "Impl" in the name, this is so widely used that it really isn't "private to the implementation" anymore. A name like `SmallVectorHeader` would be more appropriate.

<vector>

`std::vector` is well loved and respected. It is useful when `SmallVector` isn't: when the size of the vector is often large (thus the small optimization will rarely be a benefit) or if you will be allocating many instances of the vector itself (which would waste space for elements that aren't in the container). `vector` is also useful when interfacing with code that expects vectors :).

One worthwhile note about `std::vector`: avoid code like this:

```
for ( ... ) {
    std::vector<foo> V;
    // make use of V.
}
```

Instead, write this as:

```
std::vector<foo> V;
for ( ... ) {
    // make use of V.
    V.clear();
}
```

Doing so will save (at least) one heap allocation and free per iteration of the loop.

<deque>

`std::deque` is, in some senses, a generalized version of `std::vector`. Like `std::vector`, it provides constant time random access and other similar properties, but it also provides efficient access to the front of the list. It does not guarantee continuity of elements within memory.

In exchange for this extra flexibility, `std::deque` has significantly higher constant factor costs than `std::vector`. If possible, use `std::vector` or something cheaper.

<list>

`std::list` is an extremely inefficient class that is rarely useful. It performs a heap allocation for every element inserted into it, thus having an extremely high constant factor, particularly for small data types. `std::list` also only supports bidirectional iteration, not random access iteration.

In exchange for this high cost, `std::list` supports efficient access to both ends of the list (like `std::deque`, but unlike `std::vector` or `SmallVector`). In addition, the iterator invalidation characteristics of `std::list` are stronger than that of a vector class: inserting or removing an element into the list does not invalidate iterator or pointers to other elements in the list.

llvm/ADT/ilist.h

`ilist<T>` implements an ‘intrusive’ doubly-linked list. It is intrusive, because it requires the element to store and provide access to the prev/next pointers for the list.

`ilist` has the same drawbacks as `std::list`, and additionally requires an `ilist_traits` implementation for the element type, but it provides some novel characteristics. In particular, it can efficiently store polymorphic objects, the traits class is informed when an element is inserted or removed from the list, and `ilists` are guaranteed to support a constant-time splice operation.

These properties are exactly what we want for things like Instructions and basic blocks, which is why these are implemented with `ilists`.

Related classes of interest are explained in the following subsections:

- [*ilist_traits*](#)
- [*iplist*](#)
- [*llvm/ADT/ilist_node.h*](#)
- [*Sentinels*](#)

llvm/ADT/PackedVector.h

Useful for storing a vector of values using only a few number of bits for each value. Apart from the standard operations of a vector-like container, it can also perform an ‘or’ set operation.

For example:

```
enum State {
    None = 0x0,
    FirstCondition = 0x1,
    SecondCondition = 0x2,
    Both = 0x3
};

State get() {
    PackedVector<State, 2> Vec1;
    Vec1.push_back(FirstCondition);

    PackedVector<State, 2> Vec2;
    Vec2.push_back(SecondCondition);

    Vec1 |= Vec2;
    return Vec1[0]; // returns 'Both'.
}
```

`ilist_traits`

`ilist_traits<T>` is `ilist<T>`'s customization mechanism. `iplist<T>` (and consequently `ilist<T>`) publicly derive from this traits class.

`iplist`

`iplist<T>` is `ilist<T>`'s base and as such supports a slightly narrower interface. Notably, inserters from `T&` are absent.

`ilist_traits<T>` is a public base of this class and can be used for a wide variety of customizations.

`llvm/ADT/ilist_node.h`

`ilist_node<T>` implements the forward and backward links that are expected by the `ilist<T>` (and analogous containers) in the default manner.

`ilist_node<T>`s are meant to be embedded in the node type `T`, usually `T` publicly derives from `ilist_node<T>`.

Sentinels

`ilists` have another specialty that must be considered. To be a good citizen in the C++ ecosystem, it needs to support the standard container operations, such as `begin` and `end` iterators, etc. Also, the `operator--` must work correctly on the `end` iterator in the case of non-empty `ilists`.

The only sensible solution to this problem is to allocate a so-called *sentinel* along with the intrusive list, which serves as the `end` iterator, providing the back-link to the last element. However conforming to the C++ convention it is illegal to `operator++` beyond the sentinel and it also must not be dereferenced.

These constraints allow for some implementation freedom to the `ilist` how to allocate and store the sentinel. The corresponding policy is dictated by `ilist_traits<T>`. By default a `T` gets heap-allocated whenever the need for a sentinel arises.

While the default policy is sufficient in most cases, it may break down when `T` does not provide a default constructor. Also, in the case of many instances of `ilists`, the memory overhead of the associated sentinels is wasted. To alleviate the situation with numerous and voluminous `T`-sentinels, sometimes a trick is employed, leading to *ghostly sentinels*.

Ghostly sentinels are obtained by specially-crafted `ilist_traits<T>` which superpose the sentinel with the `ilist` instance in memory. Pointer arithmetic is used to obtain the sentinel, which is relative to the `ilist`'s `this` pointer. The `ilist` is augmented by an extra pointer, which serves as the back-link of the sentinel. This is the only field in the ghostly sentinel which can be legally accessed.

Other Sequential Container options

Other STL containers are available, such as `std::string`.

There are also various STL adapter classes such as `std::queue`, `std::priority_queue`, `std::stack`, etc. These provide simplified access to an underlying container but don't affect the cost of the container itself.

String-like containers

There are a variety of ways to pass around and use strings in C and C++, and LLVM adds a few new options to choose from. Pick the first option on this list that will do what you need, they are ordered according to their relative cost.

Note that it is generally preferred to *not* pass strings around as `const char*`'s. These have a number of problems, including the fact that they cannot represent embedded nul (“0”) characters, and do not have a length available efficiently. The general replacement for ‘`const char*`’ is `StringRef`.

For more information on choosing string containers for APIs, please see [Passing Strings](#).

llvm/ADT/StringRef.h

The `StringRef` class is a simple value class that contains a pointer to a character and a length, and is quite related to the [ArrayRef](#) class (but specialized for arrays of characters). Because `StringRef` carries a length with it, it safely handles strings with embedded nul characters in it, getting the length does not require a `strlen` call, and it even has very convenient APIs for slicing and dicing the character range that it represents.

`StringRef` is ideal for passing simple strings around that are known to be live, either because they are C string literals, `std::string`, a C array, or a `SmallVector`. Each of these cases has an efficient implicit conversion to `StringRef`, which doesn't result in a dynamic `strlen` being executed.

`StringRef` has a few major limitations which make more powerful string containers useful:

1. You cannot directly convert a `StringRef` to a ‘`const char*`’ because there is no way to add a trailing nul (unlike the `.c_str()` method on various stronger classes).
2. `StringRef` doesn't own or keep alive the underlying string bytes. As such it can easily lead to dangling pointers, and is not suitable for embedding in datastructures in most cases (instead, use an `std::string` or something like that).
3. For the same reason, `StringRef` cannot be used as the return value of a method if the method “computes” the result string. Instead, use `std::string`.
4. `StringRef`'s do not allow you to mutate the pointed-to string bytes and it doesn't allow you to insert or remove bytes from the range. For editing operations like this, it interoperates with the [Twine](#) class.

Because of its strengths and limitations, it is very common for a function to take a `StringRef` and for a method on an object to return a `StringRef` that points into some string that it owns.

llvm/ADT/Twine.h

The `Twine` class is used as an intermediary datatype for APIs that want to take a string that can be constructed inline with a series of concatenations. `Twine` works by forming recursive instances of the `Twine` datatype (a simple value object) on the stack as temporary objects, linking them together into a tree which is then linearized when the `Twine` is consumed. `Twine` is only safe to use as the argument to a function, and should always be a `const` reference, e.g.:

```
void foo(const Twine &T);
...
StringRef X = ...
unsigned i = ...
foo(X + "." + Twine(i));
```

This example forms a string like “blarg.42” by concatenating the values together, and does not form intermediate strings containing “blarg” or “blarg.”.

Because Twine is constructed with temporary objects on the stack, and because these instances are destroyed at the end of the current statement, it is an inherently dangerous API. For example, this simple variant contains undefined behavior and will probably crash:

```
void foo(const Twine &T);
...
StringRef X = ...
unsigned i = ...
const Twine &Tmp = X + "." + Twine(i);
foo(Tmp);
```

... because the temporaries are destroyed before the call. That said, Twine's are much more efficient than intermediate `std::string` temporaries, and they work really well with `StringRef`. Just be aware of their limitations.

llvm/ADT/SmallString.h

`SmallString` is a subclass of *SmallVector* that adds some convenience APIs like `+=` that takes `StringRef`'s. `SmallString` avoids allocating memory in the case when the preallocated space is enough to hold its data, and it calls back to general heap allocation when required. Since it owns its data, it is very safe to use and supports full mutation of the string.

Like `SmallVector`'s, the big downside to `SmallString` is their `sizeof`. While they are optimized for small strings, they themselves are not particularly small. This means that they work great for temporary scratch buffers on the stack, but should not generally be put into the heap: it is very rare to see a `SmallString` as the member of a frequently-allocated heap data structure or returned by-value.

std::string

The standard C++ `std::string` class is a very general class that (like `SmallString`) owns its underlying data. `sizeof(std::string)` is very reasonable so it can be embedded into heap data structures and returned by-value. On the other hand, `std::string` is highly inefficient for inline editing (e.g. concatenating a bunch of stuff together) and because it is provided by the standard library, its performance characteristics depend a lot of the host standard library (e.g. `libc++` and `MSVC` provide a highly optimized string class, `GCC` contains a really slow implementation).

The major disadvantage of `std::string` is that almost every operation that makes them larger can allocate memory, which is slow. As such, it is better to use `SmallVector` or `Twine` as a scratch buffer, but then use `std::string` to persist the result.

Set-Like Containers (std::set, SmallSet, SetVector, etc)

Set-like containers are useful when you need to canonicalize multiple values into a single representation. There are several different choices for how to do this, providing various trade-offs.

A sorted 'vector'

If you intend to insert a lot of elements, then do a lot of queries, a great approach is to use a vector (or other sequential container) with `std::sort+std::unique` to remove duplicates. This approach works really well if your usage pattern has these two distinct phases (insert then query), and can be coupled with a good choice of *sequential container*.

This combination provides the several nice properties: the result data is contiguous in memory (good for cache locality), has few allocations, is easy to address (iterators in the final vector are just indices or pointers), and can be efficiently queried with a standard binary search (e.g. `std::lower_bound`; if you want the whole range of elements comparing equal, use `std::equal_range`).

llvm/ADT/SmallSet.h

If you have a set-like data structure that is usually small and whose elements are reasonably small, a `SmallSet<Type, N>` is a good choice. This set has space for `N` elements in place (thus, if the set is dynamically smaller than `N`, no malloc traffic is required) and accesses them with a simple linear search. When the set grows beyond `N` elements, it allocates a more expensive representation that guarantees efficient access (for most types, it falls back to `std::set`, but for pointers it uses something far better, *SmallPtrSet*).

The magic of this class is that it handles small sets extremely efficiently, but gracefully handles extremely large sets without loss of efficiency. The drawback is that the interface is quite small: it supports insertion, queries and erasing, but does not support iteration.

llvm/ADT/SmallPtrSet.h

`SmallPtrSet` has all the advantages of `SmallSet` (and a `SmallSet` of pointers is transparently implemented with a `SmallPtrSet`), but also supports iterators. If more than `N` insertions are performed, a single quadratically probed hash table is allocated and grows as needed, providing extremely efficient access (constant time insertion/deleting/queries with low constant factors) and is very stingy with malloc traffic.

Note that, unlike `std::set`, the iterators of `SmallPtrSet` are invalidated whenever an insertion occurs. Also, the values visited by the iterators are not visited in sorted order.

llvm/ADT/StringSet.h

`StringSet` is a thin wrapper around *StringMap<char>*, and it allows efficient storage and retrieval of unique strings.

Functionally analogous to `SmallSet<StringRef>`, `StringSet` also supports iteration. (The iterator dereferences to a `StringMapEntry<char>`, so you need to call `i->getKey()` to access the item of the `StringSet`.) On the other hand, `StringSet` doesn't support range-insertion and copy-construction, which *SmallSet* and *SmallPtrSet* do support.

llvm/ADT/DenseSet.h

`DenseSet` is a simple quadratically probed hash table. It excels at supporting small values: it uses a single allocation to hold all of the pairs that are currently inserted in the set. `DenseSet` is a great way to unique small values that are not simple pointers (use *SmallPtrSet* for pointers). Note that `DenseSet` has the same requirements for the value type that *DenseMap* has.

llvm/ADT/SparseSet.h

`SparseSet` holds a small number of objects identified by unsigned keys of moderate size. It uses a lot of memory, but provides operations that are almost as fast as a vector. Typical keys are physical registers, virtual registers, or numbered basic blocks.

`SparseSet` is useful for algorithms that need very fast clear/find/insert/erase and fast iteration over small sets. It is not intended for building composite data structures.

llvm/ADT/SparseMultiSet.h

`SparseMultiSet` adds multiset behavior to `SparseSet`, while retaining `SparseSet`'s desirable attributes. Like `SparseSet`, it typically uses a lot of memory, but provides operations that are almost as fast as a vector. Typical keys are physical

registers, virtual registers, or numbered basic blocks.

`SparseMultiSet` is useful for algorithms that need very fast clear/find/insert/erase of the entire collection, and iteration over sets of elements sharing a key. It is often a more efficient choice than using composite data structures (e.g. vector-of-vectors, map-of-vectors). It is not intended for building composite data structures.

llvm/ADT/FoldingSet.h

`FoldingSet` is an aggregate class that is really good at uniquing expensive-to-create or polymorphic objects. It is a combination of a chained hash table with intrusive links (uniqued objects are required to inherit from `FoldingSetNode`) that uses *SmallVector* as part of its ID process.

Consider a case where you want to implement a “getOrCreateFoo” method for a complex object (for example, a node in the code generator). The client has a description of **what** it wants to generate (it knows the opcode and all the operands), but we don’t want to ‘new’ a node, then try inserting it into a set only to find out it already exists, at which point we would have to delete it and return the node that already exists.

To support this style of client, `FoldingSet` perform a query with a `FoldingSetNodeID` (which wraps `SmallVector`) that can be used to describe the element that we want to query for. The query either returns the element matching the ID or it returns an opaque ID that indicates where insertion should take place. Construction of the ID usually does not require heap traffic.

Because `FoldingSet` uses intrusive links, it can support polymorphic objects in the set (for example, you can have `SDNode` instances mixed with `LoadSDNodes`). Because the elements are individually allocated, pointers to the elements are stable: inserting or removing elements does not invalidate any pointers to other elements.

<set>

`std::set` is a reasonable all-around set class, which is decent at many things but great at nothing. `std::set` allocates memory for each element inserted (thus it is very malloc intensive) and typically stores three pointers per element in the set (thus adding a large amount of per-element space overhead). It offers guaranteed log(n) performance, which is not particularly fast from a complexity standpoint (particularly if the elements of the set are expensive to compare, like strings), and has extremely high constant factors for lookup, insertion and removal.

The advantages of `std::set` are that its iterators are stable (deleting or inserting an element from the set does not affect iterators or pointers to other elements) and that iteration over the set is guaranteed to be in sorted order. If the elements in the set are large, then the relative overhead of the pointers and malloc traffic is not a big deal, but if the elements of the set are small, `std::set` is almost never a good choice.

llvm/ADT/SetVector.h

LLVM’s `SetVector<Type>` is an adapter class that combines your choice of a set-like container along with a *Sequential Container*. The important property that this provides is efficient insertion with uniquing (duplicate elements are ignored) with iteration support. It implements this by inserting elements into both a set-like container and the sequential container, using the set-like container for uniquing and the sequential container for iteration.

The difference between `SetVector` and other sets is that the order of iteration is guaranteed to match the order of insertion into the `SetVector`. This property is really important for things like sets of pointers. Because pointer values are non-deterministic (e.g. vary across runs of the program on different machines), iterating over the pointers in the set will not be in a well-defined order.

The drawback of `SetVector` is that it requires twice as much space as a normal set and has the sum of constant factors from the set-like container and the sequential container that it uses. Use it **only** if you need to iterate over the

elements in a deterministic order. `SetVector` is also expensive to delete elements out of (linear time), unless you use its “`pop_back`” method, which is faster.

`SetVector` is an adapter class that defaults to using `std::vector` and a size 16 `SmallSet` for the underlying containers, so it is quite expensive. However, “`llvm/ADT/SetVector.h`” also provides a `SmallSetVector` class, which defaults to using a `SmallVector` and `SmallSet` of a specified size. If you use this, and if your sets are dynamically smaller than `N`, you will save a lot of heap traffic.

llvm/ADT/UniqueVector.h

`UniqueVector` is similar to [SetVector](#) but it retains a unique ID for each element inserted into the set. It internally contains a map and a vector, and it assigns a unique ID for each value inserted into the set.

`UniqueVector` is very expensive: its cost is the sum of the cost of maintaining both the map and vector, it has high complexity, high constant factors, and produces a lot of malloc traffic. It should be avoided.

llvm/ADT/ImmutableSet.h

`ImmutableSet` is an immutable (functional) set implementation based on an AVL tree. Adding or removing elements is done through a `Factory` object and results in the creation of a new `ImmutableSet` object. If an `ImmutableSet` already exists with the given contents, then the existing one is returned; equality is compared with a `FoldingSetNodeID`. The time and space complexity of add or remove operations is logarithmic in the size of the original set.

There is no method for returning an element of the set, you can only check for membership.

Other Set-Like Container Options

The STL provides several other options, such as `std::multiset` and the various “`hash_set`” like containers (whether from C++ TR1 or from the SGI library). We never use `hash_set` and `unordered_set` because they are generally very expensive (each insertion requires a malloc) and very non-portable.

`std::multiset` is useful if you’re not interested in elimination of duplicates, but has all the drawbacks of [std::set](#). A sorted vector (where you don’t delete duplicate entries) or some other approach is almost always better.

Map-Like Containers (std::map, DenseMap, etc)

Map-like containers are useful when you want to associate data to a key. As usual, there are a lot of different ways to do this. :)

A sorted ‘vector’

If your usage pattern follows a strict insert-then-query approach, you can trivially use the same approach as [sorted vectors for set-like containers](#). The only difference is that your query function (which uses `std::lower_bound` to get efficient $\log(n)$ lookup) should only compare the key, not both the key and value. This yields the same advantages as sorted vectors for sets.

llvm/ADT/StringMap.h

Strings are commonly used as keys in maps, and they are difficult to support efficiently: they are variable length, inefficient to hash and compare when long, expensive to copy, etc. `StringMap` is a specialized container designed to cope with these issues. It supports mapping an arbitrary range of bytes to an arbitrary other object.

The StringMap implementation uses a quadratically-probed hash table, where the buckets store a pointer to the heap allocated entries (and some other stuff). The entries in the map must be heap allocated because the strings are variable length. The string data (key) and the element object (value) are stored in the same allocation with the string data immediately after the element object. This container guarantees the “(char*) (&Value+1)” points to the key string for a value.

The StringMap is very fast for several reasons: quadratic probing is very cache efficient for lookups, the hash value of strings in buckets is not recomputed when looking up an element, StringMap rarely has to touch the memory for unrelated objects when looking up a value (even when hash collisions happen), hash table growth does not recompute the hash values for strings already in the table, and each pair in the map is store in a single allocation (the string data is stored in the same allocation as the Value of a pair).

StringMap also provides query methods that take byte ranges, so it only ever copies a string if a value is inserted into the table.

StringMap iteration order, however, is not guaranteed to be deterministic, so any uses which require that should instead use a `std::map`.

llvm/ADT/IndexedMap.h

IndexedMap is a specialized container for mapping small dense integers (or values that can be mapped to small dense integers) to some other type. It is internally implemented as a vector with a mapping function that maps the keys to the dense integer range.

This is useful for cases like virtual registers in the LLVM code generator: they have a dense mapping that is offset by a compile-time constant (the first virtual register ID).

llvm/ADT/DenseMap.h

DenseMap is a simple quadratically probed hash table. It excels at supporting small keys and values: it uses a single allocation to hold all of the pairs that are currently inserted in the map. DenseMap is a great way to map pointers to pointers, or map other small types to each other.

There are several aspects of DenseMap that you should be aware of, however. The iterators in a DenseMap are invalidated whenever an insertion occurs, unlike `map`. Also, because DenseMap allocates space for a large number of key/value pairs (it starts with 64 by default), it will waste a lot of space if your keys or values are large. Finally, you must implement a partial specialization of `DenseMapInfo` for the key that you want, if it isn't already supported. This is required to tell DenseMap about two special marker values (which can never be inserted into the map) that it needs internally.

DenseMap's `find_as()` method supports lookup operations using an alternate key type. This is useful in cases where the normal key type is expensive to construct, but cheap to compare against. The `DenseMapInfo` is responsible for defining the appropriate comparison and hashing methods for each alternate key type used.

llvm/IR/ValueMap.h

ValueMap is a wrapper around a *DenseMap* mapping `Value*s` (or subclasses) to another type. When a `Value` is deleted or RAUW'ed, ValueMap will update itself so the new version of the key is mapped to the same value, just as if the key were a `WeakVH`. You can configure exactly how this happens, and what else happens on these two events, by passing a `Config` parameter to the ValueMap template.

llvm/ADT/IntervalMap.h

IntervalMap is a compact map for small keys and values. It maps key intervals instead of single keys, and it will automatically coalesce adjacent intervals. When the map only contains a few intervals, they are stored in the map object itself to avoid allocations.

The IntervalMap iterators are quite big, so they should not be passed around as STL iterators. The heavyweight iterators allow a smaller data structure.

<map>

std::map has similar characteristics to *std::set*: it uses a single allocation per pair inserted into the map, it offers log(n) lookup with an extremely large constant factor, imposes a space penalty of 3 pointers per pair in the map, etc.

std::map is most useful when your keys or values are very large, if you need to iterate over the collection in sorted order, or if you need stable iterators into the map (i.e. they don't get invalidated if an insertion or deletion of another element takes place).

llvm/ADT/MapVector.h

MapVector<KeyT, ValueT> provides a subset of the DenseMap interface. The main difference is that the iteration order is guaranteed to be the insertion order, making it an easy (but somewhat expensive) solution for non-deterministic iteration over maps of pointers.

It is implemented by mapping from key to an index in a vector of key,value pairs. This provides fast lookup and iteration, but has two main drawbacks: the key is stored twice and removing elements takes linear time. If it is necessary to remove elements, it's best to remove them in bulk using `remove_if()`.

llvm/ADT/IntEqClasses.h

IntEqClasses provides a compact representation of equivalence classes of small integers. Initially, each integer in the range 0..n-1 has its own equivalence class. Classes can be joined by passing two class representatives to the `join(a, b)` method. Two integers are in the same class when `findLeader()` returns the same representative.

Once all equivalence classes are formed, the map can be compressed so each integer 0..n-1 maps to an equivalence class number in the range 0..m-1, where m is the total number of equivalence classes. The map must be uncompressed before it can be edited again.

llvm/ADT/ImmutableMap.h

ImmutableMap is an immutable (functional) map implementation based on an AVL tree. Adding or removing elements is done through a Factory object and results in the creation of a new ImmutableMap object. If an ImmutableMap already exists with the given key set, then the existing one is returned; equality is compared with a FoldingSetNodeID. The time and space complexity of add or remove operations is logarithmic in the size of the original map.

Other Map-Like Container Options

The STL provides several other options, such as `std::multimap` and the various “hash_map” like containers (whether from C++ TR1 or from the SGI library). We never use `hash_set` and `unordered_set` because they are generally very expensive (each insertion requires a malloc) and very non-portable.

`std::multimap` is useful if you want to map a key to multiple values, but has all the drawbacks of `std::map`. A sorted vector or some other approach is almost always better.

Bit storage containers (`BitVector`, `SparseBitVector`)

Unlike the other containers, there are only two bit storage containers, and choosing when to use each is relatively straightforward.

One additional option is `std::vector<bool>`: we discourage its use for two reasons 1) the implementation in many common compilers (e.g. commonly available versions of GCC) is extremely inefficient and 2) the C++ standards committee is likely to deprecate this container and/or change it significantly somehow. In any case, please don't use it.

`BitVector`

The `BitVector` container provides a dynamic size set of bits for manipulation. It supports individual bit setting/testing, as well as set operations. The set operations take time $O(\text{size of bitvector})$, but operations are performed one word at a time, instead of one bit at a time. This makes the `BitVector` very fast for set operations compared to other containers. Use the `BitVector` when you expect the number of set bits to be high (i.e. a dense set).

`SmallBitVector`

The `SmallBitVector` container provides the same interface as `BitVector`, but it is optimized for the case where only a small number of bits, less than 25 or so, are needed. It also transparently supports larger bit counts, but slightly less efficiently than a plain `BitVector`, so `SmallBitVector` should only be used when larger counts are rare.

At this time, `SmallBitVector` does not support set operations (and, or, xor), and its `operator[]` does not provide an assignable lvalue.

`SparseBitVector`

The `SparseBitVector` container is much like `BitVector`, with one major difference: Only the bits that are set, are stored. This makes the `SparseBitVector` much more space efficient than `BitVector` when the set is sparse, as well as making set operations $O(\text{number of set bits})$ instead of $O(\text{size of universe})$. The downside to the `SparseBitVector` is that setting and testing of random bits is $O(N)$, and on large `SparseBitVectors`, this can be slower than `BitVector`. In our implementation, setting or testing bits in sorted order (either forwards or reverse) is $O(1)$ worst case. Testing and setting bits within 128 bits (depends on size) of the current bit is also $O(1)$. As a general statement, testing/setting bits in a `SparseBitVector` is $O(\text{distance away from last set bit})$.

3.7.5 Debugging

A handful of [GDB pretty printers](#) are provided for some of the core LLVM libraries. To use them, execute the following (or add it to your `~/.gdbinit`):

```
source /path/to/llvm/src/utils/gdb-scripts/prettyprinters.py
```

It also might be handy to enable the `print pretty` option to avoid data structures being printed as a big block of text.

3.7.6 Helpful Hints for Common Operations

This section describes how to perform some very simple transformations of LLVM code. This is meant to give examples of common idioms used, showing the practical side of LLVM transformations.

Because this is a “how-to” section, you should also read about the main classes that you will be working with. The *Core LLVM Class Hierarchy Reference* contains details and descriptions of the main classes that you should know about.

Basic Inspection and Traversal Routines

The LLVM compiler infrastructure have many different data structures that may be traversed. Following the example of the C++ standard template library, the techniques used to traverse these various data structures are all basically the same. For an enumerable sequence of values, the `XXXbegin()` function (or method) returns an iterator to the start of the sequence, the `XXXend()` function returns an iterator pointing to one past the last valid element of the sequence, and there is some `XXXiterator` data type that is common between the two operations.

Because the pattern for iteration is common across many different aspects of the program representation, the standard template library algorithms may be used on them, and it is easier to remember how to iterate. First we show a few common examples of the data structures that need to be traversed. Other data structures are traversed in very similar ways.

Iterating over the BasicBlock in a Function

It’s quite common to have a `Function` instance that you’d like to transform in some way; in particular, you’d like to manipulate its `BasicBlocks`. To facilitate this, you’ll need to iterate over all of the `BasicBlocks` that constitute the `Function`. The following is an example that prints the name of a `BasicBlock` and the number of instructions it contains:

```
Function &Func = ...
for (BasicBlock &BB : Func)
    // Print out the name of the basic block if it has one, and then the
    // number of instructions that it contains
    errs() << "Basic block (name=" << BB.getName() << ") has "
            << BB.size() << " instructions.\n";
```

Iterating over the Instruction in a BasicBlock

Just like when dealing with `BasicBlocks` in `Functions`, it’s easy to iterate over the individual instructions that make up `BasicBlocks`. Here’s a code snippet that prints out each instruction in a `BasicBlock`:

```
BasicBlock& BB = ...
for (Instruction &I : BB)
    // The next statement works since operator<<(ostream&,...)
    // is overloaded for Instruction&
    errs() << I << "\n";
```

However, this isn’t really the best way to print out the contents of a `BasicBlock`! Since the ostream operators are overloaded for virtually anything you’ll care about, you could have just invoked the print routine on the basic block itself: `errs() << BB << "\n";`.

Iterating over the Instruction in a Function

If you're finding that you commonly iterate over a Function's BasicBlocks and then that BasicBlock's Instructions, `InstIterator` should be used instead. You'll need to include `llvm/IR/InstIterator.h` ([doxygen](#)) and then instantiate `InstIterators` explicitly in your code. Here's a small example that shows how to dump all instructions in a function to the standard error stream:

```
#include "llvm/IR/InstIterator.h"

// F is a pointer to a Function instance
for (inst_iterator I = inst_begin(F), E = inst_end(F); I != E; ++I)
    errs() << *I << "\n";
```

Easy, isn't it? You can also use `InstIterators` to fill a work list with its initial contents. For example, if you wanted to initialize a work list to contain all instructions in a Function `F`, all you would need to do is something like:

```
std::set<Instruction> worklist;
// or better yet, SmallPtrSet<Instruction*, 64> worklist;

for (inst_iterator I = inst_begin(F), E = inst_end(F); I != E; ++I)
    worklist.insert(&*I);
```

The STL set `worklist` would now contain all instructions in the Function pointed to by `F`.

Turning an iterator into a class pointer (and vice-versa)

Sometimes, it'll be useful to grab a reference (or pointer) to a class instance when all you've got at hand is an iterator. Well, extracting a reference or a pointer from an iterator is very straight-forward. Assuming that `i` is a `BasicBlock::iterator` and `j` is a `BasicBlock::const_iterator`:

```
Instruction& inst = *i; // Grab reference to instruction reference
Instruction* pinst = &*i; // Grab pointer to instruction reference
const Instruction& inst = *j;
```

However, the iterators you'll be working with in the LLVM framework are special: they will automatically convert to a ptr-to-instance type whenever they need to. Instead of dereferencing the iterator and then taking the address of the result, you can simply assign the iterator to the proper pointer type and you get the dereference and address-of operation as a result of the assignment (behind the scenes, this is a result of overloading casting mechanisms). Thus the second line of the last example,

```
Instruction *pinst = &*i;
```

is semantically equivalent to

```
Instruction *pinst = i;
```

It's also possible to turn a class pointer into the corresponding iterator, and this is a constant time operation (very efficient). The following code snippet illustrates use of the conversion constructors provided by LLVM iterators. By using these, you can explicitly grab the iterator of something without actually obtaining it via iteration over some structure:

```
void printNextInstruction(Instruction* inst) {
    BasicBlock::iterator it(inst);
    ++it; // After this line, it refers to the instruction after *inst
```

```
if (it != inst->getParent()->end()) errs() << *it << "\n";
}
```

Unfortunately, these implicit conversions come at a cost; they prevent these iterators from conforming to standard iterator conventions, and thus from being usable with standard algorithms and containers. For example, they prevent the following code, where `B` is a `BasicBlock`, from compiling:

```
llvm::SmallVector<llvm::Instruction *, 16>(B->begin(), B->end());
```

Because of this, these implicit conversions may be removed some day, and `operator*` changed to return a pointer instead of a reference.

Finding call sites: a slightly more complex example

Say that you're writing a `FunctionPass` and would like to count all the locations in the entire module (that is, across every `Function`) where a certain function (i.e., some `Function *`) is already in scope. As you'll learn later, you may want to use an `InstVisitor` to accomplish this in a much more straight-forward manner, but this example will allow us to explore how you'd do it if you didn't have `InstVisitor` around. In pseudo-code, this is what we want to do:

```
initialize callCounter to zero
for each Function f in the Module
  for each BasicBlock b in f
    for each Instruction i in b
      if (i is a CallInst and calls the given function)
        increment callCounter
```

And the actual code is (remember, because we're writing a `FunctionPass`, our `FunctionPass`-derived class simply has to override the `runOnFunction` method):

```
Function* targetFunc = ...;

class OurFunctionPass : public FunctionPass {
public:
    OurFunctionPass(): callCounter(0) { }

    virtual runOnFunction(Function& F) {
        for (BasicBlock &B : F) {
            for (Instruction &I: B) {
                if (auto *CallInst = dyn_cast<CallInst>(&I)) {
                    // We know we've encountered a call instruction, so we
                    // need to determine if it's a call to the
                    // function pointed to by m_func or not.
                    if (CallInst->getCalledFunction() == targetFunc)
                        ++callCounter;
                }
            }
        }
    }

private:
    unsigned callCounter;
};
```

Treating calls and invokes the same way

You may have noticed that the previous example was a bit oversimplified in that it did not deal with call sites generated by ‘invoke’ instructions. In this, and in other situations, you may find that you want to treat `CallInsts` and `InvokeInsts` the same way, even though their most-specific common base class is `Instruction`, which includes lots of less closely-related things. For these cases, LLVM provides a handy wrapper class called `CallSite` (doxygen) It is essentially a wrapper around an `Instruction` pointer, with some methods that provide functionality common to `CallInsts` and `InvokeInsts`.

This class has “value semantics”: it should be passed by value, not by reference and it should not be dynamically allocated or deallocated using `operator new` or `operator delete`. It is efficiently copyable, assignable and constructable, with costs equivalents to that of a bare pointer. If you look at its definition, it has only a single pointer member.

Iterating over def-use & use-def chains

Frequently, we might have an instance of the `Value` class (doxygen) and we want to determine which `User s` use the `Value`. The list of all `Users` of a particular `Value` is called a *def-use* chain. For example, let’s say we have a `Function*` named `F` to a particular function `foo`. Finding all of the instructions that *use* `foo` is as simple as iterating over the *def-use* chain of `F`:

```
Function *F = ...;

for (User *U : F->users()) {
    if (Instruction *Inst = dyn_cast<Instruction>(U)) {
        errs() << "F is used in instruction:\n";
        errs() << *Inst << "\n";
    }
}
```

Alternatively, it’s common to have an instance of the `User` Class (doxygen) and need to know what `Values` are used by it. The list of all `Values` used by a `User` is known as a *use-def* chain. Instances of class `Instruction` are common `User s`, so we might want to iterate over all of the values that a particular instruction uses (that is, the operands of the particular `Instruction`):

```
Instruction *pi = ...;

for (Use &U : pi->operands()) {
    Value *v = U.get();
    // ...
}
```

Declaring objects as `const` is an important tool of enforcing mutation free algorithms (such as analyses, etc.). For this purpose above iterators come in constant flavors as `Value::const_use_iterator` and `Value::const_op_iterator`. They automatically arise when calling `use/op_begin()` on `const Value*s` or `const User*s` respectively. Upon dereferencing, they return `const Use*s`. Otherwise the above patterns remain unchanged.

Iterating over predecessors & successors of blocks

Iterating over the predecessors and successors of a block is quite easy with the routines defined in `"llvm/IR/CFG.h"`. Just use code like this to iterate over all predecessors of `BB`:

```
#include "llvm/IR/CFG.h"
BasicBlock *BB = ...;
```



```
for (BasicBlock *Pred : predecessors(BB)) {
    // ...
}
```

Similarly, to iterate over successors use `successors`.

Making simple changes

There are some primitive transformation operations present in the LLVM infrastructure that are worth knowing about. When performing transformations, it's fairly common to manipulate the contents of basic blocks. This section describes some of the common methods for doing so and gives example code.

Creating and inserting new Instructions

Instantiating Instructions

Creation of `Instructions` is straight-forward: simply call the constructor for the kind of instruction to instantiate and provide the necessary parameters. For example, an `AllocaInst` only *requires* a (const-`ptr-to`) `Type`. Thus:

```
auto *ai = new AllocaInst(Type::Int32Ty);
```

will create an `AllocaInst` instance that represents the allocation of one integer in the current stack frame, at run time. Each `Instruction` subclass is likely to have varying default parameters which change the semantics of the instruction, so refer to the [doxygen documentation for the subclass of `Instruction`](#) that you're interested in instantiating.

Naming values

It is very useful to name the values of instructions when you're able to, as this facilitates the debugging of your transformations. If you end up looking at generated LLVM machine code, you definitely want to have logical names associated with the results of instructions! By supplying a value for the `Name` (default) parameter of the `Instruction` constructor, you associate a logical name with the result of the instruction's execution at run time. For example, say that I'm writing a transformation that dynamically allocates space for an integer on the stack, and that integer is going to be used as some kind of index by some other code. To accomplish this, I place an `AllocaInst` at the first point in the first `BasicBlock` of some `Function`, and I'm intending to use it within the same `Function`. I might do:

```
auto *pa = new AllocaInst(Type::Int32Ty, 0, "indexLoc");
```

where `indexLoc` is now the logical name of the instruction's execution value, which is a pointer to an integer on the run time stack.

Inserting instructions

There are essentially three ways to insert an `Instruction` into an existing sequence of instructions that form a `BasicBlock`:

- Insertion into an explicit instruction list

Given a `BasicBlock*` `pb`, an `Instruction*` `pi` within that `BasicBlock`, and a newly-created instruction we wish to insert before `*pi`, we do the following:

```
BasicBlock *pb = ...;
Instruction *pi = ...;
auto *newInst = new Instruction(...);

pb->getInstList().insert(pi, newInst); // Inserts newInst before pi in pb
```

Appending to the end of a `BasicBlock` is so common that the `Instruction` class and `Instruction`-derived classes provide constructors which take a pointer to a `BasicBlock` to be appended to. For example code that looked like:

```
BasicBlock *pb = ...;
auto *newInst = new Instruction(...);

pb->getInstList().push_back(newInst); // Appends newInst to pb
```

becomes:

```
BasicBlock *pb = ...;
auto *newInst = new Instruction(..., pb);
```

which is much cleaner, especially if you are creating long instruction streams.

- Insertion into an implicit instruction list

`Instruction` instances that are already in `BasicBlocks` are implicitly associated with an existing instruction list: the instruction list of the enclosing basic block. Thus, we could have accomplished the same thing as the above code without being given a `BasicBlock` by doing:

```
Instruction *pi = ...;
auto *newInst = new Instruction(...);

pi->getParent()->getInstList().insert(pi, newInst);
```

In fact, this sequence of steps occurs so frequently that the `Instruction` class and `Instruction`-derived classes provide constructors which take (as a default parameter) a pointer to an `Instruction` which the newly-created `Instruction` should precede. That is, `Instruction` constructors are capable of inserting the newly-created instance into the `BasicBlock` of a provided instruction, immediately before that instruction. Using an `Instruction` constructor with a `insertBefore` (default) parameter, the above code becomes:

```
Instruction* pi = ...;
auto *newInst = new Instruction(..., pi);
```

which is much cleaner, especially if you're creating a lot of instructions and adding them to `BasicBlocks`.

- Insertion using an instance of `IRBuilder`

Inserting several `Instructions` can be quite laborious using the previous methods. The `IRBuilder` is a convenience class that can be used to add several instructions to the end of a `BasicBlock` or before a particular `Instruction`. It also supports constant folding and renaming named registers (see `IRBuilder`'s template arguments).

The example below demonstrates a very simple use of the `IRBuilder` where three instructions are inserted before the instruction `pi`. The first two instructions are `Call` instructions and third instruction multiplies the return value of the two calls.

```
Instruction *pi = ...;
IRBuilder<> Builder(pi);
CallInst* callOne = Builder.CreateCall(...);
CallInst* callTwo = Builder.CreateCall(...);
Value* result = Builder.CreateMul(callOne, callTwo);
```

The example below is similar to the above example except that the created `IRBuilder` inserts instructions at the end of the `BasicBlock` `pb`.

```
BasicBlock *pb = ...;
IRBuilder<> Builder(pb);
CallInst* callOne = Builder.CreateCall(...);
CallInst* callTwo = Builder.CreateCall(...);
Value* result = Builder.CreateMul(callOne, callTwo);
```

See *Kaleidoscope: Code generation to LLVM IR* for a practical use of the `IRBuilder`.

Deleting Instructions

Deleting an instruction from an existing sequence of instructions that form a *BasicBlock* is very straight-forward: just call the instruction's `eraseFromParent()` method. For example:

```
Instruction *I = .. ;
I->eraseFromParent();
```

This unlinks the instruction from its containing basic block and deletes it. If you'd just like to unlink the instruction from its containing basic block but not delete it, you can use the `removeFromParent()` method.

Replacing an Instruction with another Value

Replacing individual instructions

Including “`llvm/Transforms/Utils/BasicBlockUtils.h`” permits use of two very useful replace functions: `ReplaceInstWithValue` and `ReplaceInstWithInst`.

Deleting Instructions

- `ReplaceInstWithValue`

This function replaces all uses of a given instruction with a value, and then removes the original instruction. The following example illustrates the replacement of the result of a particular `AllocaInst` that allocates memory for a single integer with a null pointer to an integer.

```
AllocaInst* instToReplace = ...;
BasicBlock::iterator ii(instToReplace);

ReplaceInstWithValue(instToReplace->getParent()->getInstList(), ii,
    Constant::getNullValue(PointerTy::getUnqual(Type::Int32Ty)));
```

- `ReplaceInstWithInst`

This function replaces a particular instruction with another instruction, inserting the new instruction into the basic block at the location where the old instruction was, and replacing any uses of the old instruction with the new instruction. The following example illustrates the replacement of one `AllocaInst` with another.

```
AllocaInst* instToReplace = ...;
BasicBlock::iterator ii(instToReplace);

ReplaceInstWithInst(instToReplace->getParent()->getInstList(), ii,
    new AllocaInst(Type::Int32Ty, 0, "ptrToReplacedInt"));
```

Replacing multiple uses of Users and Values

You can use `Value::replaceAllUsesWith` and `User::replaceAllUsesOfWith` to change more than one use at a time. See the [doxygen documentation](#) for the [Value Class](#) and [User Class](#), respectively, for more information.

Deleting Global Variables

Deleting a global variable from a module is just as easy as deleting an Instruction. First, you must have a pointer to the global variable that you wish to delete. You use this pointer to erase it from its parent, the module. For example:

```
GlobalVariable *GV = .. ;  
  
GV->eraseFromParent();
```

How to Create Types

In generating IR, you may need some complex types. If you know these types statically, you can use `TypeBuilder<...>::get()`, defined in `llvm/Support/TypeBuilder.h`, to retrieve them. `TypeBuilder` has two forms depending on whether you're building types for cross-compilation or native library use. `TypeBuilder<T, true>` requires that `T` be independent of the host environment, meaning that it's built out of types from the `llvm::types` ([doxygen](#)) namespace and pointers, functions, arrays, etc. built of those. `TypeBuilder<T, false>` additionally allows native C types whose size may depend on the host compiler. For example,

```
FunctionType *ft = TypeBuilder<types::i<8>(types::i<32>*), true>::get();
```

is easier to read and write than the equivalent

```
std::vector<const Type*> params;  
params.push_back(PointerType::getUnqual(Type::Int32Ty));  
FunctionType *ft = FunctionType::get(Type::Int8Ty, params, false);
```

See the [class comment](#) for more details.

3.7.7 Threads and LLVM

This section describes the interaction of the LLVM APIs with multithreading, both on the part of client applications, and in the JIT, in the hosted application.

Note that LLVM's support for multithreading is still relatively young. Up through version 2.5, the execution of threaded hosted applications was supported, but not threaded client access to the APIs. While this use case is now supported, clients *must* adhere to the guidelines specified below to ensure proper operation in multithreaded mode.

Note that, on Unix-like platforms, LLVM requires the presence of GCC's atomic intrinsics in order to support threaded operation. If you need a multithreading-capable LLVM on a platform without a suitably modern system compiler, consider compiling LLVM and LLVM-GCC in single-threaded mode, and using the resultant compiler to build a copy of LLVM with multithreading support.

Ending Execution with `llvm_shutdown()`

When you are done using the LLVM APIs, you should call `llvm_shutdown()` to deallocate memory used for internal structures.

Lazy Initialization with `ManagedStatic`

`ManagedStatic` is a utility class in LLVM used to implement static initialization of static resources, such as the global type tables. In a single-threaded environment, it implements a simple lazy initialization scheme. When LLVM is compiled with support for multi-threading, however, it uses double-checked locking to implement thread-safe lazy initialization.

Achieving Isolation with `LLVMContext`

`LLVMContext` is an opaque class in the LLVM API which clients can use to operate multiple, isolated instances of LLVM concurrently within the same address space. For instance, in a hypothetical compile-server, the compilation of an individual translation unit is conceptually independent from all the others, and it would be desirable to be able to compile incoming translation units concurrently on independent server threads. Fortunately, `LLVMContext` exists to enable just this kind of scenario!

Conceptually, `LLVMContext` provides isolation. Every LLVM entity (Modules, Values, Types, Constants, etc.) in LLVM's in-memory IR belongs to an `LLVMContext`. Entities in different contexts *cannot* interact with each other: Modules in different contexts cannot be linked together, Functions cannot be added to Modules in different contexts, etc. What this means is that it is safe to compile on multiple threads simultaneously, as long as no two threads operate on entities within the same context.

In practice, very few places in the API require the explicit specification of a `LLVMContext`, other than the Type creation/lookup APIs. Because every Type carries a reference to its owning context, most other entities can determine what context they belong to by looking at their own Type. If you are adding new entities to LLVM IR, please try to maintain this interface design.

Threads and the JIT

LLVM's "eager" JIT compiler is safe to use in threaded programs. Multiple threads can call `ExecutionEngine::getPointerToFunction()` or `ExecutionEngine::runFunction()` concurrently, and multiple threads can run code output by the JIT concurrently. The user must still ensure that only one thread accesses IR in a given `LLVMContext` while another thread might be modifying it. One way to do that is to always hold the JIT lock while accessing IR outside the JIT (the JIT *modifies* the IR by adding `CallbackVHs`). Another way is to only call `getPointerToFunction()` from the `LLVMContext`'s thread.

When the JIT is configured to compile lazily (using `ExecutionEngine::DisableLazyCompilation(false)`), there is currently a *race condition* in updating call sites after a function is lazily-jitted. It's still possible to use the lazy JIT in a threaded program if you ensure that only one thread at a time can call any particular lazy stub and that the JIT lock guards any IR access, but we suggest using only the eager JIT in threaded programs.

3.7.8 Advanced Topics

This section describes some of the advanced or obscure API's that most clients do not need to be aware of. These API's tend manage the inner workings of the LLVM system, and only need to be accessed in unusual circumstances.

The `ValueSymbolTable` class

The `ValueSymbolTable` ([doxygen](#)) class provides a symbol table that the *Function* and *Module* classes use for naming value definitions. The symbol table can provide a name for any *Value*.

Note that the `SymbolTable` class should not be directly accessed by most clients. It should only be used when iteration over the symbol table names themselves are required, which is very special purpose. Note that not all LLVM *Values* have names, and those without names (i.e. they have an empty name) do not exist in the symbol table.

Symbol tables support iteration over the values in the symbol table with `begin/end/iterator` and supports querying to see if a specific name is in the symbol table (with `lookup`). The `ValueSymbolTable` class exposes no public mutator methods, instead, simply call `setName` on a value, which will autoinsert it into the appropriate symbol table.

The `User` and owned `Use` classes' memory layout

The `User` ([doxygen](#)) class provides a basis for expressing the ownership of `User` towards other `Value` instances. The `Use` ([doxygen](#)) helper class is employed to do the bookkeeping and to facilitate $O(1)$ addition and removal.

Interaction and relationship between `User` and `Use` objects

A subclass of `User` can choose between incorporating its `Use` objects or refer to them out-of-line by means of a pointer. A mixed variant (some `Use` s inline others hung off) is impractical and breaks the invariant that the `Use` objects belonging to the same `User` form a contiguous array.

We have 2 different layouts in the `User` (sub)classes:

- Layout a)

The `Use` object(s) are inside (resp. at fixed offset) of the `User` object and there are a fixed number of them.

- Layout b)

The `Use` object(s) are referenced by a pointer to an array from the `User` object and there may be a variable number of them.

As of v2.4 each layout still possesses a direct pointer to the start of the array of `Uses`. Though not mandatory for layout a), we stick to this redundancy for the sake of simplicity. The `User` object also stores the number of `Use` objects it has. (Theoretically this information can also be calculated given the scheme presented below.)

Special forms of allocation operators (`operator new`) enforce the following memory layouts:

- Layout a) is modelled by prepending the `User` object by the `Use []` array.

```
.....
| P | P | P | P | User
'-----'
```

- Layout b) is modelled by pointing at the `Use []` array.

```
.....
| User
'-----'
|
v
.....
| P | P | P | P |
'-----'
```

(In the above figures 'P' stands for the `Use**` that is stored in each `Use` object in the member `Use::Prev`)

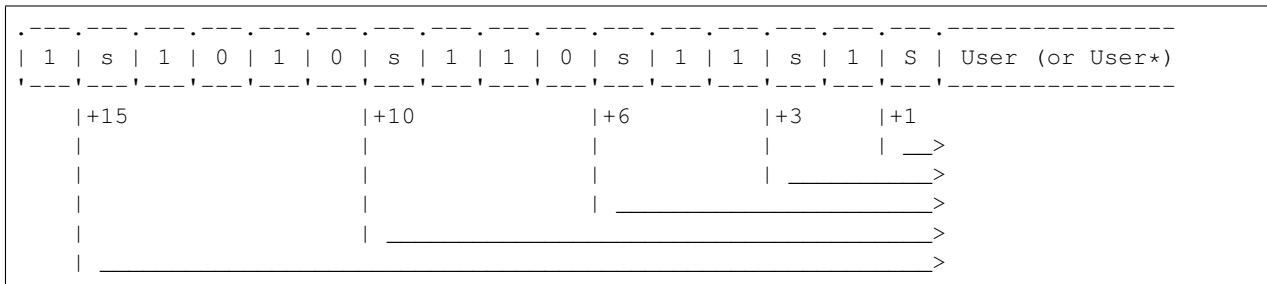
The waymarking algorithm

Since the `Use` objects are deprived of the direct (back)pointer to their `User` objects, there must be a fast and exact method to recover it. This is accomplished by the following scheme:

A bit-encoding in the 2 LSBits (least significant bits) of the `Use::Prev` allows to find the start of the `User` object:

- 00 — binary digit 0
- 01 — binary digit 1
- 10 — stop and calculate (s)
- 11 — full stop (S)

Given a `Use*`, all we have to do is to walk till we get a stop and we either have a `User` immediately behind or we have to walk to the next stop picking up digits and calculating the offset:



Only the significant number of bits need to be stored between the stops, so that the *worst case is 20 memory accesses* when there are 1000 `Use` objects associated with a `User`.

Reference implementation

The following literate Haskell fragment demonstrates the concept:

```
> import Test.QuickCheck
>
> digits :: Int -> [Char] -> [Char]
> digits 0 acc = '0' : acc
> digits 1 acc = '1' : acc
> digits n acc = digits (n `div` 2) $ digits (n `mod` 2) acc
>
> dist :: Int -> [Char] -> [Char]
> dist 0 [] = ['S']
> dist 0 acc = acc
> dist 1 acc = let r = dist 0 acc in 's' : digits (length r) r
> dist n acc = dist (n - 1) $ dist 1 acc
>
> takeLast n ss = reverse $ take n $ reverse ss
>
> test = takeLast 40 $ dist 20 []
>
```

Printing `<test>` gives: `"1s100000s11010s10100s1111s1010s110s11s1S"`

The reverse algorithm computes the length of the string just by examining a certain prefix:

```
> pref :: [Char] -> Int
> pref "S" = 1
> pref ('s':'1':rest) = decode 2 1 rest
> pref (_:rest) = 1 + pref rest
>
> decode walk acc ('0':rest) = decode (walk + 1) (acc * 2) rest
> decode walk acc ('1':rest) = decode (walk + 1) (acc * 2 + 1) rest
```

```
> decode walk acc _ = walk + acc
>
```

Now, as expected, printing <pref test> gives 40.

We can *quickCheck* this with following property:

```
> testcase = dist 2000 []
> testcaseLength = length testcase
>
> identityProp n = n > 0 && n <= testcaseLength ==> length arr == pref arr
>   where arr = takeLast n testcase
>
```

As expected <quickCheck identityProp> gives:

```
*Main> quickCheck identityProp
OK, passed 100 tests.
```

Let's be a bit more exhaustive:

```
>
> deepCheck p = check (defaultConfig { configMaxTest = 500 }) p
>
```

And here is the result of <deepCheck identityProp>:

```
*Main> deepCheck identityProp
OK, passed 500 tests.
```

Tagging considerations

To maintain the invariant that the 2 LSBits of each `Use**` in `Use` never change after being set up, setters of `Use :: Prev` must re-tag the new `Use**` on every modification. Accordingly getters must strip the tag bits.

For layout b) instead of the `User` we find a pointer (`User*` with LSBit set). Following this pointer brings us to the `User`. A portable trick ensures that the first bytes of `User` (if interpreted as a pointer) never has the LSBit set. (Portability is relying on the fact that all known compilers place the `vptr` in the first word of the instances.)

Designing Type Hiercharies and Polymorphic Interfaces

There are two different design patterns that tend to result in the use of virtual dispatch for methods in a type hierarchy in C++ programs. The first is a genuine type hierarchy where different types in the hierarchy model a specific subset of the functionality and semantics, and these types nest strictly within each other. Good examples of this can be seen in the `Value` or `Type` type hierarchies.

A second is the desire to dispatch dynamically across a collection of polymorphic interface implementations. This latter use case can be modeled with virtual dispatch and inheritance by defining an abstract interface base class which all implementations derive from and override. However, this implementation strategy forces an “**is-a**” relationship to exist that is not actually meaningful. There is often not some nested hierarchy of useful generalizations which code might interact with and move up and down. Instead, there is a singular interface which is dispatched across a range of implementations.

The preferred implementation strategy for the second use case is that of generic programming (sometimes called “compile-time duck typing” or “static polymorphism”). For example, a template over some type parameter `T` can be

instantiated across any particular implementation that conforms to the interface or *concept*. A good example here is the highly generic properties of any type which models a node in a directed graph. LLVM models these primarily through templates and generic programming. Such templates include the `LoopInfoBase` and `DominatorTreeBase`. When this type of polymorphism truly needs **dynamic** dispatch you can generalize it using a technique called *concept-based polymorphism*. This pattern emulates the interfaces and behaviors of templates using a very limited form of virtual dispatch for type erasure inside its implementation. You can find examples of this technique in the `PassManager.h` system, and there is a more detailed introduction to it by Sean Parent in several of his talks and papers:

1. [Inheritance Is The Base Class of Evil](#) - The GoingNative 2013 talk describing this technique, and probably the best place to start.
2. [Value Semantics and Concepts-based Polymorphism](#) - The C++Now! 2012 talk describing this technique in more detail.
3. [Sean Parent's Papers and Presentations](#) - A Github project full of links to slides, video, and sometimes code.

When deciding between creating a type hierarchy (with either tagged or virtual dispatch) and using templates or concepts-based polymorphism, consider whether there is some refinement of an abstract base class which is a semantically meaningful type on an interface boundary. If anything more refined than the root abstract interface is meaningless to talk about as a partial extension of the semantic model, then your use case likely fits better with polymorphism and you should avoid using virtual dispatch. However, there may be some exigent circumstances that require one technique or the other to be used.

If you do need to introduce a type hierarchy, we prefer to use explicitly closed type hierarchies with manual tagged dispatch and/or RTTI rather than the open inheritance model and virtual dispatch that is more common in C++ code. This is because LLVM rarely encourages library consumers to extend its core types, and leverages the closed and tag-dispatched nature of its hierarchies to generate significantly more efficient code. We have also found that a large amount of our usage of type hierarchies fits better with tag-based pattern matching rather than dynamic dispatch across a common interface. Within LLVM we have built custom helpers to facilitate this design. See this document's section on *isa and dyn_cast* and our [detailed document](#) which describes how you can implement this pattern for use with the LLVM helpers.

ABI Breaking Checks

Checks and asserts that alter the LLVM C++ ABI are predicated on the preprocessor symbol `LLVM_ENABLE_ABI_BREAKING_CHECKS` – LLVM libraries built with `LLVM_ENABLE_ABI_BREAKING_CHECKS` are not ABI compatible LLVM libraries built without it defined. By default, turning on assertions also turns on `LLVM_ENABLE_ABI_BREAKING_CHECKS` so a default `+Asserts` build is not ABI compatible with a default `-Asserts` build. Clients that want ABI compatibility between `+Asserts` and `-Asserts` builds should use the CMake or autoconf build systems to set `LLVM_ENABLE_ABI_BREAKING_CHECKS` independently of `LLVM_ENABLE_ASSERTIONS`.

3.7.9 The Core LLVM Class Hierarchy Reference

```
#include "llvm/IR/Type.h"
```

header source: [Type.h](#)

doxygen info: [Type Clases](#)

The Core LLVM classes are the primary means of representing the program being inspected or transformed. The core LLVM classes are defined in header files in the `include/llvm/IR` directory, and implemented in the `lib/IR` directory. It's worth noting that, for historical reasons, this library is called `libLLVMCore.so`, not `libLLVMIR.so` as you might expect.

The Type class and Derived Types

`Type` is a superclass of all type classes. Every `Value` has a `Type`. `Type` cannot be instantiated directly but only through its subclasses. Certain primitive types (`VoidType`, `LabelType`, `FloatType` and `DoubleType`) have hidden subclasses. They are hidden because they offer no useful functionality beyond what the `Type` class offers except to distinguish themselves from other subclasses of `Type`.

All other types are subclasses of `DerivedType`. Types can be named, but this is not a requirement. There exists exactly one instance of a given shape at any one time. This allows type equality to be performed with address equality of the `Type` Instance. That is, given two `Type*` values, the types are identical if the pointers are identical.

Important Public Methods

- `bool isIntegerTy() const`: Returns true for any integer type.
- `bool isFloatingPointTy()`: Return true if this is one of the five floating point types.
- `bool isSized()`: Return true if the type has known size. Things that don't have a size are abstract types, labels and void.

Important Derived Types

IntegerType Subclass of `DerivedType` that represents integer types of any bit width. Any bit width between `IntegerType::MIN_INT_BITS` (1) and `IntegerType::MAX_INT_BITS` (~8 million) can be represented.

- `static const IntegerType* get(unsigned NumBits)`: get an integer type of a specific bit width.
- `unsigned getBitWidth() const`: Get the bit width of an integer type.

SequentialType This is subclassed by `ArrayType` and `VectorType`.

- `const Type * getElementTy() const`: Returns the type of each of the elements in the sequential type.
- `uint64_t getNumElements() const`: Returns the number of elements in the sequential type.

ArrayType This is a subclass of `SequentialType` and defines the interface for array types.

PointerType Subclass of `Type` for pointer types.

VectorType Subclass of `SequentialType` for vector types. A vector type is similar to an `ArrayType` but is distinguished because it is a first class type whereas `ArrayType` is not. Vector types are used for vector operations and are usually small vectors of an integer or floating point type.

StructType Subclass of `DerivedTypes` for struct types.

FunctionType Subclass of `DerivedTypes` for function types.

- `bool isVarArg() const`: Returns true if it's a vararg function.
- `const Type * getReturnType() const`: Returns the return type of the function.
- `const Type * getParamTy(unsigned i)`: Returns the type of the *i*th parameter.
- `const unsigned getNumParams() const`: Returns the number of formal parameters.

The Module class

```
#include "llvm/IR/Module.h"
```

header source: [Module.h](#)

doxygen info: [Module Class](#)

The `Module` class represents the top level structure present in LLVM programs. An LLVM module is effectively either a translation unit of the original program or a combination of several translation units merged by the linker. The `Module` class keeps track of a list of [Functions](#), a list of [GlobalVariables](#), and a [SymbolTable](#). Additionally, it contains a few helpful member functions that try to make common operations easy.

Important Public Members of the Module class

- `Module::Module(std::string name = "")`

Constructing a [Module](#) is easy. You can optionally provide a name for it (probably based on the name of the translation unit).

- `Module::iterator` - Typedef for function list iterator
`Module::const_iterator` - Typedef for const_iterator.
`begin()`, `end()`, `size()`, `empty()`

These are forwarding methods that make it easy to access the contents of a `Module` object's [Function](#) list.

- `Module::FunctionListType &getFunctionList()`

Returns the list of [Functions](#). This is necessary to use when you need to update the list or perform a complex action that doesn't have a forwarding method.

- `Module::global_iterator` - Typedef for global variable list iterator
`Module::const_global_iterator` - Typedef for const_iterator.
`global_begin()`, `global_end()`, `global_size()`, `global_empty()`

These are forwarding methods that make it easy to access the contents of a `Module` object's [GlobalVariable](#) list.

- `Module::GlobalListType &getGlobalList()`

Returns the list of [GlobalVariables](#). This is necessary to use when you need to update the list or perform a complex action that doesn't have a forwarding method.

- `SymbolTable *getSymbolTable()`

Return a reference to the [SymbolTable](#) for this `Module`.

- `Function *getFunction(StringRef Name) const`

Look up the specified function in the `Module` [SymbolTable](#). If it does not exist, return null.

- `Function *getOrInsertFunction(const std::string &Name, const FunctionType *T)`

Look up the specified function in the `Module` [SymbolTable](#). If it does not exist, add an external declaration for the function and return it.

- `std::string getTypeName(const Type *Ty)`

If there is at least one entry in the *SymbolTable* for the specified *Type*, return it. Otherwise return the empty string.

- `bool addTypeName(const std::string &Name, const Type *Ty)`

Insert an entry in the *SymbolTable* mapping *Name* to *Ty*. If there is already an entry for this name, true is returned and the *SymbolTable* is not modified.

The Value class

#include "llvm/IR/Value.h"

header source: [Value.h](#)

doxygen info: [Value Class](#)

The *Value* class is the most important class in the LLVM Source base. It represents a typed value that may be used (among other things) as an operand to an instruction. There are many different types of *Values*, such as *Constants*, *Arguments*. Even *Instructions* and *Functions* are *Values*.

A particular *Value* may be used many times in the LLVM representation for a program. For example, an incoming argument to a function (represented with an instance of the *Argument* class) is “used” by every instruction in the function that references the argument. To keep track of this relationship, the *Value* class keeps a list of all of the *Users* that is using it (the *User* class is a base class for all nodes in the LLVM graph that can refer to *Values*). This use list is how LLVM represents def-use information in the program, and is accessible through the *use_** methods, shown below.

Because LLVM is a typed representation, every LLVM *Value* is typed, and this *Type* is available through the *getType()* method. In addition, all LLVM values can be named. The “name” of the *Value* is a symbolic string printed in the LLVM code:

```
%foo = add i32 1, 2
```

The name of this instruction is “foo”. **NOTE** that the name of any value may be missing (an empty string), so names should **ONLY** be used for debugging (making the source code easier to read, debugging printouts), they should not be used to keep track of values or map between them. For this purpose, use a `std::map` of pointers to the *Value* itself instead.

One important aspect of LLVM is that there is no distinction between an SSA variable and the operation that produces it. Because of this, any reference to the value produced by an instruction (or the value available as an incoming argument, for example) is represented as a direct pointer to the instance of the class that represents this value. Although this may take some getting used to, it simplifies the representation and makes it easier to manipulate.

Important Public Members of the Value class

- *Value::use_iterator* - Typedef for iterator over the use-list
- Value::const_use_iterator* - Typedef for const_iterator over the use-list
- unsigned use_size()* - Returns the number of users of the value.
- bool use_empty()* - Returns true if there are no users.
- use_iterator use_begin()* - Get an iterator to the start of the use-list.
- use_iterator use_end()* - Get an iterator to the end of the use-list.
- User *use_back()* - Returns the last element in the list.

These methods are the interface to access the def-use information in LLVM. As with all other iterators in LLVM, the naming conventions follow the conventions defined by the *STL*.

- `Type *getType() const` This method returns the Type of the Value.
- `bool hasName() const`
`std::string getName() const`
`void setName(const std::string &Name)`

This family of methods is used to access and assign a name to a Value, be aware of the *precaution above*.

- `void replaceAllUsesWith(Value *V)`

This method traverses the use list of a Value changing all *Users* of the current value to refer to “V” instead. For example, if you detect that an instruction always produces a constant value (for example through constant folding), you can replace all uses of the instruction with the constant like this:

```
Inst->replaceAllUsesWith(ConstVal);
```

The User class

```
#include "llvm/IR/User.h"
```

header source: [User.h](#)

doxygen info: [User Class](#)

Superclass: [Value](#)

The `User` class is the common base class of all LLVM nodes that may refer to Values. It exposes a list of “Operands” that are all of the Values that the User is referring to. The `User` class itself is a subclass of `Value`.

The operands of a `User` point directly to the LLVM `Value` that it refers to. Because LLVM uses Static Single Assignment (SSA) form, there can only be one definition referred to, allowing this direct connection. This connection provides the use-def information in LLVM.

Important Public Members of the User class

The `User` class exposes the operand list in two ways: through an index access interface and through an iterator based interface.

- `Value *getOperand(unsigned i)`
`unsigned getNumOperands()`

These two methods expose the operands of the `User` in a convenient form for direct access.

- `User::op_iterator` - Typedef for iterator over the operand list
`op_iterator op_begin()` - Get an iterator to the start of the operand list.
`op_iterator op_end()` - Get an iterator to the end of the operand list.

Together, these methods make up the iterator based interface to the operands of a `User`.

The Instruction class

```
#include "llvm/IR/Instruction.h"
```

header source: [Instruction.h](#)

doxygen info: [Instruction Class](#)

Superclasses: [User](#), [Value](#)

The `Instruction` class is the common base class for all LLVM instructions. It provides only a few methods, but is a very commonly used class. The primary data tracked by the `Instruction` class itself is the opcode (instruction type) and the parent *BasicBlock* the `Instruction` is embedded into. To represent a specific type of instruction, one of many subclasses of `Instruction` are used.

Because the `Instruction` class subclasses the *User* class, its operands can be accessed in the same way as for other Users (with the `getOperand()`/`getNumOperands()` and `op_begin()`/`op_end()` methods). An important file for the `Instruction` class is the `llvm/Instruction.def` file. This file contains some meta-data about the various different types of instructions in LLVM. It describes the enum values that are used as opcodes (for example `Instruction::Add` and `Instruction::ICmp`), as well as the concrete sub-classes of `Instruction` that implement the instruction (for example *BinaryOperator* and *CmpInst*). Unfortunately, the use of macros in this file confuses doxygen, so these enum values don't show up correctly in the [doxygen output](#).

Important Subclasses of the `Instruction` class

- `BinaryOperator`

This subclasses represents all two operand instructions whose operands must be the same type, except for the comparison instructions.

- `CastInst` This subclass is the parent of the 12 casting instructions. It provides common operations on cast instructions.

- `CmpInst`

This subclass represents the two comparison instructions, *ICmpInst* (integer operands), and *FCmpInst* (floating point operands).

- `TerminatorInst`

This subclass is the parent of all terminator instructions (those which can terminate a block).

Important Public Members of the `Instruction` class

- `BasicBlock *getParent()`

Returns the *BasicBlock* that this `Instruction` is embedded into.

- `bool mayWriteToMemory()`

Returns true if the instruction writes to memory, i.e. it is a call, free, invoke, or store.

- `unsigned getOpcode()`

Returns the opcode for the `Instruction`.

- `Instruction *clone() const`

Returns another instance of the specified instruction, identical in all ways to the original except that the instruction has no parent (i.e. it's not embedded into a *BasicBlock*), and it has no name.

The `Constant` class and subclasses

`Constant` represents a base class for different types of constants. It is subclassed by `ConstantInt`, `ConstantArray`, etc. for representing the various types of Constants. *GlobalValue* is also a subclass, which represents the address of a global variable or function.

Important Subclasses of Constant

- **ConstantInt** : This subclass of Constant represents an integer constant of any width.
 - `const APInt& getValue() const`: Returns the underlying value of this constant, an APInt value.
 - `int64_t getSExtValue() const`: Converts the underlying APInt value to an int64_t via sign extension. If the value (not the bit width) of the APInt is too large to fit in an int64_t, an assertion will result. For this reason, use of this method is discouraged.
 - `uint64_t getZExtValue() const`: Converts the underlying APInt value to a uint64_t via zero extension. If the value (not the bit width) of the APInt is too large to fit in a uint64_t, an assertion will result. For this reason, use of this method is discouraged.
 - `static ConstantInt* get(const APInt& Val)`: Returns the ConstantInt object that represents the value provided by Val. The type is implied as the IntegerType that corresponds to the bit width of Val.
 - `static ConstantInt* get(const Type *Ty, uint64_t Val)`: Returns the ConstantInt object that represents the value provided by Val for integer type Ty.
- **ConstantFP** : This class represents a floating point constant.
 - `double getValue() const`: Returns the underlying value of this constant.
- **ConstantArray** : This represents a constant array.
 - `const std::vector<Use> &getValues() const`: Returns a vector of component constants that makeup this array.
- **ConstantStruct** : This represents a constant struct.
 - `const std::vector<Use> &getValues() const`: Returns a vector of component constants that makeup this array.
- **GlobalValue** : This represents either a global variable or a function. In either case, the value is a constant fixed address (after linking).

The GlobalValue class

#include "llvm/IR/GlobalValue.h"

header source: [GlobalValue.h](#)

doxygen info: [GlobalValue Class](#)

Superclasses: [Constant](#), [User](#), [Value](#)

Global values ([GlobalVariables](#) or [Functions](#)) are the only LLVM values that are visible in the bodies of all [Functions](#). Because they are visible at global scope, they are also subject to linking with other globals defined in different translation units. To control the linking process, GlobalValues know their linkage rules. Specifically, GlobalValues know whether they have internal or external linkage, as defined by the `LinkageTypes` enumeration.

If a GlobalValue has internal linkage (equivalent to being `static` in C), it is not visible to code outside the current translation unit, and does not participate in linking. If it has external linkage, it is visible to external code, and does participate in linking. In addition to linkage information, GlobalValues keep track of which [Module](#) they are currently part of.

Because GlobalValues are memory objects, they are always referred to by their **address**. As such, the [Type](#) of a global is always a pointer to its contents. It is important to remember this when using the `GetElementPtrInst` instruction because this pointer must be dereferenced first. For example, if you have a [GlobalVariable](#) (a subclass of GlobalValue) that is an array of 24 ints, type `[24 x i32]`, then the GlobalVariable is a pointer to that

array. Although the address of the first element of this array and the value of the `GlobalVariable` are the same, they have different types. The `GlobalVariable`'s type is `[24 x i32]`. The first element's type is `i32`. Because of this, accessing a global value requires you to dereference the pointer with `GetElementPtrInst` first, then its elements can be accessed. This is explained in the [LLVM Language Reference Manual](#).

Important Public Members of the `GlobalValue` class

- `bool hasInternalLinkage() const`
`bool hasExternalLinkage() const`
`void setInternalLinkage(bool HasInternalLinkage)`
These methods manipulate the linkage characteristics of the `GlobalValue`.
- `Module *getParent()`
This returns the *Module* that the `GlobalValue` is currently embedded into.

The `Function` class

`#include "llvm/IR/Function.h"`

header source: [Function.h](#)

doxygen info: [Function Class](#)

Superclasses: [GlobalValue](#), [Constant](#), [User](#), [Value](#)

The `Function` class represents a single procedure in LLVM. It is actually one of the more complex classes in the LLVM hierarchy because it must keep track of a large amount of data. The `Function` class keeps track of a list of *BasicBlocks*, a list of formal *Arguments*, and a *SymbolTable*.

The list of *BasicBlocks* is the most commonly used part of `Function` objects. The list imposes an implicit ordering of the blocks in the function, which indicate how the code will be laid out by the backend. Additionally, the first *BasicBlock* is the implicit entry node for the `Function`. It is not legal in LLVM to explicitly branch to this initial block. There are no implicit exit nodes, and in fact there may be multiple exit nodes from a single `Function`. If the *BasicBlock* list is empty, this indicates that the `Function` is actually a function declaration: the actual body of the function hasn't been linked in yet.

In addition to a list of *BasicBlocks*, the `Function` class also keeps track of the list of formal *Arguments* that the function receives. This container manages the lifetime of the *Argument* nodes, just like the *BasicBlock* list does for the *BasicBlocks*.

The *SymbolTable* is a very rarely used LLVM feature that is only used when you have to look up a value by name. Aside from that, the *SymbolTable* is used internally to make sure that there are not conflicts between the names of *Instructions*, *BasicBlocks*, or *Arguments* in the function body.

Note that `Function` is a *GlobalValue* and therefore also a *Constant*. The value of the function is its address (after linking) which is guaranteed to be constant.

Important Public Members of the `Function`

- `Function(const FunctionType *Ty, LinkageTypes Linkage, const std::string &N = "", Module* Parent = 0)`

Constructor used when you need to create new `Functions` to add the program. The constructor must specify the type of the function to create and what type of linkage the function should have. The *FunctionType* argument specifies the formal arguments and return value for the function. The same *FunctionType* value can be used to

create multiple functions. The `Parent` argument specifies the Module in which the function is defined. If this argument is provided, the function will automatically be inserted into that module’s list of functions.

- `bool isDeclaration()`

Return whether or not the `Function` has a body defined. If the function is “external”, it does not have a body, and thus must be resolved by linking with a function defined in a different translation unit.

- `Function::iterator` - Typedef for basic block list iterator
`Function::const_iterator` - Typedef for `const_iterator`.
`begin()`, `end()`, `size()`, `empty()`

These are forwarding methods that make it easy to access the contents of a `Function` object’s *BasicBlock* list.

- `Function::BasicBlockListType &getBasicBlockList()`

Returns the list of *BasicBlocks*. This is necessary to use when you need to update the list or perform a complex action that doesn’t have a forwarding method.

- `Function::arg_iterator` - Typedef for the argument list iterator
`Function::const_arg_iterator` - Typedef for `const_iterator`.
`arg_begin()`, `arg_end()`, `arg_size()`, `arg_empty()`

These are forwarding methods that make it easy to access the contents of a `Function` object’s *Argument* list.

- `Function::ArgumentListType &getArgumentList()`

Returns the list of *Argument*. This is necessary to use when you need to update the list or perform a complex action that doesn’t have a forwarding method.

- `BasicBlock &getEntryBlock()`

Returns the entry `BasicBlock` for the function. Because the entry block for the function is always the first block, this returns the first block of the `Function`.

- `Type *getReturnType()`
`FunctionType *getFunctionType()`

This traverses the *Type* of the `Function` and returns the return type of the function, or the *FunctionType* of the actual function.

- `SymbolTable *getSymbolTable()`

Return a pointer to the *SymbolTable* for this `Function`.

The GlobalVariable class

```
#include "llvm/IR/GlobalVariable.h"
```

header source: [GlobalVariable.h](#)

doxygen info: [GlobalVariable Class](#)

Superclasses: [GlobalValue](#), [Constant](#), [User](#), [Value](#)

Global variables are represented with the (surprise surprise) `GlobalVariable` class. Like functions, `GlobalVariables` are also subclasses of *GlobalValue*, and as such are always referenced by their address (global values must live in memory, so their “name” refers to their constant address). See *GlobalValue* for more on this. Global variables may have an initial value (which must be a *Constant*), and if they have an initializer, they may be marked as “constant” themselves (indicating that their contents never change at runtime).

Important Public Members of the `GlobalVariable` class

- `GlobalVariable(const Type *Ty, bool isConstant, LinkageTypes &Linkage, Constant *Initializer = 0, const std::string &Name = "", Module* Parent = 0)`

Create a new global variable of the specified type. If `isConstant` is true then the global variable will be marked as unchanging for the program. The `Linkage` parameter specifies the type of linkage (internal, external, weak, linkonce, appending) for the variable. If the linkage is `InternalLinkage`, `WeakAnyLinkage`, `WeakODRLinkage`, `LinkOnceAnyLinkage` or `LinkOnceODRLinkage`, then the resultant global variable will have internal linkage. `AppendingLinkage` concatenates together all instances (in different translation units) of the variable into a single variable but is only applicable to arrays. See the [LLVM Language Reference](#) for further details on linkage types. Optionally an initializer, a name, and the module to put the variable into may be specified for the global variable as well.

- `bool isConstant() const`
Returns true if this is a global variable that is known not to be modified at runtime.
- `bool hasInitializer()`
Returns true if this `GlobalVariable` has an initializer.
- `Constant *getInitializer()`
Returns the initial value for a `GlobalVariable`. It is not legal to call this method if there is no initializer.

The `BasicBlock` class

`#include "llvm/IR/BasicBlock.h"`

header source: [BasicBlock.h](#)

doxygen info: [BasicBlock Class](#)

Superclass: [Value](#)

This class represents a single entry single exit section of the code, commonly known as a basic block by the compiler community. The `BasicBlock` class maintains a list of [Instructions](#), which form the body of the block. Matching the language definition, the last element of this list of instructions is always a terminator instruction (a subclass of the [TerminatorInst](#) class).

In addition to tracking the list of instructions that make up the block, the `BasicBlock` class also keeps track of the [Function](#) that it is embedded into.

Note that `BasicBlocks` themselves are [Values](#), because they are referenced by instructions like branches and can go in the switch tables. `BasicBlocks` have type `label`.

Important Public Members of the `BasicBlock` class

- `BasicBlock(const std::string &Name = "", Function *Parent = 0)`

The `BasicBlock` constructor is used to create new basic blocks for insertion into a function. The constructor optionally takes a name for the new block, and a [Function](#) to insert it into. If the `Parent` parameter is specified, the new `BasicBlock` is automatically inserted at the end of the specified [Function](#), if not specified, the `BasicBlock` must be manually inserted into the [Function](#).

- `BasicBlock::iterator` - Typedef for instruction list iterator
`BasicBlock::const_iterator` - Typedef for const_iterator.

`begin()`, `end()`, `front()`, `back()`, `size()`, `empty()` STL-style functions for accessing the instruction list.

These methods and typedefs are forwarding functions that have the same semantics as the standard library methods of the same names. These methods expose the underlying instruction list of a basic block in a way that is easy to manipulate. To get the full complement of container operations (including operations to update the list), you must use the `getInstList()` method.

- `BasicBlock::InstListType &getInstList()`

This method is used to get access to the underlying container that actually holds the Instructions. This method must be used when there isn't a forwarding function in the `BasicBlock` class for the operation that you would like to perform. Because there are no forwarding functions for “updating” operations, you need to use this if you want to update the contents of a `BasicBlock`.

- `Function *getParent()`

Returns a pointer to [Function](#) the block is embedded into, or a null pointer if it is homeless.

- `TerminatorInst *getTerminator()`

Returns a pointer to the terminator instruction that appears at the end of the `BasicBlock`. If there is no terminator instruction, or if the last instruction in the block is not a terminator, then a null pointer is returned.

The Argument class

This subclass of `Value` defines the interface for incoming formal arguments to a function. A `Function` maintains a list of its formal arguments. An argument has a pointer to the parent `Function`.

3.8 LLVM Extensions

- *Introduction*
- *General Assembly Syntax*
 - *C99-style Hexadecimal Floating-point Constants*
- *Machine-specific Assembly Syntax*
 - *X86/COFF-Dependent*
 - * *Relocations*
 - * *.linkonce Directive*
 - * *.section Directive*
 - *ELF-Dependent*
 - * *.section Directive*
- *Target Specific Behaviour*
 - *X86*
 - * *Relocations*
 - *Windows on ARM*
 - * *Stack Probe Emission*

* *Variable Length Arrays*

3.8.1 Introduction

This document describes extensions to tools and formats LLVM seeks compatibility with.

3.8.2 General Assembly Syntax

C99-style Hexadecimal Floating-point Constants

LLVM's assemblers allow floating-point constants to be written in C99's hexadecimal format instead of decimal if desired.

```
.section .data
.float 0x1c2.2ap3
```

3.8.3 Machine-specific Assembly Syntax

X86/COFF-Dependent

Relocations

The following additional relocation types are supported:

@IMGREL (AT&T syntax only) generates an image-relative relocation that corresponds to the COFF relocation types IMAGE_REL_I386_DIR32NB (32-bit) or IMAGE_REL_AMD64_ADDR32NB (64-bit).

```
.text
fun:
    mov foo@IMGREL(%ebx, %ecx, 4), %eax

.section .pdata
    .long fun@IMGREL
    .long (fun@imgrel + 0x3F)
    .long $unwind$fun@imgrel
```

.secrel32 generates a relocation that corresponds to the COFF relocation types IMAGE_REL_I386_SECREL (32-bit) or IMAGE_REL_AMD64_SECREL (64-bit).

.secidx relocation generates an index of the section that contains the target. It corresponds to the COFF relocation types IMAGE_REL_I386_SECTION (32-bit) or IMAGE_REL_AMD64_SECTION (64-bit).

```
.section .debug$S, "rn"
    .long 4
    .long 242
    .long 40
    .secrel32 __function_name + 0
    .secidx  __function_name
    ...
```

.linkonce Directive

Syntax:

```
.linkonce [ comdat type ]
```

Supported COMDAT types:

discard Discards duplicate sections with the same COMDAT symbol. This is the default if no type is specified.

one_only If the symbol is defined multiple times, the linker issues an error.

same_size Duplicates are discarded, but the linker issues an error if any have different sizes.

same_contents Duplicates are discarded, but the linker issues an error if any duplicates do not have exactly the same content.

largest Links the largest section from among the duplicates.

newest Links the newest section from among the duplicates.

```
.section .text$foo
.linkonce
...
```

.section Directive

MC supports passing the information in `.linkonce` at the end of `.section`. For example, these two codes are equivalent

```
.section secName, "dr", discard, "Symbol1"
.globl Symbol1
Symbol1:
.long 1
```

```
.section secName, "dr"
.linkonce discard
.globl Symbol1
Symbol1:
.long 1
```

Note that in the combined form the COMDAT symbol is explicit. This extension exists to support multiple sections with the same name in different COMDATs:

```
.section secName, "dr", discard, "Symbol1"
.globl Symbol1
Symbol1:
.long 1

.section secName, "dr", discard, "Symbol2"
.globl Symbol2
Symbol2:
.long 1
```

In addition to the types allowed with `.linkonce`, `.section` also accepts `associative`. The meaning is that the section is linked if a certain other COMDAT section is linked. This other section is indicated by the `comdat` symbol in this directive. It can be any symbol defined in the associated section, but is usually the associated section's `comdat`.

The following restrictions apply to the associated section:

1. It must be a COMDAT section.
2. It cannot be another associative COMDAT section.

In the following example the symbol `sym` is the comdat symbol of `.foo` and `.bar` is associated to `.foo`.

```
.section      .foo,"bw",discard, "sym"  
.section      .bar,"rd",associative, "sym"
```

MC supports these flags in the COFF `.section` directive:

- `b`: BSS section (`IMAGE_SCN_CNT_INITIALIZED_DATA`)
- `d`: Data section (`IMAGE_SCN_CNT_UNINITIALIZED_DATA`)
- `n`: Section is not loaded (`IMAGE_SCN_LNK_REMOVE`)
- `r`: Read-only
- `s`: Shared section
- `w`: Writable
- `x`: Executable section
- `y`: Not readable
- `D`: Discardable (`IMAGE_SCN_MEM_DISCARDABLE`)

These flags are all compatible with gas, with the exception of the `D` flag, which `gnu as` does not support. For gas compatibility, sections with a name starting with `“.debug”` are implicitly discardable.

ELF-Dependent

`.section` Directive

In order to support creating multiple sections with the same name and comdat, it is possible to add an unique number at the end of the `.section` directive. For example, the following code creates two sections named `.text`.

```
.section      .text,"ax",@progbits,unique,1  
nop  
  
.section      .text,"ax",@progbits,unique,2  
nop
```

The unique number is not present in the resulting object at all. It is just used in the assembler to differentiate the sections.

The `‘o’` flag is mapped to `SHF_LINK_ORDER`. If it is present, a symbol must be given that identifies the section to be placed is the `.sh_link`.

```
.section .foo,"a",@progbits  
.Ltmp:  
.section .bar,"ao",@progbits,.Ltmp
```

which is equivalent to just

```
.section .foo,"a",@progbits  
.section .bar,"ao",@progbits,.foo
```

3.8.4 Target Specific Behaviour

X86

Relocations

@ABS8 can be applied to symbols which appear as immediate operands to instructions that have an 8-bit immediate form for that operand. It causes the assembler to use the 8-bit form and an 8-bit relocation (e.g. R_386_8 or R_X86_64_8) for the symbol.

For example:

```
cmpq $foo@ABS8, %rdi
```

This causes the assembler to select the form of the 64-bit `cmpq` instruction that takes an 8-bit immediate operand that is sign extended to 64 bits, as opposed to `cmpq $foo, %rdi` which takes a 32-bit immediate operand. This is also not the same as `cmpb $foo, %dil`, which is an 8-bit comparison.

Windows on ARM

Stack Probe Emission

The reference implementation (Microsoft Visual Studio 2012) emits stack probes in the following fashion:

```
movw r4, #constant
bl __chkstk
sub.w sp, sp, r4
```

However, this has the limitation of 32 MiB (± 16 MiB). In order to accommodate larger binaries, LLVM supports the use of `-mcode-model=large` to allow a 4GiB range via a slight deviation. It will generate an indirect jump as follows:

```
movw r4, #constant
movw r12, :lower16:__chkstk
movt r12, :upper16:__chkstk
blx r12
sub.w sp, sp, r4
```

Variable Length Arrays

The reference implementation (Microsoft Visual Studio 2012) does not permit the emission of Variable Length Arrays (VLAs).

The Windows ARM Itanium ABI extends the base ABI by adding support for emitting a dynamic stack allocation. When emitting a variable stack allocation, a call to `__chkstk` is emitted unconditionally to ensure that guard pages are setup properly. The emission of this stack probe emission is handled similar to the standard stack probe emission.

The MSVC environment does not emit code for VLAs currently.

3.9 libFuzzer – a library for coverage-guided fuzz testing.

- *Introduction*
- *Versions*
- *Getting Started*
- *Options*
- *Output*
- *Examples*
- *Advanced features*
- *Developing libFuzzer*
- *Fuzzing components of LLVM*
- *FAQ*
- *Trophies*

3.9.1 Introduction

LibFuzzer is in-process, coverage-guided, evolutionary fuzzing engine.

LibFuzzer is linked with the library under test, and feeds fuzzed inputs to the library via a specific fuzzing entrypoint (aka “target function”); the fuzzer then tracks which areas of the code are reached, and generates mutations on the corpus of input data in order to maximize the code coverage. The code coverage information for libFuzzer is provided by LLVM’s [SanitizerCoverage](#) instrumentation.

Contact: [libfuzzer\(#\)googlegroups.com](mailto:libfuzzer@googlegroups.com)

3.9.2 Versions

LibFuzzer is under active development so you will need the current (or at least a very recent) version of the Clang compiler.

(If [building Clang from trunk](#) is too time-consuming or difficult, then the Clang binaries that the Chromium developers build are likely to be fairly recent:

```
mkdir TMP_CLANG
cd TMP_CLANG
git clone https://chromium.googlesource.com/chromium/src/tools/clang
cd ..
TMP_CLANG/clang/scripts/update.py
```

This installs the Clang binary as `./third_party/llvm-build/Release+Asserts/bin/clang`)

The libFuzzer code resides in the LLVM repository, and requires a recent Clang compiler to build (and is used to *fuzz various parts of LLVM itself*). However the fuzzer itself does not (and should not) depend on any part of LLVM infrastructure and can be used for other projects without requiring the rest of LLVM.

3.9.3 Getting Started

- *Fuzz Target*
- *Fuzzer Usage*
- *Corpus*
- *Running*
- *Parallel Fuzzing*

Fuzz Target

The first step in using libFuzzer on a library is to implement a *fuzz target* – a function that accepts an array of bytes and does something interesting with these bytes using the API under test. Like this:

```
// fuzz_target.cc
extern "C" int LLVMFuzzerTestOneInput (const uint8_t *Data, size_t Size) {
    DoSomethingInterestingWithMyAPI (Data, Size);
    return 0; // Non-zero return values are reserved for future use.
}
```

Note that this fuzz target does not depend on libFuzzer in any way and so it is possible and even desirable to use it with other fuzzing engines e.g. [AFL](#) and/or [Radamsa](#).

Some important things to remember about fuzz targets:

- The fuzzing engine will execute the fuzz target many times with different inputs in the same process.
- It must tolerate any kind of input (empty, huge, malformed, etc).
- It must not *exit()* on any input.
- It may use threads but ideally all threads should be joined at the end of the function.
- It must be as deterministic as possible. Non-determinism (e.g. random decisions not based on the input bytes) will make fuzzing inefficient.
- It must be fast. Try avoiding cubic or greater complexity, logging, or excessive memory consumption.
- Ideally, it should not modify any global state (although that's not strict).
- Usually, the narrower the target the better. E.g. if your target can parse several data formats, split it into several targets, one per format.

Fuzzer Usage

Very recent versions of Clang (> April 20 2017) include libFuzzer, and no installation is necessary. In order to fuzz your binary, use the `-fsanitize=fuzzer` flag during the compilation:

```
clang -fsanitize=fuzzer,address mytarget.c
```

Otherwise, build the libFuzzer library as a static archive, without any sanitizer options. Note that the libFuzzer library contains the `main()` function:

```
svn co http://llvm.org/svn/llvm-project/llvm/trunk/lib/Fuzzer # or git clone https://
→chromium.googlesource.com/chromium/llvm-project/llvm/lib/Fuzzer
./Fuzzer/build.sh # Produces libFuzzer.a
```

Then build the fuzzing target function and the library under test using the `SanitizerCoverage` option, which instruments the code so that the fuzzer can retrieve code coverage information (to guide the fuzzing). Linking with the `libFuzzer` code then gives a fuzzer executable.

You should also enable one or more of the *sanitizers*, which help to expose latent bugs by making incorrect behavior generate errors at runtime:

- `AddressSanitizer` (ASAN) detects memory access errors. Use `-fsanitize=address`.
- `UndefinedBehaviorSanitizer` (UBSAN) detects the use of various features of C/C++ that are explicitly listed as resulting in undefined behavior. Use `-fsanitize=undefined -fno-sanitize-recover=undefined` or any individual UBSAN check, e.g. `-fsanitize=signed-integer-overflow -fno-sanitize-recover=undefined`. You may combine ASAN and UBSAN in one build.
- `MemorySanitizer` (MSAN) detects uninitialized reads: code whose behavior relies on memory contents that have not been initialized to a specific value. Use `-fsanitize=memory`. MSAN can not be combined with other sanitizers and should be used as a separate build.

Finally, link with `libFuzzer.a`:

```
clang -fsanitize-coverage=trace-pc-guard -fsanitize=address your_lib.cc fuzz_target.  
↪cc libFuzzer.a -o my_fuzzer
```

Corpus

Coverage-guided fuzzers like `libFuzzer` rely on a corpus of sample inputs for the code under test. This corpus should ideally be seeded with a varied collection of valid and invalid inputs for the code under test; for example, for a graphics library the initial corpus might hold a variety of different small PNG/JPG/GIF files. The fuzzer generates random mutations based around the sample inputs in the current corpus. If a mutation triggers execution of a previously-uncovered path in the code under test, then that mutation is saved to the corpus for future variations.

`LibFuzzer` will work without any initial seeds, but will be less efficient if the library under test accepts complex, structured inputs.

The corpus can also act as a sanity/regression check, to confirm that the fuzzing entrypoint still works and that all of the sample inputs run through the code under test without problems.

If you have a large corpus (either generated by fuzzing or acquired by other means) you may want to minimize it while still preserving the full coverage. One way to do that is to use the `-merge=1` flag:

```
mkdir NEW_CORPUS_DIR # Store minimized corpus here.  
./my_fuzzer -merge=1 NEW_CORPUS_DIR FULL_CORPUS_DIR
```

You may use the same flag to add more interesting items to an existing corpus. Only the inputs that trigger new coverage will be added to the first corpus.

```
./my_fuzzer -merge=1 CURRENT_CORPUS_DIR NEW_POTENTIALLY_INTERESTING_INPUTS_DIR
```

Running

To run the fuzzer, first create a *Corpus* directory that holds the initial “seed” sample inputs:

```
mkdir CORPUS_DIR  
cp /some/input/samples/* CORPUS_DIR
```

Then run the fuzzer on the corpus directory:

```
./my_fuzzer CORPUS_DIR # -max_len=1000 -jobs=20 ...
```

As the fuzzer discovers new interesting test cases (i.e. test cases that trigger coverage of new paths through the code under test), those test cases will be added to the corpus directory.

By default, the fuzzing process will continue indefinitely – at least until a bug is found. Any crashes or sanitizer failures will be reported as usual, stopping the fuzzing process, and the particular input that triggered the bug will be written to disk (typically as `crash-<sha1>`, `leak-<sha1>`, or `timeout-<sha1>`).

Parallel Fuzzing

Each libFuzzer process is single-threaded, unless the library under test starts its own threads. However, it is possible to run multiple libFuzzer processes in parallel with a shared corpus directory; this has the advantage that any new inputs found by one fuzzer process will be available to the other fuzzer processes (unless you disable this with the `-reload=0` option).

This is primarily controlled by the `-jobs=N` option, which indicates that that N fuzzing jobs should be run to completion (i.e. until a bug is found or time/iteration limits are reached). These jobs will be run across a set of worker processes, by default using half of the available CPU cores; the count of worker processes can be overridden by the `-workers=N` option. For example, running with `-jobs=30` on a 12-core machine would run 6 workers by default, with each worker averaging 5 bugs by completion of the entire process.

3.9.4 Options

To run the fuzzer, pass zero or more corpus directories as command line arguments. The fuzzer will read test inputs from each of these corpus directories, and any new test inputs that are generated will be written back to the first corpus directory:

```
./fuzzer [-flag1=val1 [-flag2=val2 ...] ] [dir1 [dir2 ...] ]
```

If a list of files (rather than directories) are passed to the fuzzer program, then it will re-run those files as test inputs but will not perform any fuzzing. In this mode the fuzzer binary can be used as a regression test (e.g. on a continuous integration system) to check the target function and saved inputs still work.

The most important command line options are:

- help** Print help message.
- seed** Random seed. If 0 (the default), the seed is generated.
- runs** Number of individual test runs, -1 (the default) to run indefinitely.
- max_len** Maximum length of a test input. If 0 (the default), libFuzzer tries to guess a good value based on the corpus (and reports it).
- timeout** Timeout in seconds, default 1200. If an input takes longer than this timeout, the process is treated as a failure case.
- rss_limit_mb** Memory usage limit in Mb, default 2048. Use 0 to disable the limit. If an input requires more than this amount of RSS memory to execute, the process is treated as a failure case. The limit is checked in a separate thread every second. If running w/o ASAN/MSAN, you may use ‘ulimit -v’ instead.
- timeout_exitcode** Exit code (default 77) used if libFuzzer reports a timeout.
- error_exitcode** Exit code (default 77) used if libFuzzer itself (not a sanitizer) reports a bug (leak, OOM, etc).
- max_total_time** If positive, indicates the maximum total time in seconds to run the fuzzer. If 0 (the default), run indefinitely.

- merge** If set to 1, any corpus inputs from the 2nd, 3rd etc. corpus directories that trigger new code coverage will be merged into the first corpus directory. Defaults to 0. This flag can be used to minimize a corpus.
- minimize_crash** If 1, minimizes the provided crash input. Use with `-runs=N` or `-max_total_time=N` to limit the number of attempts.
- reload** If set to 1 (the default), the corpus directory is re-read periodically to check for new inputs; this allows detection of new inputs that were discovered by other fuzzing processes.
- jobs** Number of fuzzing jobs to run to completion. Default value is 0, which runs a single fuzzing process until completion. If the value is ≥ 1 , then this number of jobs performing fuzzing are run, in a collection of parallel separate worker processes; each such worker process has its `stdout/stderr` redirected to `fuzz-<JOB>.log`.
- workers** Number of simultaneous worker processes to run the fuzzing jobs to completion in. If 0 (the default), `min(jobs, NumberOfCpuCores() / 2)` is used.
- dict** Provide a dictionary of input keywords; see [Dictionaries](#).
- use_counters** Use [coverage counters](#) to generate approximate counts of how often code blocks are hit; defaults to 1.
- use_value_profile** Use [value profile](#) to guide corpus expansion; defaults to 0.
- only_ascii** If 1, generate only ASCII (`isprint`+`isspace`) inputs. Defaults to 0.
- artifact_prefix** Provide a prefix to use when saving fuzzing artifacts (crash, timeout, or slow inputs) as `$(artifact_prefix)file`. Defaults to empty.
- exact_artifact_path** Ignored if empty (the default). If non-empty, write the single artifact on failure (crash, timeout) as `$(exact_artifact_path)`. This overrides `-artifact_prefix` and will not use checksum in the file name. Do not use the same path for several parallel processes.
- print_pcs** If 1, print out newly covered PCs. Defaults to 0.
- print_final_stats** If 1, print statistics at exit. Defaults to 0.
- detect_leaks** If 1 (default) and if LeakSanitizer is enabled try to detect memory leaks during fuzzing (i.e. not only at shut down).
- close_fd_mask** Indicate output streams to close at startup. Be careful, this will remove diagnostic output from target code (e.g. messages on assert failure).
- 0 (default): close neither `stdout` nor `stderr`
 - 1: close `stdout`
 - 2: close `stderr`
 - 3: close both `stdout` and `stderr`.
- print_coverage** If 1, print coverage information as text at exit.
- dump_coverage** If 1, dump coverage information as a `.sancov` file at exit.

For the full list of flags run the fuzzer binary with `-help=1`.

3.9.5 Output

During operation the fuzzer prints information to `stderr`, for example:

```

INFO: Seed: 1523017872
INFO: Loaded 1 modules (16 guards): [0x744e60, 0x744ea0),
INFO: -max_len is not provided, using 64
INFO: A corpus is not provided, starting from an empty corpus
#0   READ units: 1
#1   INITED cov: 3 ft: 2 corp: 1/1b exec/s: 0 rss: 24Mb
#3811 NEW    cov: 4 ft: 3 corp: 2/2b exec/s: 0 rss: 25Mb L: 1 MS: 5 ChangeBit-
    ↪ChangeByte-ChangeBit-ShuffleBytes-ChangeByte-
#3827 NEW    cov: 5 ft: 4 corp: 3/4b exec/s: 0 rss: 25Mb L: 2 MS: 1 CopyPart-
#3963 NEW    cov: 6 ft: 5 corp: 4/6b exec/s: 0 rss: 25Mb L: 2 MS: 2 ShuffleBytes-
    ↪ChangeBit-
#4167 NEW    cov: 7 ft: 6 corp: 5/9b exec/s: 0 rss: 25Mb L: 3 MS: 1 InsertByte-
...

```

The early parts of the output include information about the fuzzer options and configuration, including the current random seed (in the `Seed:` line; this can be overridden with the `-seed=N` flag).

Further output lines have the form of an event code and statistics. The possible event codes are:

READ The fuzzer has read in all of the provided input samples from the corpus directories.

INITED The fuzzer has completed initialization, which includes running each of the initial input samples through the code under test.

NEW The fuzzer has created a test input that covers new areas of the code under test. This input will be saved to the primary corpus directory.

pulse The fuzzer has generated 2^n inputs (generated periodically to reassure the user that the fuzzer is still working).

DONE The fuzzer has completed operation because it has reached the specified iteration limit (`-runs`) or time limit (`-max_total_time`).

RELOAD The fuzzer is performing a periodic reload of inputs from the corpus directory; this allows it to discover any inputs discovered by other fuzzer processes (see [Parallel Fuzzing](#)).

Each output line also reports the following statistics (when non-zero):

cov: Total number of code blocks or edges covered by the executing the current corpus.

ft: libFuzzer uses different signals to evaluate the code coverage: edge coverage, edge counters, value profiles, indirect caller/callee pairs, etc. These signals combined are called *features* (*ft*).

corp: Number of entries in the current in-memory test corpus and its size in bytes.

exec/s: Number of fuzzer iterations per second.

rss: Current memory consumption.

For **NEW** events, the output line also includes information about the mutation operation that produced the new input:

L: Size of the new input in bytes.

MS: `<n> <operations>` Count and list of the mutation operations used to generate the input.

3.9.6 Examples

- [Toy example](#)
- [More examples](#)

Toy example

A simple function that does something interesting if it receives the input “HI!”:

```
cat << EOF > test_fuzzer.cc
#include <stdint.h>
#include <stddef.h>
extern "C" int LLVMFuzzerTestOneInput(const uint8_t *data, size_t size) {
    if (size > 0 && data[0] == 'H')
        if (size > 1 && data[1] == 'I')
            if (size > 2 && data[2] == '!')
                __builtin_trap();
    return 0;
}
EOF
# Build test_fuzzer.cc with asan and link against libFuzzer.a
clang++ -fsanitize=address -fsanitize-coverage=trace-pc-guard test_fuzzer.cc_
↳ libFuzzer.a
# Run the fuzzer with no corpus.
./a.out
```

You should get an error pretty quickly:

```
INFO: Seed: 1523017872
INFO: Loaded 1 modules (16 guards): [0x744e60, 0x744ea0),
INFO: -max_len is not provided, using 64
INFO: A corpus is not provided, starting from an empty corpus
#0    READ units: 1
#1    INITED cov: 3 ft: 2 corp: 1/1b exec/s: 0 rss: 24Mb
#3811 NEW    cov: 4 ft: 3 corp: 2/2b exec/s: 0 rss: 25Mb L: 1 MS: 5 ChangeBit-
↳ ChangeByte-ChangeBit-ShuffleBytes-ChangeByte-
#3827 NEW    cov: 5 ft: 4 corp: 3/4b exec/s: 0 rss: 25Mb L: 2 MS: 1 CopyPart-
#3963 NEW    cov: 6 ft: 5 corp: 4/6b exec/s: 0 rss: 25Mb L: 2 MS: 2 ShuffleBytes-
↳ ChangeBit-
#4167 NEW    cov: 7 ft: 6 corp: 5/9b exec/s: 0 rss: 25Mb L: 3 MS: 1 InsertByte-
==31511== ERROR: libFuzzer: deadly signal
...
artifact_prefix='./'; Test unit written to ./crash-
↳ b13e8756b13a00cf168300179061fb4b91fefbed
```

More examples

Examples of real-life fuzz targets and the bugs they find can be found at <http://tutorial.libfuzzer.info>. Among other things you can learn how to detect [Heartbleed](#) in one second.

3.9.7 Advanced features

- *Dictionaries*
- *Tracing CMP instructions*
- *Value Profile*
- *Fuzzer-friendly build mode*

- *AFL compatibility*
- *How good is my fuzzer?*
- *User-supplied mutators*
- *Startup initialization*
- *Leaks*

Dictionaries

LibFuzzer supports user-supplied dictionaries with input language keywords or other interesting byte sequences (e.g. multi-byte magic values). Use `-dict=DICTIONARY_FILE`. For some input languages using a dictionary may significantly improve the search speed. The dictionary syntax is similar to that used by [AFL](#) for its `-x` option:

```
# Lines starting with '#' and empty lines are ignored.

# Adds "blah" (w/o quotes) to the dictionary.
kw1="blah"
# Use \\ for backslash and \" for quotes.
kw2=\"\\ac\\dc\"
# Use \xAB for hex values
kw3=\"\xF7\xF8\"
# the name of the keyword followed by '=' may be omitted:
"foo\x0Abar"
```

Tracing CMP instructions

With an additional compiler flag `-fsanitize-coverage=trace-cmp` (see [SanitizerCoverageTraceDataFlow](#)) libFuzzer will intercept CMP instructions and guide mutations based on the arguments of intercepted CMP instructions. This may slow down the fuzzing but is very likely to improve the results.

Value Profile

EXPERIMENTAL. With `-fsanitize-coverage=trace-cmp` and extra run-time flag `-use_value_profile=1` the fuzzer will collect value profiles for the parameters of compare instructions and treat some new values as new coverage.

The current implementation does roughly the following:

- The compiler instruments all CMP instructions with a callback that receives both CMP arguments.
- The callback computes $(caller_pc \& 4095) \mid (popcnt(Arg1 \wedge Arg2) \ll 12)$ and uses this value to set a bit in a bitset.
- Every new observed bit in the bitset is treated as new coverage.

This feature has a potential to discover many interesting inputs, but there are two downsides. First, the extra instrumentation may bring up to 2x additional slowdown. Second, the corpus may grow by several times.

Fuzzer-friendly build mode

Sometimes the code under test is not fuzzing-friendly. Examples:

- The target code uses a PRNG seeded e.g. by system time and thus two consequent invocations may potentially execute different code paths even if the end result will be the same. This will cause a fuzzer to treat two similar inputs as significantly different and it will blow up the test corpus. E.g. libxml uses `rand()` inside its hash table.
- The target code uses checksums to protect from invalid inputs. E.g. png checks CRC for every chunk.

In many cases it makes sense to build a special fuzzing-friendly build with certain fuzzing-unfriendly features disabled. We propose to use a common build macro for all such cases for consistency: `FUZZING_BUILD_MODE_UNSAFE_FOR_PRODUCTION`.

```
void MyInitPRNG() {  
#ifdef FUZZING_BUILD_MODE_UNSAFE_FOR_PRODUCTION  
    // In fuzzing mode the behavior of the code should be deterministic.  
    srand(0);  
#else  
    srand(time(0));  
#endif  
}
```

AFL compatibility

LibFuzzer can be used together with [AFL](#) on the same test corpus. Both fuzzers expect the test corpus to reside in a directory, one file per input. You can run both fuzzers on the same corpus, one after another:

```
./afl-fuzz -i testcase_dir -o findings_dir /path/to/program @@  
./llvm-fuzz testcase_dir findings_dir # Will write new tests to testcase_dir
```

Periodically restart both fuzzers so that they can use each other's findings. Currently, there is no simple way to run both fuzzing engines in parallel while sharing the same corpus dir.

You may also use AFL on your target function `LLVMFuzzerTestOneInput`: see an example [here](#).

How good is my fuzzer?

Once you implement your target function `LLVMFuzzerTestOneInput` and fuzz it to death, you will want to know whether the function or the corpus can be improved further. One easy to use metric is, of course, code coverage. You can get the coverage for your corpus like this:

```
./fuzzer CORPUS_DIR -runs=0 -print_coverage=1
```

This will run all tests in the `CORPUS_DIR` but will not perform any fuzzing. At the end of the process it will print text describing what code has been covered and what hasn't.

Alternatively, use

```
./fuzzer CORPUS_DIR -runs=0 -dump_coverage=1
```

which will dump a `.sancov` file with coverage information. See [SanitizerCoverage](#) for details on querying the file using the `sancov` tool.

You may also use other ways to visualize coverage, e.g. using [Clang coverage](#), but those will require you to rebuild the code with different compiler flags.

User-supplied mutators

LibFuzzer allows to use custom (user-supplied) mutators, see [FuzzerInterface.h](#)

Startup initialization

If the library being tested needs to be initialized, there are several options.

The simplest way is to have a statically initialized global object inside *LLVMFuzzerTestOneInput* (or in global scope if that works for you):

```
extern "C" int LLVMFuzzerTestOneInput (const uint8_t *Data, size_t Size) {
    static bool Initialized = DoInitialization();
    ...
}
```

Alternatively, you may define an optional init function and it will receive the program arguments that you can read and modify. Do this **only** if you really need to access argv/argc.

```
extern "C" int LLVMFuzzerInitialize(int *argc, char ***argv) {
    ReadAndMaybeModify(argc, argv);
    return 0;
}
```

Leaks

Binaries built with [AddressSanitizer](#) or [LeakSanitizer](#) will try to detect memory leaks at the process shutdown. For in-process fuzzing this is inconvenient since the fuzzer needs to report a leak with a reproducer as soon as the leaky mutation is found. However, running full leak detection after every mutation is expensive.

By default (`-detect_leaks=1`) libFuzzer will count the number of `malloc` and `free` calls when executing every mutation. If the numbers don't match (which by itself doesn't mean there is a leak) libFuzzer will invoke the more expensive [LeakSanitizer](#) pass and if the actual leak is found, it will be reported with the reproducer and the process will exit.

If your target has massive leaks and the leak detection is disabled you will eventually run out of RAM (see the `-rss_limit_mb` flag).

3.9.8 Developing libFuzzer

Building libFuzzer as a part of LLVM project and running its test requires fresh clang as the host compiler and special CMake configuration:

```
cmake -GNinja -DCMAKE_C_COMPILER=clang -DCMAKE_CXX_COMPILER=clang++ -DLLVM_USE_
↪SANITIZER=Address -DLLVM_USE_SANITIZER_COVERAGE=YES -DCMAKE_BUILD_TYPE=Release -
↪DLLVM_ENABLE_ASSERTIONS=ON /path/to/llvm
ninja check-fuzzer
```

3.9.9 Fuzzing components of LLVM

- *clang-format-fuzzer*

- *clang-fuzzer*
- *llvm-as-fuzzer*
- *llvm-mc-fuzzer*
- *Buildbot*

To build any of the LLVM fuzz targets use the build instructions above.

clang-format-fuzzer

The inputs are random pieces of C++-like text.

```
ninja clang-format-fuzzer
mkdir CORPUS_DIR
./bin/clang-format-fuzzer CORPUS_DIR
```

Optionally build other kinds of binaries (ASan+Debug, MSan, UBSan, etc).

Tracking bug: https://llvm.org/bugs/show_bug.cgi?id=23052

clang-fuzzer

The behavior is very similar to `clang-format-fuzzer`.

Tracking bug: https://llvm.org/bugs/show_bug.cgi?id=23057

llvm-as-fuzzer

Tracking bug: https://llvm.org/bugs/show_bug.cgi?id=24639

llvm-mc-fuzzer

This tool fuzzes the MC layer. Currently it is only able to fuzz the disassembler but it is hoped that assembly, and round-trip verification will be added in future.

When run in disassembly mode, the inputs are opcodes to be disassembled. The fuzzer will consume as many instructions as possible and will stop when it finds an invalid instruction or runs out of data.

Please note that the command line interface differs slightly from that of other fuzzers. The fuzzer arguments should follow `--fuzzer-args` and should have a single dash, while other arguments control the operation mode and target in a similar manner to `llvm-mc` and should have two dashes. For example:

```
llvm-mc-fuzzer --triple=aarch64-linux-gnu --disassemble --fuzzer-args -max_len=4 -
↪ jobs=10
```

Buildbot

A buildbot continuously runs the above fuzzers for LLVM components, with results shown at http://lab.llvm.org:8011/builders/sanitizer-x86_64-linux-fuzzer.

3.9.10 FAQ

Q. Why doesn't libFuzzer use any of the LLVM support?

There are two reasons.

First, we want this library to be used outside of the LLVM without users having to build the rest of LLVM. This may sound unconvincing for many LLVM folks, but in practice the need for building the whole LLVM frightens many potential users – and we want more users to use this code.

Second, there is a subtle technical reason not to rely on the rest of LLVM, or any other large body of code (maybe not even STL). When coverage instrumentation is enabled, it will also instrument the LLVM support code which will blow up the coverage set of the process (since the fuzzer is in-process). In other words, by using more external dependencies we will slow down the fuzzer while the main reason for it to exist is extreme speed.

Q. What about Windows then? The fuzzer contains code that does not build on Windows.

Volunteers are welcome.

Q. When libFuzzer is not a good solution for a problem?

- If the test inputs are validated by the target library and the validator asserts/crashes on invalid inputs, in-process fuzzing is not applicable.
- Bugs in the target library may accumulate without being detected. E.g. a memory corruption that goes undetected at first and then leads to a crash while testing another input. This is why it is highly recommended to run this in-process fuzzer with all sanitizers to detect most bugs on the spot.
- It is harder to protect the in-process fuzzer from excessive memory consumption and infinite loops in the target library (still possible).
- The target library should not have significant global state that is not reset between the runs.
- Many interesting target libraries are not designed in a way that supports the in-process fuzzer interface (e.g. require a file path instead of a byte array).
- If a single test run takes a considerable fraction of a second (or more) the speed benefit from the in-process fuzzer is negligible.
- If the target library runs persistent threads (that outlive execution of one test) the fuzzing results will be unreliable.

Q. So, what exactly this Fuzzer is good for?

This Fuzzer might be a good choice for testing libraries that have relatively small inputs, each input takes < 10ms to run, and the library code is not expected to crash on invalid inputs. Examples: regular expression matchers, text or binary format parsers, compression, network, crypto.

3.9.11 Trophies

- GLIBC: <https://sourceware.org/glibc/wiki/FuzzingLibc>
- MUSL LIBC: [1] [2]
- pugixml

- PCRE: Search for “LLVM fuzzer” in <http://vcs.pcre.org/pcre2/code/trunk/ChangeLog?view=markup>; also in [bugzilla](#)
- ICU
- Freetype
- Harfbuzz
- SQLite
- Python
- OpenSSL/BoringSSL: [1] [2] [3] [4] [5] [6]
- Libxml2 and [HT206167] (CVE-2015-5312, CVE-2015-7500, CVE-2015-7942)
- Linux Kernel’s BPF verifier
- Capstone: [1] [2]
- file:[1] [2] [3] [4]
- Radare2: [1]
- gRPC: [1] [2] [3] [4] [5] [6]
- WOFF2: [1]
- LLVM: Clang, Clang-format, libc++, llvm-as, Demangler, Disassembler: <http://reviews.llvm.org/rL247405>, <http://reviews.llvm.org/rL247414>, <http://reviews.llvm.org/rL247416>, <http://reviews.llvm.org/rL247417>, <http://reviews.llvm.org/rL247420>, <http://reviews.llvm.org/rL247422>.
- Tensorflow: [1]
- Ffmpeg: [1] [2] [3] [4]
- Wireshark

3.10 Scudo Hardened Allocator

- *Introduction*
- *Design*
- *Usage*

3.10.1 Introduction

The Scudo Hardened Allocator is a user-mode allocator based on LLVM Sanitizer’s CombinedAllocator, which aims at providing additional mitigations against heap based vulnerabilities, while maintaining good performance.

Currently, the allocator supports (was tested on) the following architectures:

- i386 (& i686) (32-bit);
- x86_64 (64-bit);
- armhf (32-bit);
- AArch64 (64-bit).

The name “Scudo” has been retained from the initial implementation (Escudo meaning Shield in Spanish and Portuguese).

3.10.2 Design

Chunk Header

Every chunk of heap memory will be preceded by a chunk header. This has two purposes, the first one being to store various information about the chunk, the second one being to detect potential heap overflows. In order to achieve this, the header will be checksummed, involving the pointer to the chunk itself and a global secret. Any corruption of the header will be detected when said header is accessed, and the process terminated.

The following information is stored in the header:

- the 16-bit checksum;
- the unused bytes amount for that chunk, which is necessary for computing the size of the chunk;
- the state of the chunk (available, allocated or quarantined);
- the allocation type (malloc, new, new[] or memalign), to detect potential mismatches in the allocation APIs used;
- the offset of the chunk, which is the distance in bytes from the beginning of the returned chunk to the beginning of the backend allocation;
- a 8-bit salt.

This header fits within 8 bytes, on all platforms supported.

The checksum is computed as a CRC32 (made faster with hardware support) of the global secret, the chunk pointer itself, and the 8 bytes of header with the checksum field zeroed out.

The header is atomically loaded and stored to prevent races. This is important as two consecutive chunks could belong to different threads. We also want to avoid any type of double fetches of information located in the header, and use local copies of the header for this purpose.

Delayed Freelist

A delayed freelist allows us to not return a chunk directly to the backend, but to keep it aside for a while. Once a criterion is met, the delayed freelist is emptied, and the quarantined chunks are returned to the backend. This helps mitigate use-after-free vulnerabilities by reducing the determinism of the allocation and deallocation patterns.

This feature is using the Sanitizer’s Quarantine as its base, and the amount of memory that it can hold is configurable by the user (see the Options section below).

Randomness

It is important for the allocator to not make use of fixed addresses. We use the dynamic base option for the SizeClassAllocator, allowing us to benefit from the randomness of mmap.

3.10.3 Usage

Library

The allocator static library can be built from the LLVM build tree thanks to the `scudo` CMake rule. The associated tests can be exercised thanks to the `check-scudo` CMake rule.

Linking the static library to your project can require the use of the `whole-archive` linker flag (or equivalent), depending on your linker. Additional flags might also be necessary.

Your linked binary should now make use of the Scudo allocation and deallocation functions.

You may also build Scudo like this:

```
cd $LLVM/projects/compiler-rt/lib
clang++ -fPIC -std=c++11 -msse4.2 -O2 -I. scudo/*.cpp \
  $(\ls sanitizer_common/*.{cc,S} | grep -v "sanitizer_termination\|sanitizer_common_
↪nolibc") \
  -shared -o scudo-allocator.so -pthread
```

and then use it with existing binaries as follows:

```
LD_PRELOAD=`pwd`/scudo-allocator.so ./a.out
```

Options

Several aspects of the allocator can be configured through the following ways:

- by defining a `__scudo_default_options` function in one's program that returns the options string to be parsed. Said function must have the following prototype: `extern "C" const char* __scudo_default_options()`.
- through the environment variable `SCUDO_OPTIONS`, containing the options string to be parsed. Options defined this way will override any definition made through `__scudo_default_options`;

The options string follows a syntax similar to ASan, where distinct options can be assigned in the same string, separated by colons.

For example, using the environment variable:

```
SCUDO_OPTIONS="DeleteSizeMismatch=1:QuarantineSizeMb=16" ./a.out
```

Or using the function:

```
extern "C" const char *__scudo_default_options() {
    return "DeleteSizeMismatch=1:QuarantineSizeMb=16";
}
```

The following options are available:

Option	64-bit default	32-bit default	Description
QuarantineSizeMb	64	16	The size (in Mb) of quarantine used to delay the actual deallocation of chunks. Lower value may reduce memory usage but decrease the effectiveness of the mitigation; a negative value will fallback to a default of 64Mb.
ThreadLocalQuarantineSizeKb	1024	256	The size (in Kb) of per-thread cache use to offload the global quarantine. Lower value may reduce memory usage but might increase contention on the global quarantine.
Deallocation-TypeMismatch	true	true	Whether or not we report errors on malloc/delete, new/free, new/delete[], etc.
Delete-SizeMismatch	true	true	Whether or not we report errors on mismatch between sizes of new and delete.
ZeroContents	false	false	Whether or not we zero chunk contents on allocation and deallocation.

Allocator related common Sanitizer options can also be passed through Scudo options, such as `allocator_may_return_null`. A detailed list including those can be found here: <https://github.com/google/sanitizers/wiki/SanitizerCommonFlags>.

3.11 Using -opt-bisect-limit to debug optimization errors

- *Introduction*
- *Getting Started*
- *Bisection Index Values*
- *Example Usage*
- *Pass Skipping Implementation*
- *Adding Finer Granularity*

3.11.1 Introduction

The `-opt-bisect-limit` option provides a way to disable all optimization passes above a specified limit without modifying the way in which the Pass Managers are populated. The intention of this option is to assist in tracking down problems where incorrect transformations during optimization result in incorrect run-time behavior.

This feature is implemented on an opt-in basis. Passes which can be safely skipped while still allowing correct code generation call a function to check the opt-bisect limit before performing optimizations. Passes which either must be run or do not modify the IR do not perform this check and are therefore never skipped. Generally, this means analysis passes, passes that are run at `CodeGenOpt::None` and passes which are required for register allocation.

The `-opt-bisect-limit` option can be used with any tool, including front ends such as clang, that uses the core LLVM library for optimization and code generation. The exact syntax for invoking the option is discussed below.

This feature is not intended to replace other debugging tools such as `bugpoint`. Rather it provides an alternate course of action when reproducing the problem requires a complex build infrastructure that would make using `bugpoint` impractical or when reproducing the failure requires a sequence of transformations that is difficult to replicate with tools like `opt` and `llc`.

3.11.2 Getting Started

The `-opt-bisect-limit` command line option can be passed directly to tools such as `opt`, `llc` and `lli`. The syntax is as follows:

```
<tool name> [other options] -opt-bisect-limit=<limit>
```

If a value of `-1` is used the tool will perform all optimizations but a message will be printed to `stderr` for each optimization that could be skipped indicating the index value that is associated with that optimization. To skip optimizations, pass the value of the last optimization to be performed as the `opt-bisect-limit`. All optimizations with a higher index value will be skipped.

In order to use the `-opt-bisect-limit` option with a driver that provides a wrapper around the LLVM core library, an additional prefix option may be required, as defined by the driver. For example, to use this option with `clang`, the “`-mllvm`” prefix must be used. A typical `clang` invocation would look like this:

```
clang -O2 -mllvm -opt-bisect-limit=256 my_file.c
```

The `-opt-bisect-limit` option may also be applied to link-time optimizations by using a prefix to indicate that this is a plug-in option for the linker. The following syntax will set a bisect limit for LTO transformations:

```
# When using lld, or ld64 (macOS)
clang -flto -Wl,-mllvm,-opt-bisect-limit=256 my_file.o my_other_file.o
# When using Gold
clang -flto -Wl,-plugin-opt,-opt-bisect-limit=256 my_file.o my_other_file.o
```

LTO passes are run by a library instance invoked by the linker. Therefore any passes run in the primary driver compilation phase are not affected by options passed via ‘`-Wl,-plugin-opt`’ and LTO passes are not affected by options passed to the driver-invoked LLVM invocation via ‘`-mllvm`’.

3.11.3 Bisection Index Values

The granularity of the optimizations associated with a single index value is variable. Depending on how the optimization pass has been instrumented the value may be associated with as much as all transformations that would have been performed by an optimization pass on an IR unit for which it is invoked (for instance, during a single call of `runOnFunction` for a `FunctionPass`) or as little as a single transformation. The index values may also be nested so that if an invocation of the pass is not skipped individual transformations within that invocation may still be skipped.

The order of the values assigned is guaranteed to remain stable and consistent from one run to the next up to and including the value specified as the limit. Above the limit value skipping of optimizations can cause a change in the numbering, but because all optimizations above the limit are skipped this is not a problem.

When an `opt-bisect` index value refers to an entire invocation of the run function for a pass, the pass will query whether or not it should be skipped each time it is invoked and each invocation will be assigned a unique value. For example, if a `FunctionPass` is used with a module containing three functions a different index value will be assigned to the pass for each of the functions as the pass is run. The pass may be run on two functions but skipped for the third.

If the pass internally performs operations on a smaller IR unit the pass must be specifically instrumented to enable bisection at this finer level of granularity (see below for details).

3.11.4 Example Usage

```
$ opt -O2 -o test-opt.bc -opt-bisect-limit=16 test.ll

BISECT: running pass (1) Simplify the CFG on function (g)
BISECT: running pass (2) SROA on function (g)
BISECT: running pass (3) Early CSE on function (g)
BISECT: running pass (4) Infer set function attributes on module (test.ll)
BISECT: running pass (5) Interprocedural Sparse Conditional Constant Propagation on
↳module (test.ll)
BISECT: running pass (6) Global Variable Optimizer on module (test.ll)
BISECT: running pass (7) Promote Memory to Register on function (g)
BISECT: running pass (8) Dead Argument Elimination on module (test.ll)
BISECT: running pass (9) Combine redundant instructions on function (g)
BISECT: running pass (10) Simplify the CFG on function (g)
BISECT: running pass (11) Remove unused exception handling info on SCC (<<null
↳function>>)
BISECT: running pass (12) Function Integration/Inlining on SCC (<<null function>>)
BISECT: running pass (13) Deduce function attributes on SCC (<<null function>>)
BISECT: running pass (14) Remove unused exception handling info on SCC (f)
BISECT: running pass (15) Function Integration/Inlining on SCC (f)
BISECT: running pass (16) Deduce function attributes on SCC (f)
BISECT: NOT running pass (17) Remove unused exception handling info on SCC (g)
BISECT: NOT running pass (18) Function Integration/Inlining on SCC (g)
BISECT: NOT running pass (19) Deduce function attributes on SCC (g)
BISECT: NOT running pass (20) SROA on function (g)
BISECT: NOT running pass (21) Early CSE on function (g)
BISECT: NOT running pass (22) Speculatively execute instructions if target has
↳divergent branches on function (g)
... etc. ...
```

3.11.5 Pass Skipping Implementation

The `-opt-bisect-limit` implementation depends on individual passes opting in to the `opt-bisect` process. The `OptBisect` object that manages the process is entirely passive and has no knowledge of how any pass is implemented. When a pass is run if the pass may be skipped, it should call the `OptBisect` object to see if it should be skipped.

The `OptBisect` object is intended to be accessed through `LLVMContext` and each `Pass` base class contains a helper function that abstracts the details in order to make this check uniform across all passes. These helper functions are:

```
bool ModulePass::skipModule(Module &M);
bool CallGraphSCCPass::skipSCC(CallGraphSCC &SCC);
bool FunctionPass::skipFunction(const Function &F);
bool BasicBlockPass::skipBasicBlock(const BasicBlock &BB);
bool LoopPass::skipLoop(const Loop *L);
```

A `MachineFunctionPass` should use `FunctionPass::skipFunction()` as such:

```
bool MyMachineFunctionPass::runOnMachineFunction(Function &MF) {
    if (skipFunction(*MF.getFunction()))
        return false;
    // Otherwise, run the pass normally.
}
```

In addition to checking with the `OptBisect` class to see if the pass should be skipped, the `skipFunction()`, `skipLoop()` and `skipBasicBlock()` helper functions also look for the presence of the “`optnone`” function attribute. The calling

pass will be unable to determine whether it is being skipped because the “optnone” attribute is present or because the opt-bisect-limit has been reached. This is desirable because the behavior should be the same in either case.

The majority of LLVM passes which can be skipped have already been instrumented in the manner described above. If you are adding a new pass or believe you have found a pass which is not being included in the opt-bisect process but should be, you can add it as described above.

3.11.6 Adding Finer Granularity

Once the pass in which an incorrect transformation is performed has been determined, it may be useful to perform further analysis in order to determine which specific transformation is causing the problem. Debug counters can be used for this purpose.

LLVM Language Reference Manual Defines the LLVM intermediate representation and the assembly form of the different nodes.

LLVM Atomic Instructions and Concurrency Guide Information about LLVM’s concurrency model.

LLVM Programmer’s Manual Introduction to the general layout of the LLVM sourcebase, important classes and APIs, and some tips & tricks.

LLVM Extensions LLVM-specific extensions to tools and formats LLVM seeks compatibility with.

CommandLine 2.0 Library Manual Provides information on using the command line parsing library.

LLVM Coding Standards Details the LLVM coding standards and provides useful information on writing efficient C++ code.

How to set up LLVM-style RTTI for your class hierarchy How to make `isa<>`, `dyn_cast<>`, etc. available for clients of your class hierarchy.

Extending LLVM: Adding instructions, intrinsics, types, etc. Look here to see how to add instructions and intrinsics to LLVM.

Doxygen generated documentation (classes) (tarball)

Documentation for Go bindings

ViewVC Repository Browser

Architecture & Platform Information for Compiler Writers A list of helpful links for compiler writers.

libFuzzer – a library for coverage-guided fuzz testing. A library for writing in-process guided fuzzers.

Scudo Hardened Allocator A library that implements a security-hardened `malloc()`.

Using -opt-bisect-limit to debug optimization errors A command line option for debugging optimization-induced failures.

For API clients and LLVM developers.

4.1 LLVM Alias Analysis Infrastructure

- *Introduction*
- *AliasAnalysis Class Overview*
 - *Representation of Pointers*
 - *The alias method*
 - * *Must, May, and No Alias Responses*
 - *The getModRefInfo methods*
 - *Other useful AliasAnalysis methods*
 - * *The pointsToConstantMemory method*
 - * *The doesNotAccessMemory and onlyReadsMemory methods*
- *Writing a new AliasAnalysis Implementation*
 - *Different Pass styles*
 - *Required initialization calls*
 - *Required methods to override*
 - *Interfaces which may be specified*
 - *AliasAnalysis chaining behavior*
 - *Updating analysis results for transformations*

- * *The deleteValue method*
- * *The copyValue method*
- * *The replaceWithNewValue method*
- * *The addEscapingUse method*
- *Efficiency Issues*
- *Limitations*
- *Using alias analysis results*
 - *Using the MemoryDependenceAnalysis Pass*
 - *Using the AliasSetTracker class*
 - * *The AliasSetTracker implementation*
 - *Using the AliasAnalysis interface directly*
- *Existing alias analysis implementations and clients*
 - *Available AliasAnalysis implementations*
 - * *The -no-aa pass*
 - * *The -basicaa pass*
 - * *The -globalsmodref-aa pass*
 - * *The -steens-aa pass*
 - * *The -ds-aa pass*
 - * *The -scev-aa pass*
 - *Alias analysis driven transformations*
 - * *The -adce pass*
 - * *The -licm pass*
 - * *The -argpromotion pass*
 - * *The -gvn, -memcpyopt, and -dse passes*
 - *Clients for debugging and evaluation of implementations*
 - * *The -print-alias-sets pass*
 - * *The -count-aa pass*
 - * *The -aa-eval pass*
- *Memory Dependence Analysis*

4.1.1 Introduction

Alias Analysis (aka Pointer Analysis) is a class of techniques which attempt to determine whether or not two pointers ever can point to the same object in memory. There are many different algorithms for alias analysis and many different ways of classifying them: flow-sensitive vs. flow-insensitive, context-sensitive vs. context-insensitive, field-sensitive vs. field-insensitive, unification-based vs. subset-based, etc. Traditionally, alias analyses respond to a query with a *Must, May, or No* alias response, indicating that two pointers always point to the same object, might point to the same object, or are known to never point to the same object.

The LLVM `AliasAnalysis` class is the primary interface used by clients and implementations of alias analyses in the LLVM system. This class is the common interface between clients of alias analysis information and the implementations providing it, and is designed to support a wide range of implementations and clients (but currently all clients are assumed to be flow-insensitive). In addition to simple alias analysis information, this class exposes Mod/Ref information from those implementations which can provide it, allowing for powerful analyses and transformations to work well together.

This document contains information necessary to successfully implement this interface, use it, and to test both sides. It also explains some of the finer points about what exactly results mean.

4.1.2 AliasAnalysis Class Overview

The `AliasAnalysis` class defines the interface that the various alias analysis implementations should support. This class exports two important enums: `AliasResult` and `ModRefResult` which represent the result of an alias query or a mod/ref query, respectively.

The `AliasAnalysis` interface exposes information about memory, represented in several different ways. In particular, memory objects are represented as a starting address and size, and function calls are represented as the actual call or invoke instructions that performs the call. The `AliasAnalysis` interface also exposes some helper methods which allow you to get mod/ref information for arbitrary instructions.

All `AliasAnalysis` interfaces require that in queries involving multiple values, values which are not *constants* are all defined within the same function.

Representation of Pointers

Most importantly, the `AliasAnalysis` class provides several methods which are used to query whether or not two memory objects alias, whether function calls can modify or read a memory object, etc. For all of these queries, memory objects are represented as a pair of their starting address (a symbolic LLVM `Value*`) and a static size.

Representing memory objects as a starting address and a size is critically important for correct Alias Analyses. For example, consider this (silly, but possible) C code:

```
int i;
char C[2];
char A[10];
/* ... */
for (i = 0; i != 10; ++i) {
    C[0] = A[i];           /* One byte store */
    C[1] = A[9-i];        /* One byte store */
}
```

In this case, the `basicaa` pass will disambiguate the stores to `C[0]` and `C[1]` because they are accesses to two distinct locations one byte apart, and the accesses are each one byte. In this case, the Loop Invariant Code Motion (LICM) pass can use store motion to remove the stores from the loop. In contrast, the following code:

```
int i;
char C[2];
char A[10];
/* ... */
for (i = 0; i != 10; ++i) {
    ((short*)C)[0] = A[i]; /* Two byte store! */
    C[1] = A[9-i];        /* One byte store */
}
```

In this case, the two stores to `C` do alias each other, because the access to the `&C[0]` element is a two byte access. If size information wasn't available in the query, even the first case would have to conservatively assume that the accesses alias.

The `alias` method

The `alias` method is the primary interface used to determine whether or not two memory objects alias each other. It takes two memory objects as input and returns `MustAlias`, `PartialAlias`, `MayAlias`, or `NoAlias` as appropriate.

Like all `AliasAnalysis` interfaces, the `alias` method requires that either the two pointer values be defined within the same function, or at least one of the values is a *constant*.

Must, May, and No Alias Responses

The `NoAlias` response may be used when there is never an immediate dependence between any memory reference *based* on one pointer and any memory reference *based* the other. The most obvious example is when the two pointers point to non-overlapping memory ranges. Another is when the two pointers are only ever used for reading memory. Another is when the memory is freed and reallocated between accesses through one pointer and accesses through the other — in this case, there is a dependence, but it's mediated by the free and reallocation.

As an exception to this is with the *noalias* keyword; the “irrelevant” dependencies are ignored.

The `MayAlias` response is used whenever the two pointers might refer to the same object.

The `PartialAlias` response is used when the two memory objects are known to be overlapping in some way, regardless whether they start at the same address or not.

The `MustAlias` response may only be returned if the two memory objects are guaranteed to always start at exactly the same location. A `MustAlias` response does not imply that the pointers compare equal.

The `getModRefInfo` methods

The `getModRefInfo` methods return information about whether the execution of an instruction can read or modify a memory location. Mod/Ref information is always conservative: if an instruction **might** read or write a location, `ModRef` is returned.

The `AliasAnalysis` class also provides a `getModRefInfo` method for testing dependencies between function calls. This method takes two call sites (`CS1` & `CS2`), returns `NoModRef` if neither call writes to memory read or written by the other, `Ref` if `CS1` reads memory written by `CS2`, `Mod` if `CS1` writes to memory read or written by `CS2`, or `ModRef` if `CS1` might read or write memory written to by `CS2`. Note that this relation is not commutative.

Other useful `AliasAnalysis` methods

Several other tidbits of information are often collected by various alias analysis implementations and can be put to good use by various clients.

The `pointsToConstantMemory` method

The `pointsToConstantMemory` method returns true if and only if the analysis can prove that the pointer only points to unchanging memory locations (functions, constant global variables, and the null pointer). This information can be used to refine mod/ref information: it is impossible for an unchanging memory location to be modified.

The `doesNotAccessMemory` and `onlyReadsMemory` methods

These methods are used to provide very simple mod/ref information for function calls. The `doesNotAccessMemory` method returns true for a function if the analysis can prove that the function never reads or writes to memory, or if the function only reads from constant memory. Functions with this property are side-effect free and only depend on their input arguments, allowing them to be eliminated if they form common subexpressions or be hoisted out of loops. Many common functions behave this way (e.g., `sin` and `cos`) but many others do not (e.g., `acos`, which modifies the `errno` variable).

The `onlyReadsMemory` method returns true for a function if analysis can prove that (at most) the function only reads from non-volatile memory. Functions with this property are side-effect free, only depending on their input arguments and the state of memory when they are called. This property allows calls to these functions to be eliminated and moved around, as long as there is no store instruction that changes the contents of memory. Note that all functions that satisfy the `doesNotAccessMemory` method also satisfy `onlyReadsMemory`.

4.1.3 Writing a new `AliasAnalysis` Implementation

Writing a new alias analysis implementation for LLVM is quite straight-forward. There are already several implementations that you can use for examples, and the following information should help fill in any details. For a examples, take a look at the *various alias analysis implementations* included with LLVM.

Different Pass styles

The first step to determining what type of *LLVM pass* you need to use for your Alias Analysis. As is the case with most other analyses and transformations, the answer should be fairly obvious from what type of problem you are trying to solve:

1. If you require interprocedural analysis, it should be a `Pass`.
2. If you are a function-local analysis, subclass `FunctionPass`.
3. If you don't need to look at the program at all, subclass `ImmutablePass`.

In addition to the pass that you subclass, you should also inherit from the `AliasAnalysis` interface, of course, and use the `RegisterAnalysisGroup` template to register as an implementation of `AliasAnalysis`.

Required initialization calls

Your subclass of `AliasAnalysis` is required to invoke two methods on the `AliasAnalysis` base class: `getAnalysisUsage` and `InitializeAliasAnalysis`. In particular, your implementation of `getAnalysisUsage` should explicitly call into the `AliasAnalysis::getAnalysisUsage` method in addition to doing any declaring any pass dependencies your pass has. Thus you should have something like this:

```
void getAnalysisUsage(AnalysisUsage &AU) const {
    AliasAnalysis::getAnalysisUsage(AU);
    // declare your dependencies here.
}
```

Additionally, you must invoke the `InitializeAliasAnalysis` method from your analysis run method (run for a `Pass`, `runOnFunction` for a `FunctionPass`, or `InitializePass` for an `ImmutablePass`). For example (as part of a `Pass`):

```
bool run(Module &M) {
    InitializeAliasAnalysis(this);
    // Perform analysis here...
```

```
    return false;
}
```

Required methods to override

You must override the `getAdjustedAnalysisPointer` method on all subclasses of `AliasAnalysis`. An example implementation of this method would look like:

```
void *getAdjustedAnalysisPointer(const void* ID) override {
    if (ID == &AliasAnalysis::ID)
        return (AliasAnalysis*)this;
    return this;
}
```

Interfaces which may be specified

All of the `AliasAnalysis` virtual methods default to providing *chaining* to another alias analysis implementation, which ends up returning conservatively correct information (returning “May” Alias and “Mod/Ref” for alias and mod/ref queries respectively). Depending on the capabilities of the analysis you are implementing, you just override the interfaces you can improve.

AliasAnalysis chaining behavior

With only one special exception (the `-no-aa` pass) every alias analysis pass chains to another alias analysis implementation (for example, the user can specify “`-basicaa -ds-aa -licm`” to get the maximum benefit from both alias analyses). The alias analysis class automatically takes care of most of this for methods that you don’t override. For methods that you do override, in code paths that return a conservative MayAlias or Mod/Ref result, simply return whatever the superclass computes. For example:

```
AliasResult alias(const Value *V1, unsigned V1Size,
                  const Value *V2, unsigned V2Size) {
    if (...)
        return NoAlias;
    ...

    // Couldn't determine a must or no-alias result.
    return AliasAnalysis::alias(V1, V1Size, V2, V2Size);
}
```

In addition to analysis queries, you must make sure to unconditionally pass LLVM *update notification* methods to the superclass as well if you override them, which allows all alias analyses in a change to be updated.

Updating analysis results for transformations

Alias analysis information is initially computed for a static snapshot of the program, but clients will use this information to make transformations to the code. All but the most trivial forms of alias analysis will need to have their analysis results updated to reflect the changes made by these transformations.

The `AliasAnalysis` interface exposes four methods which are used to communicate program changes from the clients to the analysis implementations. Various alias analysis implementations should use these methods to ensure that their internal data structures are kept up-to-date as the program changes (for example, when an instruction is deleted), and clients of alias analysis must be sure to call these interfaces appropriately.

The `deleteValue` method

The `deleteValue` method is called by transformations when they remove an instruction or any other value from the program (including values that do not use pointers). Typically alias analyses keep data structures that have entries for each value in the program. When this method is called, they should remove any entries for the specified value, if they exist.

The `copyValue` method

The `copyValue` method is used when a new value is introduced into the program. There is no way to introduce a value into the program that did not exist before (this doesn't make sense for a safe compiler transformation), so this is the only way to introduce a new value. This method indicates that the new value has exactly the same properties as the value being copied.

The `replaceWithNewValue` method

This method is a simple helper method that is provided to make clients easier to use. It is implemented by copying the old analysis information to the new value, then deleting the old value. This method cannot be overridden by alias analysis implementations.

The `addEscapingUse` method

The `addEscapingUse` method is used when the uses of a pointer value have changed in ways that may invalidate precomputed analysis information. Implementations may either use this callback to provide conservative responses for points whose uses have changed since analysis time, or may recompute some or all of their internal state to continue providing accurate responses.

In general, any new use of a pointer value is considered an escaping use, and must be reported through this callback, *except* for the uses below:

- A `bitcast` or `getelementptr` of the pointer
- A `store` through the pointer (but not a `store of` the pointer)
- A `load` through the pointer

Efficiency Issues

From the LLVM perspective, the only thing you need to do to provide an efficient alias analysis is to make sure that alias analysis **queries** are serviced quickly. The actual calculation of the alias analysis results (the “run” method) is only performed once, but many (perhaps duplicate) queries may be performed. Because of this, try to move as much computation to the run method as possible (within reason).

Limitations

The `AliasAnalysis` infrastructure has several limitations which make writing a new `AliasAnalysis` implementation difficult.

There is no way to override the default alias analysis. It would be very useful to be able to do something like “`opt -my-aa -O2`” and have it use `-my-aa` for all passes which need `AliasAnalysis`, but there is currently no support for that, short of changing the source code and recompiling. Similarly, there is also no way of setting a chain of analyses as the default.

There is no way for transform passes to declare that they preserve `AliasAnalysis` implementations. The `AliasAnalysis` interface includes `deleteValue` and `copyValue` methods which are intended to allow a pass to keep an `AliasAnalysis` consistent, however there's no way for a pass to declare in its `getAnalysisUsage` that it does so. Some passes attempt to use `AU.addPreserved<AliasAnalysis>`, however this doesn't actually have any effect.

`AliasAnalysisCounter(-count-aa)` are implemented as `ModulePass` classes, so if your alias analysis uses `FunctionPass`, it won't be able to use these utilities. If you try to use them, the pass manager will silently route alias analysis queries directly to `BasicAliasAnalysis` instead.

Similarly, the `opt -p` option introduces `ModulePass` passes between each pass, which prevents the use of `FunctionPass` alias analysis passes.

The `AliasAnalysis` API does have functions for notifying implementations when values are deleted or copied, however these aren't sufficient. There are many other ways that LLVM IR can be modified which could be relevant to `AliasAnalysis` implementations which can not be expressed.

The `AliasAnalysisDebugger` utility seems to suggest that `AliasAnalysis` implementations can expect that they will be informed of any relevant `Value` before it appears in an alias query. However, popular clients such as `GVN` don't support this, and are known to trigger errors when run with the `AliasAnalysisDebugger`.

Due to several of the above limitations, the most obvious use for the `AliasAnalysisCounter` utility, collecting stats on all alias queries in a compilation, doesn't work, even if the `AliasAnalysis` implementations don't use `FunctionPass`. There's no way to set a default, much less a default sequence, and there's no way to preserve it.

The `AliasSetTracker` class (which is used by `LICM`) makes a non-deterministic number of alias queries. This can cause stats collected by `AliasAnalysisCounter` to have fluctuations among identical runs, for example. Another consequence is that debugging techniques involving pausing execution after a predetermined number of queries can be unreliable.

Many alias queries can be reformulated in terms of other alias queries. When multiple `AliasAnalysis` queries are chained together, it would make sense to start those queries from the beginning of the chain, with care taken to avoid infinite looping, however currently an implementation which wants to do this can only start such queries from itself.

4.1.4 Using alias analysis results

There are several different ways to use alias analysis results. In order of preference, these are:

Using the `MemoryDependenceAnalysis` Pass

The `memdep` pass uses alias analysis to provide high-level dependence information about memory-using instructions. This will tell you which store feeds into a load, for example. It uses caching and other techniques to be efficient, and is used by `Dead Store Elimination`, `GVN`, and `memcpy` optimizations.

Using the `AliasSetTracker` class

Many transformations need information about alias **sets** that are active in some scope, rather than information about pairwise aliasing. The `AliasSetTracker` class is used to efficiently build these Alias Sets from the pairwise alias analysis information provided by the `AliasAnalysis` interface.

First you initialize the `AliasSetTracker` by using the “add” methods to add information about various potentially aliasing instructions in the scope you are interested in. Once all of the alias sets are completed, your pass should simply iterate through the constructed alias sets, using the `AliasSetTracker begin()/end()` methods.

The `AliasSets` formed by the `AliasSetTracker` are guaranteed to be disjoint, calculate mod/ref information and volatility for the set, and keep track of whether or not all of the pointers in the set are Must aliases. The `AliasSetTracker` also makes sure that sets are properly folded due to call instructions, and can provide a list of pointers in each set.

As an example user of this, the `Loop Invariant Code Motion` pass uses `AliasSetTrackers` to calculate alias sets for each loop nest. If an `AliasSet` in a loop is not modified, then all load instructions from that set may be hoisted out of the loop. If any alias sets are stored to **and** are must alias sets, then the stores may be sunk to outside of the loop, promoting the memory location to a register for the duration of the loop nest. Both of these transformations only apply if the pointer argument is loop-invariant.

The `AliasSetTracker` implementation

The `AliasSetTracker` class is implemented to be as efficient as possible. It uses the union-find algorithm to efficiently merge `AliasSets` when a pointer is inserted into the `AliasSetTracker` that aliases multiple sets. The primary data structure is a hash table mapping pointers to the `AliasSet` they are in.

The `AliasSetTracker` class must maintain a list of all of the LLVM `Value*`s that are in each `AliasSet`. Since the hash table already has entries for each LLVM `Value*` of interest, the `AliasSets` thread the linked list through these hash-table nodes to avoid having to allocate memory unnecessarily, and to make merging alias sets extremely efficient (the linked list merge is constant time).

You shouldn't need to understand these details if you are just a client of the `AliasSetTracker`, but if you look at the code, hopefully this brief description will help make sense of why things are designed the way they are.

Using the `AliasAnalysis` interface directly

If neither of these utility class are what your pass needs, you should use the interfaces exposed by the `AliasAnalysis` class directly. Try to use the higher-level methods when possible (e.g., use mod/ref information instead of the `alias` method directly if possible) to get the best precision and efficiency.

4.1.5 Existing alias analysis implementations and clients

If you're going to be working with the LLVM alias analysis infrastructure, you should know what clients and implementations of alias analysis are available. In particular, if you are implementing an alias analysis, you should be aware of the *the clients* that are useful for monitoring and evaluating different implementations.

Available `AliasAnalysis` implementations

This section lists the various implementations of the `AliasAnalysis` interface. With the exception of the `-no-aa` implementation, all of these *chain* to other alias analysis implementations.

The `-no-aa` pass

The `-no-aa` pass is just like what it sounds: an alias analysis that never returns any useful information. This pass can be useful if you think that alias analysis is doing something wrong and are trying to narrow down a problem.

The `-basicaa` pass

The `-basicaa` pass is an aggressive local analysis that *knows* many important facts:

- Distinct globals, stack allocations, and heap allocations can never alias.
- Globals, stack allocations, and heap allocations never alias the null pointer.
- Different fields of a structure do not alias.
- Indexes into arrays with statically differing subscripts cannot alias.
- Many common standard C library functions *never access memory or only read memory*.
- Pointers that obviously point to constant globals “`pointToConstantMemory`”.
- Function calls can not modify or references stack allocations if they never escape from the function that allocates them (a common case for automatic arrays).

The `-globalmodref-aa` pass

This pass implements a simple context-sensitive mod/ref and alias analysis for internal global variables that don’t “have their address taken”. If a global does not have its address taken, the pass knows that no pointers alias the global. This pass also keeps track of functions that it knows never access memory or never read memory. This allows certain optimizations (e.g. GVN) to eliminate call instructions entirely.

The real power of this pass is that it provides context-sensitive mod/ref information for call instructions. This allows the optimizer to know that calls to a function do not clobber or read the value of the global, allowing loads and stores to be eliminated.

Note: This pass is somewhat limited in its scope (only support non-address taken globals), but is very quick analysis.

The `-steens-aa` pass

The `-steens-aa` pass implements a variation on the well-known “Steensgaard’s algorithm” for interprocedural alias analysis. Steensgaard’s algorithm is a unification-based, flow-insensitive, context-insensitive, and field-insensitive alias analysis that is also very scalable (effectively linear time).

The LLVM `-steens-aa` pass implements a “speculatively field-**sensitive**” version of Steensgaard’s algorithm using the Data Structure Analysis framework. This gives it substantially more precision than the standard algorithm while maintaining excellent analysis scalability.

Note: `-steens-aa` is available in the optional “poolalloc” module. It is not part of the LLVM core.

The `-ds-aa` pass

The `-ds-aa` pass implements the full Data Structure Analysis algorithm. Data Structure Analysis is a modular unification-based, flow-insensitive, context-**sensitive**, and speculatively field-**sensitive** alias analysis that is also quite scalable, usually at $O(n * \log(n))$.

This algorithm is capable of responding to a full variety of alias analysis queries, and can provide context-sensitive mod/ref information as well. The only major facility not implemented so far is support for must-alias information.

Note: `-ds-aa` is available in the optional “poolalloc” module. It is not part of the LLVM core.

The `-scev-aa` pass

The `-scev-aa` pass implements `AliasAnalysis` queries by translating them into `ScalarEvolution` queries. This gives it a more complete understanding of `getelementptr` instructions and loop induction variables than other alias analyses have.

Alias analysis driven transformations

LLVM includes several alias-analysis driven transformations which can be used with any of the implementations above.

The `-adce` pass

The `-adce` pass, which implements Aggressive Dead Code Elimination uses the `AliasAnalysis` interface to delete calls to functions that do not have side-effects and are not used.

The `-licm` pass

The `-licm` pass implements various Loop Invariant Code Motion related transformations. It uses the `AliasAnalysis` interface for several different transformations:

- It uses mod/ref information to hoist or sink load instructions out of loops if there are no instructions in the loop that modifies the memory loaded.
- It uses mod/ref information to hoist function calls out of loops that do not write to memory and are loop-invariant.
- It uses alias information to promote memory objects that are loaded and stored to in loops to live in a register instead. It can do this if there are no may aliases to the loaded/stored memory location.

The `-argpromotion` pass

The `-argpromotion` pass promotes by-reference arguments to be passed in by-value instead. In particular, if pointer arguments are only loaded from it passes in the value loaded instead of the address to the function. This pass uses alias information to make sure that the value loaded from the argument pointer is not modified between the entry of the function and any load of the pointer.

The `-gvn`, `-memcpyopt`, and `-dse` passes

These passes use `AliasAnalysis` information to reason about loads and stores.

Clients for debugging and evaluation of implementations

These passes are useful for evaluating the various alias analysis implementations. You can use them with commands like:

```
% opt -ds-aa -aa-eval foo.bc -disable-output -stats
```

The `-print-alias-sets` pass

The `-print-alias-sets` pass is exposed as part of the `opt` tool to print out the Alias Sets formed by the *AliasSetTracker* class. This is useful if you're using the *AliasSetTracker* class. To use it, use something like:

```
% opt -ds-aa -print-alias-sets -disable-output
```

The `-count-aa` pass

The `-count-aa` pass is useful to see how many queries a particular pass is making and what responses are returned by the alias analysis. As an example:

```
% opt -basicaa -count-aa -ds-aa -count-aa -licm
```

will print out how many queries (and what responses are returned) by the `-licm` pass (of the `-ds-aa` pass) and how many queries are made of the `-basicaa` pass by the `-ds-aa` pass. This can be useful when debugging a transformation or an alias analysis implementation.

The `-aa-eval` pass

The `-aa-eval` pass simply iterates through all pairs of pointers in a function and asks an alias analysis whether or not the pointers alias. This gives an indication of the precision of the alias analysis. Statistics are printed indicating the percent of no/may/must aliases found (a more precise algorithm will have a lower number of may aliases).

4.1.6 Memory Dependence Analysis

Note: We are currently in the process of migrating things from `MemoryDependenceAnalysis` to *MemorySSA*. Please try to use that instead.

If you're just looking to be a client of alias analysis information, consider using the Memory Dependence Analysis interface instead. MemDep is a lazy, caching layer on top of alias analysis that is able to answer the question of what preceding memory operations a given instruction depends on, either at an intra- or inter-block level. Because of its laziness and caching policy, using MemDep can be a significant performance win over accessing alias analysis directly.

4.2 MemorySSA

- *Introduction*
- *MemorySSA Structure*
- *Design of MemorySSA*

- *The walker*
 - * *Locating clobbers yourself*
- *Build-time use optimization*
- *Invalidation and updating*
 - * *Phi placement*
- *Non-Goals*
- *Design tradeoffs*
 - * *Precision*
 - * *Use Optimization*

4.2.1 Introduction

MemorySSA is an analysis that allows us to cheaply reason about the interactions between various memory operations. Its goal is to replace MemoryDependenceAnalysis for most (if not all) use-cases. This is because, unless you're very careful, use of MemoryDependenceAnalysis can easily result in quadratic-time algorithms in LLVM. Additionally, MemorySSA doesn't have as many arbitrary limits as MemoryDependenceAnalysis, so you should get better results, too.

At a high level, one of the goals of MemorySSA is to provide an SSA based form for memory, complete with def-use and use-def chains, which enables users to quickly find may-def and may-uses of memory operations. It can also be thought of as a way to cheaply give versions to the complete state of heap memory, and associate memory operations with those versions.

This document goes over how MemorySSA is structured, and some basic intuition on how MemorySSA works.

A paper on MemorySSA (with notes about how it's implemented in GCC) [can be found here](#). Though, it's relatively out-of-date; the paper references multiple heap partitions, but GCC eventually swapped to just using one, like we now have in LLVM. Like GCC's, LLVM's MemorySSA is intraprocedural.

4.2.2 MemorySSA Structure

MemorySSA is a virtual IR. After it's built, MemorySSA will contain a structure that maps Instructions to MemoryAccesses, which are MemorySSA's parallel to LLVM Instructions.

Each MemoryAccess can be one of three types:

- MemoryPhi
- MemoryUse
- MemoryDef

MemoryPhis are PhiNodes, but for memory operations. If at any point we have two (or more) MemoryDefs that could flow into a BasicBlock, the block's top MemoryAccess will be a MemoryPhi. As in LLVM IR, MemoryPhis don't correspond to any concrete operation. As such, BasicBlocks are mapped to MemoryPhis inside MemorySSA, whereas Instructions are mapped to MemoryUses and MemoryDefs.

Note also that in SSA, Phi nodes merge must-reach definitions (that is, definitions that *must* be new versions of variables). In MemorySSA, PHI nodes merge may-reach definitions (that is, until disambiguated, the versions that reach a phi node may or may not clobber a given variable).

MemoryUses are operations which use but don't modify memory. An example of a MemoryUse is a load, or a readonly function call.

MemoryDefs are operations which may either modify memory, or which introduce some kind of ordering constraints. Examples of MemoryDefs include stores, function calls, loads with acquire (or higher) ordering, volatile operations, memory fences, etc.

Every function that exists has a special MemoryDef called `liveOnEntry`. It dominates every MemoryAccess in the function that MemorySSA is being run on, and implies that we've hit the top of the function. It's the only MemoryDef that maps to no Instruction in LLVM IR. Use of `liveOnEntry` implies that the memory being used is either undefined or defined before the function begins.

An example of all of this overlaid on LLVM IR (obtained by running `opt -passes='print<memoryssa>' -disable-output` on an `.ll` file) is below. When viewing this example, it may be helpful to view it in terms of clobbers. The operands of a given MemoryAccess are all (potential) clobbers of said MemoryAccess, and the value produced by a MemoryAccess can act as a clobber for other MemoryAccesses. Another useful way of looking at it is in terms of heap versions. In that view, operands of a given MemoryAccess are the version of the heap before the operation, and if the access produces a value, the value is the new version of the heap after the operation.

```
define void @foo() {
entry:
    %p1 = alloca i8
    %p2 = alloca i8
    %p3 = alloca i8
    ; 1 = MemoryDef(liveOnEntry)
    store i8 0, i8* %p3
    br label %while.cond

while.cond:
    ; 6 = MemoryPhi({%0,1},{if.end,4})
    br i1 undef, label %if.then, label %if.else

if.then:
    ; 2 = MemoryDef(6)
    store i8 0, i8* %p1
    br label %if.end

if.else:
    ; 3 = MemoryDef(6)
    store i8 1, i8* %p2
    br label %if.end

if.end:
    ; 5 = MemoryPhi({if.then,2},{if.else,3})
    ; MemoryUse(5)
    %1 = load i8, i8* %p1
    ; 4 = MemoryDef(5)
    store i8 2, i8* %p2
    ; MemoryUse(1)
    %2 = load i8, i8* %p3
    br label %while.cond
}
```

The MemorySSA IR is shown in comments that precede the instructions they map to (if such an instruction exists). For example, `1 = MemoryDef(liveOnEntry)` is a MemoryAccess (specifically, a MemoryDef), and it describes the LLVM instruction `store i8 0, i8* %p3`. Other places in MemorySSA refer to this particular MemoryDef as `1` (much like how one can refer to `load i8, i8* %p1` in LLVM with `%1`). Again, MemoryPhis don't correspond to any LLVM Instruction, so the line directly below a MemoryPhi isn't special.

Going from the top down:

- `6 = MemoryPhi({entry, 1}, {if.end, 4})` notes that, when entering `while.cond`, the reaching definition for it is either 1 or 4. This `MemoryPhi` is referred to in the textual IR by the number 6.
- `2 = MemoryDef(6)` notes that `store i8 0, i8* %p1` is a definition, and its reaching definition before it is 6, or the `MemoryPhi` after `while.cond`. (See the [Build-time use optimization](#) and [Precision](#) sections below for why this `MemoryDef` isn't linked to a separate, disambiguated `MemoryPhi`.)
- `3 = MemoryDef(6)` notes that `store i8 0, i8* %p2` is a definition; its reaching definition is also 6.
- `5 = MemoryPhi({if.then, 2}, {if.else, 3})` notes that the clobber before this block could either be 2 or 3.
- `MemoryUse(5)` notes that `load i8, i8* %p1` is a use of memory, and that it's clobbered by 5.
- `4 = MemoryDef(5)` notes that `store i8 2, i8* %p2` is a definition; its reaching definition is 5.
- `MemoryUse(1)` notes that `load i8, i8* %p3` is just a user of memory, and the last thing that could clobber this use is above `while.cond` (e.g. the store to `%p3`). In heap versioning parlance, it really only depends on the heap version 1, and is unaffected by the new heap versions generated since then.

As an aside, `MemoryAccess` is a `Value` mostly for convenience; it's not meant to interact with LLVM IR.

4.2.3 Design of MemorySSA

`MemorySSA` is an analysis that can be built for any arbitrary function. When it's built, it does a pass over the function's IR in order to build up its mapping of `MemoryAccesses`. You can then query `MemorySSA` for things like the dominance relation between `MemoryAccesses`, and get the `MemoryAccess` for any given `Instruction`.

When `MemorySSA` is done building, it also hands you a `MemorySSAWalker` that you can use (see below).

The walker

A structure that helps `MemorySSA` do its job is the `MemorySSAWalker`, or the walker, for short. The goal of the walker is to provide answers to clobber queries beyond what's represented directly by `MemoryAccesses`. For example, given:

```
define void @foo() {
  %a = alloca i8
  %b = alloca i8

  ; 1 = MemoryDef(liveOnEntry)
  store i8 0, i8* %a
  ; 2 = MemoryDef(1)
  store i8 0, i8* %b
}
```

The store to `%a` is clearly not a clobber for the store to `%b`. It would be the walker's goal to figure this out, and return `liveOnEntry` when queried for the clobber of `MemoryAccess 2`.

By default, `MemorySSA` provides a walker that can optimize `MemoryDefs` and `MemoryUses` by consulting whatever alias analysis stack you happen to be using. Walkers were built to be flexible, though, so it's entirely reasonable (and expected) to create more specialized walkers (e.g. one that specifically queries `GlobalsAA`, one that always stops at `MemoryPhi` nodes, etc).

Locating clobbers yourself

If you choose to make your own walker, you can find the clobber for a `MemoryAccess` by walking every `MemoryDef` that dominates said `MemoryAccess`. The structure of `MemoryDefs` makes this relatively simple; they ultimately form a linked list of every clobber that dominates the `MemoryAccess` that you're trying to optimize. In other words, the `definingAccess` of a `MemoryDef` is always the nearest dominating `MemoryDef` or `MemoryPhi` of said `MemoryDef`.

Build-time use optimization

`MemorySSA` will optimize some `MemoryAccesses` at build-time. Specifically, we optimize the operand of every `MemoryUse` to point to the actual clobber of said `MemoryUse`. This can be seen in the above example; the second `MemoryUse` in `if.end` has an operand of `1`, which is a `MemoryDef` from the entry block. This is done to make walking, value numbering, etc, faster and easier.

It is not possible to optimize `MemoryDef` in the same way, as we restrict `MemorySSA` to one heap variable and, thus, one Phi node per block.

Invalidation and updating

Because `MemorySSA` keeps track of LLVM IR, it needs to be updated whenever the IR is updated. “Update”, in this case, includes the addition, deletion, and motion of `Instructions`. The update API is being made on an as-needed basis. If you'd like examples, `GVNHoist` is a user of `MemorySSA`'s update API.

Phi placement

`MemorySSA` only places `MemoryPhis` where they're actually needed. That is, it is a pruned SSA form, like LLVM's SSA form. For example, consider:

```
define void @foo() {
entry:
    %p1 = alloca i8
    %p2 = alloca i8
    %p3 = alloca i8
    ; 1 = MemoryDef(liveOnEntry)
    store i8 0, i8* %p3
    br label %while.cond

while.cond:
    ; 3 = MemoryPhi({%0,1},{if.end,2})
    br i1 undef, label %if.then, label %if.else

if.then:
    br label %if.end

if.else:
    br label %if.end

if.end:
    ; MemoryUse(1)
    %1 = load i8, i8* %p1
    ; 2 = MemoryDef(3)
    store i8 2, i8* %p2
    ; MemoryUse(1)
```

```

%2 = load i8, i8* %p3
br label %while.cond
}

```

Because we removed the stores from `if.then` and `if.else`, a `MemoryPhi` for `if.end` would be pointless, so we don't place one. So, if you need to place a `MemoryDef` in `if.then` or `if.else`, you'll need to also create a `MemoryPhi` for `if.end`.

If it turns out that this is a large burden, we can just place `MemoryPhis` everywhere. Because we have Walkers that are capable of optimizing above said phis, doing so shouldn't prohibit optimizations.

Non-Goals

`MemorySSA` is meant to reason about the relation between memory operations, and enable quicker querying. It isn't meant to be the single source of truth for all potential memory-related optimizations. Specifically, care must be taken when trying to use `MemorySSA` to reason about atomic or volatile operations, as in:

```

define i8 @foo(i8* %a) {
entry:
  br i1 undef, label %if.then, label %if.end

if.then:
  ; l = MemoryDef(liveOnEntry)
  %0 = load volatile i8, i8* %a
  br label %if.end

if.end:
  %av = phi i8 [0, %entry], [%0, %if.then]
  ret i8 %av
}

```

Going solely by `MemorySSA`'s analysis, hoisting the load to `entry` may seem legal. Because it's a volatile load, though, it's not.

Design tradeoffs

Precision

`MemorySSA` in LLVM deliberately trades off precision for speed. Let us think about memory variables as if they were disjoint partitions of the heap (that is, if you have one variable, as above, it represents the entire heap, and if you have multiple variables, each one represents some disjoint portion of the heap)

First, because alias analysis results conflict with each other, and each result may be what an analysis wants (IE TBAA may say no-alias, and something else may say must-alias), it is not possible to partition the heap the way every optimization wants. Second, some alias analysis results are not transitive (IE A noalias B, and B noalias C, does not mean A noalias C), so it is not possible to come up with a precise partitioning in all cases without variables to represent every pair of possible aliases. Thus, partitioning precisely may require introducing at least N^2 new virtual variables, phi nodes, etc.

Each of these variables may be clobbered at multiple def sites.

To give an example, if you were to split up struct fields into individual variables, all aliasing operations that may-def multiple struct fields, will may-def more than one of them. This is pretty common (calls, copies, field stores, etc).

Experience with SSA forms for memory in other compilers has shown that it is simply not possible to do this precisely, and in fact, doing it precisely is not worth it, because now all the optimizations have to walk tons and tons of virtual variables and phi nodes.

So we partition. At the point at which you partition, again, experience has shown us there is no point in partitioning to more than one variable. It simply generates more IR, and optimizations still have to query something to disambiguate further anyway.

As a result, LLVM partitions to one variable.

Use Optimization

Unlike other partitioned forms, LLVM's `MemorySSA` does make one useful guarantee - all loads are optimized to point at the thing that actually clobbers them. This gives some nice properties. For example, for a given store, you can find all loads actually clobbered by that store by walking the immediate uses of the store.

4.3 LLVM Bitcode File Format

- *Abstract*
- *Overview*
- *Bitstream Format*
 - *Magic Numbers*
 - *Primitives*
 - * *Fixed Width Integers*
 - * *Variable Width Integers*
 - * *6-bit characters*
 - * *Word Alignment*
 - *Abbreviation IDs*
 - *Blocks*
 - * *ENTER_SUBBLOCK Encoding*
 - * *END_BLOCK Encoding*
 - *Data Records*
 - * *UNABBREV_RECORD Encoding*
 - * *Abbreviated Record Encoding*
 - *Abbreviations*
 - * *DEFINE_ABBREV Encoding*
 - *Standard Blocks*
 - * *#0 - BLOCKINFO Block*
- *Bitcode Wrapper Format*
- *Native Object File Wrapper Format*

- *LLVM IR Encoding*
 - *Basics*
 - * *LLVM IR Magic Number*
 - * *Signed VBRs*
 - * *LLVM IR Blocks*
 - *MODULE_BLOCK Contents*
 - * *MODULE_CODE_VERSION Record*
 - * *MODULE_CODE_TRIPLE Record*
 - * *MODULE_CODE_DATA_LAYOUT Record*
 - * *MODULE_CODE_ASM Record*
 - * *MODULE_CODE_SECTIONNAME Record*
 - * *MODULE_CODE_DEPLIB Record*
 - * *MODULE_CODE_GLOBALVAR Record*
 - * *MODULE_CODE_FUNCTION Record*
 - * *MODULE_CODE_ALIAS Record*
 - * *MODULE_CODE_GCNAME Record*
 - *PARAMATTR_BLOCK Contents*
 - * *PARAMATTR_CODE_ENTRY Record*
 - * *PARAMATTR_CODE_ENTRY_OLD Record*
 - *PARAMATTR_GROUP_BLOCK Contents*
 - * *PARAMATTR_GRP_CODE_ENTRY Record*
 - *TYPE_BLOCK Contents*
 - * *TYPE_CODE_NUMENTRY Record*
 - * *TYPE_CODE_VOID Record*
 - * *TYPE_CODE_HALF Record*
 - * *TYPE_CODE_FLOAT Record*
 - * *TYPE_CODE_DOUBLE Record*
 - * *TYPE_CODE_LABEL Record*
 - * *TYPE_CODE_OPAQUE Record*
 - * *TYPE_CODE_INTEGER Record*
 - * *TYPE_CODE_POINTER Record*
 - * *TYPE_CODE_FUNCTION_OLD Record*
 - * *TYPE_CODE_ARRAY Record*
 - * *TYPE_CODE_VECTOR Record*
 - * *TYPE_CODE_X86_FP80 Record*

- * *TYPE_CODE_FPI28 Record*
- * *TYPE_CODE_PPC_FPI28 Record*
- * *TYPE_CODE_METADATA Record*
- * *TYPE_CODE_X86_MMX Record*
- * *TYPE_CODE_STRUCT_ANON Record*
- * *TYPE_CODE_STRUCT_NAME Record*
- * *TYPE_CODE_STRUCT_NAMED Record*
- * *TYPE_CODE_FUNCTION Record*
- *CONSTANTS_BLOCK Contents*
- *FUNCTION_BLOCK Contents*
- *VALUE_SYMTAB_BLOCK Contents*
- *METADATA_BLOCK Contents*
- *METADATA_ATTACHMENT Contents*
- *STRTAB_BLOCK Contents*

4.3.1 Abstract

This document describes the LLVM bitstream file format and the encoding of the LLVM IR into it.

4.3.2 Overview

What is commonly known as the LLVM bitcode file format (also, sometimes anachronistically known as bytecode) is actually two things: a *bitstream container format* and an *encoding of LLVM IR* into the container format.

The bitstream format is an abstract encoding of structured data, very similar to XML in some ways. Like XML, bitstream files contain tags, and nested structures, and you can parse the file without having to understand the tags. Unlike XML, the bitstream format is a binary encoding, and unlike XML it provides a mechanism for the file to self-describe “abbreviations”, which are effectively size optimizations for the content.

LLVM IR files may be optionally embedded into a *wrapper* structure, or in a *native object file*. Both of these mechanisms make it easy to embed extra data along with LLVM IR files.

This document first describes the LLVM bitstream format, describes the wrapper format, then describes the record structure used by LLVM IR files.

4.3.3 Bitstream Format

The bitstream format is literally a stream of bits, with a very simple structure. This structure consists of the following concepts:

- A “*magic number*” that identifies the contents of the stream.
- Encoding *primitives* like variable bit-rate integers.
- *Blocks*, which define nested content.
- *Data Records*, which describe entities within the file.

- Abbreviations, which specify compression optimizations for the file.

Note that the *llvm-bcanalyzer* tool can be used to dump and inspect arbitrary bitstreams, which is very useful for understanding the encoding.

Magic Numbers

The first two bytes of a bitcode file are ‘BC’ (0x42, 0x43). The second two bytes are an application-specific magic number. Generic bitcode tools can look at only the first two bytes to verify the file is bitcode, while application-specific programs will want to look at all four.

Primitives

A bitstream literally consists of a stream of bits, which are read in order starting with the least significant bit of each byte. The stream is made up of a number of primitive values that encode a stream of unsigned integer values. These integers are encoded in two ways: either as *Fixed Width Integers* or as *Variable Width Integers*.

Fixed Width Integers

Fixed-width integer values have their low bits emitted directly to the file. For example, a 3-bit integer value encodes 1 as 001. Fixed width integers are used when there are a well-known number of options for a field. For example, boolean values are usually encoded with a 1-bit wide integer.

Variable Width Integers

Variable-width integer (VBR) values encode values of arbitrary size, optimizing for the case where the values are small. Given a 4-bit VBR field, any 3-bit value (0 through 7) is encoded directly, with the high bit set to zero. Values larger than N-1 bits emit their bits in a series of N-1 bit chunks, where all but the last set the high bit.

For example, the value 27 (0x1B) is encoded as 1011 0011 when emitted as a vbr4 value. The first set of four bits indicates the value 3 (011) with a continuation piece (indicated by a high bit of 1). The next word indicates a value of 24 (011 << 3) with no continuation. The sum (3+24) yields the value 27.

6-bit characters

6-bit characters encode common characters into a fixed 6-bit field. They represent the following characters with the following 6-bit values:

'a'	..	'z'	---	0	..	25
'A'	..	'Z'	---	26	..	51
'0'	..	'9'	---	52	..	61
		'.'	---			62
		'_'	---			63

This encoding is only suitable for encoding characters and strings that consist only of the above characters. It is completely incapable of encoding characters not in the set.

Word Alignment

Occasionally, it is useful to emit zero bits until the bitstream is a multiple of 32 bits. This ensures that the bit position in the stream can be represented as a multiple of 32-bit words.

Abbreviation IDs

A bitstream is a sequential series of *Blocks* and *Data Records*. Both of these start with an abbreviation ID encoded as a fixed-bitwidth field. The width is specified by the current block, as described below. The value of the abbreviation ID specifies either a builtin ID (which have special meanings, defined below) or one of the abbreviation IDs defined for the current block by the stream itself.

The set of builtin abbrev IDs is:

- 0 - *END_BLOCK* — This abbrev ID marks the end of the current block.
- 1 - *ENTER_SUBBLOCK* — This abbrev ID marks the beginning of a new block.
- 2 - *DEFINE_ABBREV* — This defines a new abbreviation.
- 3 - *UNABBREV_RECORD* — This ID specifies the definition of an unabbreviated record.

Abbreviation IDs 4 and above are defined by the stream itself, and specify an *abbreviated record encoding*.

Blocks

Blocks in a bitstream denote nested regions of the stream, and are identified by a content-specific id number (for example, LLVM IR uses an ID of 12 to represent function bodies). Block IDs 0-7 are reserved for *standard blocks* whose meaning is defined by Bitcode; block IDs 8 and greater are application specific. Nested blocks capture the hierarchical structure of the data encoded in it, and various properties are associated with blocks as the file is parsed. Block definitions allow the reader to efficiently skip blocks in constant time if the reader wants a summary of blocks, or if it wants to efficiently skip data it does not understand. The LLVM IR reader uses this mechanism to skip function bodies, lazily reading them on demand.

When reading and encoding the stream, several properties are maintained for the block. In particular, each block maintains:

1. A current abbrev id width. This value starts at 2 at the beginning of the stream, and is set every time a block record is entered. The block entry specifies the abbrev id width for the body of the block.
2. A set of abbreviations. Abbreviations may be defined within a block, in which case they are only defined in that block (neither subblocks nor enclosing blocks see the abbreviation). Abbreviations can also be defined inside a *BLOCKINFO* block, in which case they are defined in all blocks that match the ID that the *BLOCKINFO* block is describing.

As sub blocks are entered, these properties are saved and the new sub-block has its own set of abbreviations, and its own abbrev id width. When a sub-block is popped, the saved values are restored.

ENTER_SUBBLOCK Encoding

[ENTER_SUBBLOCK, blockid_{vbr8}, newabbrevlen_{vbr4}, <align32bits>, blocklen₃₂]

The *ENTER_SUBBLOCK* abbreviation ID specifies the start of a new block record. The *blockid* value is encoded as an 8-bit VBR identifier, and indicates the type of block being entered, which can be a *standard block* or an application-specific block. The *newabbrevlen* value is a 4-bit VBR, which specifies the abbrev id width for the sub-block. The *blocklen* value is a 32-bit aligned value that specifies the size of the subblock in 32-bit words. This value allows the reader to skip over the entire block in one jump.

END_BLOCK Encoding

[END_BLOCK, <align32bits>]

The `END_BLOCK` abbreviation ID specifies the end of the current block record. Its end is aligned to 32-bits to ensure that the size of the block is an even multiple of 32-bits.

Data Records

Data records consist of a record code and a number of (up to) 64-bit integer values. The interpretation of the code and values is application specific and may vary between different block types. Records can be encoded either using an unabbrev record, or with an abbreviation. In the LLVM IR format, for example, there is a record which encodes the target triple of a module. The code is `MODULE_CODE_TRIPLE`, and the values of the record are the ASCII codes for the characters in the string.

UNABBREV_RECORD Encoding

```
[UNABBREV_RECORD, codevbr6, numopsvbr6, op0vbr6, op1vbr6, ...]
```

An `UNABBREV_RECORD` provides a default fallback encoding, which is both completely general and extremely inefficient. It can describe an arbitrary record by emitting the code and operands as VBRs.

For example, emitting an LLVM IR target triple as an unabbreviated record requires emitting the `UNABBREV_RECORD` abbrevid, a `vbr6` for the `MODULE_CODE_TRIPLE` code, a `vbr6` for the length of the string, which is equal to the number of operands, and a `vbr6` for each character. Because there are no letters with values less than 32, each letter would need to be emitted as at least a two-part VBR, which means that each letter would require at least 12 bits. This is not an efficient encoding, but it is fully general.

Abbreviated Record Encoding

```
[<abbrevid>, fields...]
```

An abbreviated record is a abbreviation id followed by a set of fields that are encoded according to the *abbreviation definition*. This allows records to be encoded significantly more densely than records encoded with the `UNABBREV_RECORD` type, and allows the abbreviation types to be specified in the stream itself, which allows the files to be completely self describing. The actual encoding of abbreviations is defined below.

The record code, which is the first field of an abbreviated record, may be encoded in the abbreviation definition (as a literal operand) or supplied in the abbreviated record (as a Fixed or VBR operand value).

Abbreviations

Abbreviations are an important form of compression for bitstreams. The idea is to specify a dense encoding for a class of records once, then use that encoding to emit many records. It takes space to emit the encoding into the file, but the space is recouped (hopefully plus some) when the records that use it are emitted.

Abbreviations can be determined dynamically per client, per file. Because the abbreviations are stored in the bitstream itself, different streams of the same format can contain different sets of abbreviations according to the needs of the specific stream. As a concrete example, LLVM IR files usually emit an abbreviation for binary operators. If a specific LLVM module contained no or few binary operators, the abbreviation does not need to be emitted.

DEFINE_ABBREV Encoding

```
[DEFINE_ABBREV, numabbrevopsvbr5, abbrevop0, abbrevop1, ...]
```

A `DEFINE_ABBREV` record adds an abbreviation to the list of currently defined abbreviations in the scope of this block. This definition only exists inside this immediate block — it is not visible in subblocks or enclosing blocks.

Abbreviations are implicitly assigned IDs sequentially starting from 4 (the first application-defined abbreviation ID). Any abbreviations defined in a `BLOCKINFO` record for the particular block type receive IDs first, in order, followed by any abbreviations defined within the block itself. Abbreviated data records reference this ID to indicate what abbreviation they are invoking.

An abbreviation definition consists of the `DEFINE_ABBREV` abbrevid followed by a VBR that specifies the number of abbrev operands, then the abbrev operands themselves. Abbreviation operands come in three forms. They all start with a single bit that indicates whether the abbrev operand is a literal operand (when the bit is 1) or an encoding operand (when the bit is 0).

1. Literal operands — $[1_1, \text{litvalue}_{\text{vbr8}}]$ — Literal operands specify that the value in the result is always a single specific value. This specific value is emitted as a `vbr8` after the bit indicating that it is a literal operand.
2. Encoding info without data — $[0_1, \text{encoding}_3]$ — Operand encodings that do not have extra data are just emitted as their code.
3. Encoding info with data — $[0_1, \text{encoding}_3, \text{value}_{\text{vbr5}}]$ — Operand encodings that do have extra data are emitted as their code, followed by the extra data.

The possible operand encodings are:

- Fixed (code 1): The field should be emitted as a *fixed-width value*, whose width is specified by the operand's extra data.
- VBR (code 2): The field should be emitted as a *variable-width value*, whose width is specified by the operand's extra data.
- Array (code 3): This field is an array of values. The array operand has no extra data, but expects another operand to follow it, indicating the element type of the array. When reading an array in an abbreviated record, the first integer is a `vbr6` that indicates the array length, followed by the encoded elements of the array. An array may only occur as the last operand of an abbreviation (except for the one final operand that gives the array's type).
- Char6 (code 4): This field should be emitted as a *char6-encoded value*. This operand type takes no extra data. Char6 encoding is normally used as an array element type.
- Blob (code 5): This field is emitted as a `vbr6`, followed by padding to a 32-bit boundary (for alignment) and an array of 8-bit objects. The array of bytes is further followed by tail padding to ensure that its total length is a multiple of 4 bytes. This makes it very efficient for the reader to decode the data without having to make a copy of it: it can use a pointer to the data in the mapped in file and poke directly at it. A blob may only occur as the last operand of an abbreviation.

For example, target triples in LLVM modules are encoded as a record of the form `[TRIPLE, 'a', 'b', 'c', 'd']`. Consider if the bitstream emitted the following abbrev entry:

```
[0, Fixed, 4]
[0, Array]
[0, Char6]
```

When emitting a record with this abbreviation, the above entry would be emitted as:

```
[4abbrevwidth, 24, 4vbr6, 06, 16, 26, 36]
```

These values are:

1. The first value, 4, is the abbreviation ID for this abbreviation.
2. The second value, 2, is the record code for `TRIPLE` records within LLVM IR file `MODULE_BLOCK` blocks.
3. The third value, 4, is the length of the array.
4. The rest of the values are the char6 encoded values for "abcd".

With this abbreviation, the triple is emitted with only 37 bits (assuming a abbrev id width of 3). Without the abbreviation, significantly more space would be required to emit the target triple. Also, because the `TRIPLE` value is not emitted as a literal in the abbreviation, the abbreviation can also be used for any other string value.

Standard Blocks

In addition to the basic block structure and record encodings, the bitstream also defines specific built-in block types. These block types specify how the stream is to be decoded or other metadata. In the future, new standard blocks may be added. Block IDs 0-7 are reserved for standard blocks.

#0 - BLOCKINFO Block

The `BLOCKINFO` block allows the description of metadata for other blocks. The currently specified records are:

```
[SETBID (#1), blockid]
[DEFINE_ABBREV, ...]
[BLOCKNAME, ...name...]
[SETRECORDNAME, RecordID, ...name...]
```

The `SETBID` record (code 1) indicates which block ID is being described. `SETBID` records can occur multiple times throughout the block to change which block ID is being described. There must be a `SETBID` record prior to any other records.

Standard `DEFINE_ABBREV` records can occur inside `BLOCKINFO` blocks, but unlike their occurrence in normal blocks, the abbreviation is defined for blocks matching the block ID we are describing, *not* the `BLOCKINFO` block itself. The abbreviations defined in `BLOCKINFO` blocks receive abbreviation IDs as described in [DEFINE_ABBREV](#).

The `BLOCKNAME` record (code 2) can optionally occur in this block. The elements of the record are the bytes of the string name of the block. `llvm-bcanalyzer` can use this to dump out bitcode files symbolically.

The `SETRECORDNAME` record (code 3) can also optionally occur in this block. The first operand value is a record ID number, and the rest of the elements of the record are the bytes for the string name of the record. `llvm-bcanalyzer` can use this to dump out bitcode files symbolically.

Note that although the data in `BLOCKINFO` blocks is described as “metadata,” the abbreviations they contain are essential for parsing records from the corresponding blocks. It is not safe to skip them.

4.3.4 Bitcode Wrapper Format

Bitcode files for LLVM IR may optionally be wrapped in a simple wrapper structure. This structure contains a simple header that indicates the offset and size of the embedded BC file. This allows additional information to be stored alongside the BC file. The structure of this file header is:

```
[Magic32, Version32, Offset32, Size32, CPUType32]
```

Each of the fields are 32-bit fields stored in little endian form (as with the rest of the bitcode file fields). The Magic number is always `0x0B17C0DE` and the version is currently always 0. The Offset field is the offset in bytes to the start of the bitcode stream in the file, and the Size field is the size in bytes of the stream. `CPUType` is a target-specific value that can be used to encode the CPU of the target.

4.3.5 Native Object File Wrapper Format

Bitcode files for LLVM IR may also be wrapped in a native object file (i.e. ELF, COFF, Mach-O). The bitcode must be stored in a section of the object file named `__LLVM`, `__bitcode` for MachO and `.llvmbc` for the other object

formats. This wrapper format is useful for accommodating LTO in compilation pipelines where intermediate objects must be native object files which contain metadata in other sections.

Not all tools support this format.

4.3.6 LLVM IR Encoding

LLVM IR is encoded into a bitstream by defining blocks and records. It uses blocks for things like constant pools, functions, symbol tables, etc. It uses records for things like instructions, global variable descriptors, type descriptions, etc. This document does not describe the set of abbreviations that the writer uses, as these are fully self-described in the file, and the reader is not allowed to build in any knowledge of this.

Basics

LLVM IR Magic Number

The magic number for LLVM IR files is:

[0x04, 0xC4, 0xE4, 0xD4]

When combined with the bitcode magic number and viewed as bytes, this is "BC 0xCODE".

Signed VBRs

Variable Width Integer encoding is an efficient way to encode arbitrary sized unsigned values, but is an extremely inefficient for encoding signed values, as signed values are otherwise treated as maximally large unsigned values.

As such, signed VBR values of a specific width are emitted as follows:

- Positive values are emitted as VBRs of the specified width, but with their value shifted left by one.
- Negative values are emitted as VBRs of the specified width, but the negated value is shifted left by one, and the low bit is set.

With this encoding, small positive and small negative values can both be emitted efficiently. Signed VBR encoding is used in `CST_CODE_INTEGER` and `CST_CODE_WIDE_INTEGER` records within `CONSTANTS_BLOCK` blocks. It is also used for phi instruction operands in *MODULE_CODE_VERSION* 1.

LLVM IR Blocks

LLVM IR is defined with the following blocks:

- 8 — *MODULE_BLOCK* — This is the top-level block that contains the entire module, and describes a variety of per-module information.
- 9 — *PARAMATTR_BLOCK* — This enumerates the parameter attributes.
- 10 — *PARAMATTR_GROUP_BLOCK* — This describes the attribute group table.
- 11 — *CONSTANTS_BLOCK* — This describes constants for a module or function.
- 12 — *FUNCTION_BLOCK* — This describes a function body.
- 14 — *VALUE_SYMTAB_BLOCK* — This describes a value symbol table.
- 15 — *METADATA_BLOCK* — This describes metadata items.

- 16 — *METADATA_ATTACHMENT* — This contains records associating metadata with function instruction values.
- 17 — *TYPE_BLOCK* — This describes all of the types in the module.
- 23 — *STRTAB_BLOCK* — The bitcode file’s string table.

MODULE_BLOCK Contents

The `MODULE_BLOCK` block (id 8) is the top-level block for LLVM bitcode files, and each bitcode file must contain exactly one. In addition to records (described below) containing information about the module, a `MODULE_BLOCK` block may contain the following sub-blocks:

- *BLOCKINFO*
- *PARAMATTR_BLOCK*
- *PARAMATTR_GROUP_BLOCK*
- *TYPE_BLOCK*
- *VALUE_SYMTAB_BLOCK*
- *CONSTANTS_BLOCK*
- *FUNCTION_BLOCK*
- *METADATA_BLOCK*

MODULE_CODE_VERSION Record

[VERSION, version#]

The `VERSION` record (code 1) contains a single value indicating the format version. Versions 0, 1 and 2 are supported at this time. The difference between version 0 and 1 is in the encoding of instruction operands in each *FUNCTION_BLOCK*.

In version 0, each value defined by an instruction is assigned an ID unique to the function. Function-level value IDs are assigned starting from `NumModuleValues` since they share the same namespace as module-level values. The value enumerator resets after each function. When a value is an operand of an instruction, the value ID is used to represent the operand. For large functions or large modules, these operand values can be large.

The encoding in version 1 attempts to avoid large operand values in common cases. Instead of using the value ID directly, operands are encoded as relative to the current instruction. Thus, if an operand is the value defined by the previous instruction, the operand will be encoded as 1.

For example, instead of

```
#n = load #n-1
#n+1 = icmp eq #n, #const0
br #n+1, label #(bb1), label #(bb2)
```

version 1 will encode the instructions as

```
#n = load #1
#n+1 = icmp eq #1, (#n+1)-#const0
br #1, label #(bb1), label #(bb2)
```

Note in the example that operands which are constants also use the relative encoding, while operands like basic block labels do not use the relative encoding.

Forward references will result in a negative value. This can be inefficient, as operands are normally encoded as unsigned VBRs. However, forward references are rare, except in the case of phi instructions. For phi instructions, operands are encoded as *Signed VBRs* to deal with forward references.

In version 2, the meaning of module records `FUNCTION`, `GLOBALVAR`, `ALIAS`, `IFUNC` and `COMDAT` change such that the first two operands specify an offset and size of a string in a string table (see *STRTAB_BLOCK Contents*), the function name is removed from the `FNENTRY` record in the value symbol table, and the top-level `VALUE_SYMTAB_BLOCK` may only contain `FNENTRY` records.

MODULE_CODE_TRIPLE Record

```
[TRIPLE, ...string...]
```

The `TRIPLE` record (code 2) contains a variable number of values representing the bytes of the `target triple` specification string.

MODULE_CODE_DATA_LAYOUT Record

```
[DATA_LAYOUT, ...string...]
```

The `DATA_LAYOUT` record (code 3) contains a variable number of values representing the bytes of the `target datalayout` specification string.

MODULE_CODE_ASM Record

```
[ASM, ...string...]
```

The `ASM` record (code 4) contains a variable number of values representing the bytes of `module asm` strings, with individual assembly blocks separated by newline (ASCII 10) characters.

MODULE_CODE_SECTIONNAME Record

```
[SECTIONNAME, ...string...]
```

The `SECTIONNAME` record (code 5) contains a variable number of values representing the bytes of a single section name string. There should be one `SECTIONNAME` record for each section name referenced (e.g., in global variable or function `section` attributes) within the module. These records can be referenced by the 1-based index in the *section* fields of `GLOBALVAR` or `FUNCTION` records.

MODULE_CODE_DEPLIB Record

```
[DEPLIB, ...string...]
```

The `DEPLIB` record (code 6) contains a variable number of values representing the bytes of a single dependent library name string, one of the libraries mentioned in a `deplibs` declaration. There should be one `DEPLIB` record for each library name referenced.

MODULE_CODE_GLOBALVAR Record

[GLOBALVAR, strtab offset, strtab size, pointer type, isconst, initid, linkage, alignment, section, visibility, threadlocal, unnamed_addr, externally_initialized, dllstorageclass, comdat]

The GLOBALVAR record (code 7) marks the declaration or definition of a global variable. The operand fields are:

- *strtab offset, strtab size*: Specifies the name of the global variable. See *STRTAB_BLOCK Contents*.
- *pointer type*: The type index of the pointer type used to point to this global variable
- *isconst*: Non-zero if the variable is treated as constant within the module, or zero if it is not
- *initid*: If non-zero, the value index of the initializer for this variable, plus 1.
- *linkage*: An encoding of the linkage type for this variable:
 - external: code 0
 - weak: code 1
 - appending: code 2
 - internal: code 3
 - linkonce: code 4
 - dllimport: code 5
 - dllexport: code 6
 - extern_weak: code 7
 - common: code 8
 - private: code 9
 - weak_odr: code 10
 - linkonce_odr: code 11
 - available_externally: code 12
 - deprecated : code 13
 - deprecated : code 14
- *alignment**: The logarithm base 2 of the variable's requested alignment, plus 1
- *section*: If non-zero, the 1-based section index in the table of *MODULE_CODE_SECTIONNAME* entries.
- *visibility*: If present, an encoding of the visibility of this variable:
 - default: code 0
 - hidden: code 1
 - protected: code 2
- *threadlocal*: If present, an encoding of the thread local storage mode of the variable:
 - not thread local: code 0
 - thread local; default TLS model: code 1
 - localdynamic: code 2
 - initialexec: code 3

- localexec: code 4
- *unnamed_addr*: If present, an encoding of the *unnamed_addr* attribute of this variable:
 - not *unnamed_addr*: code 0
 - *unnamed_addr*: code 1
 - *local_unnamed_addr*: code 2
- *dllstorageclass*: If present, an encoding of the DLL storage class of this variable:
 - default: code 0
 - *dllimport*: code 1
 - *dlexport*: code 2
- *comdat*: An encoding of the COMDAT of this function

MODULE_CODE_FUNCTION Record

[FUNCTION, strtab offset, strtab size, type, callingconv, isproto, linkage, paramattr, alignment, section, visibility, gc, prologuedata, dllstorageclass, comdat, prefixdata, personalityfn]

The FUNCTION record (code 8) marks the declaration or definition of a function. The operand fields are:

- *strtab offset*, *strtab size*: Specifies the name of the function. See [STRTAB_BLOCK Contents](#).
- *type*: The type index of the function type describing this function
- *callingconv*: The calling convention number: * ccc: code 0 * fastcc: code 8 * coldcc: code 9 * webkit_jscc: code 12 * anyregcc: code 13 * preserve_mostcc: code 14 * preserve_allcc: code 15 * swiftcc : code 16 * cxx_fast_tlscc: code 17 * x86_stdcallcc: code 64 * x86_fastcallcc: code 65 * arm_apcsc: code 66 * arm_aapcsc: code 67 * arm_aapcs_vfpcc: code 68
- *isproto**: Non-zero if this entry represents a declaration rather than a definition
- *linkage*: An encoding of the [linkage type](#) for this function
- *paramattr*: If nonzero, the 1-based parameter attribute index into the table of [PARAMATTR_CODE_ENTRY](#) entries.
- *alignment*: The logarithm base 2 of the function's requested alignment, plus 1
- *section*: If non-zero, the 1-based section index in the table of [MODULE_CODE_SECTIONNAME](#) entries.
- *visibility*: An encoding of the [visibility](#) of this function
- *gc*: If present and nonzero, the 1-based garbage collector index in the table of [MODULE_CODE_GCNAME](#) entries.
- *unnamed_addr*: If present, an encoding of the [unnamed_addr](#) attribute of this function
- *prologuedata*: If non-zero, the value index of the prologue data for this function, plus 1.
- *dllstorageclass*: An encoding of the [dllstorageclass](#) of this function
- *comdat*: An encoding of the COMDAT of this function
- *prefixdata*: If non-zero, the value index of the prefix data for this function, plus 1.
- *personalityfn*: If non-zero, the value index of the personality function for this function, plus 1.

MODULE_CODE_ALIAS Record

[ALIAS, strtab offset, strtab size, alias type, aliasee val#, linkage, visibility, dllstorageclass, threadlocal, unnamed_addr]

The ALIAS record (code 9) marks the definition of an alias. The operand fields are

- *strtab offset, strtab size*: Specifies the name of the alias. See *STRTAB_BLOCK Contents*.
- *alias type*: The type index of the alias
- *aliasee val#*: The value index of the aliased value
- *linkage*: An encoding of the *linkage type* for this alias
- *visibility*: If present, an encoding of the *visibility* of the alias
- *dllstorageclass*: If present, an encoding of the *dllstorageclass* of the alias
- *threadlocal*: If present, an encoding of the *thread local property* of the alias
- *unnamed_addr*: If present, an encoding of the *unnamed_addr* attribute of this alias

MODULE_CODE_GCNAME Record

[GCNAME, ...string...]

The GCNAME record (code 11) contains a variable number of values representing the bytes of a single garbage collector name string. There should be one GCNAME record for each garbage collector name referenced in function *gc* attributes within the module. These records can be referenced by 1-based index in the *gc* fields of FUNCTION records.

PARAMATTR_BLOCK Contents

The PARAMATTR_BLOCK block (id 9) contains a table of entries describing the attributes of function parameters. These entries are referenced by 1-based index in the *paramattr* field of module block *FUNCTION* records, or within the *attr* field of function block INST_INVOKE and INST_CALL records.

Entries within PARAMATTR_BLOCK are constructed to ensure that each is unique (i.e., no two indices represent equivalent attribute lists).

PARAMATTR_CODE_ENTRY Record

[ENTRY, attrgrp0, attrgrp1, ...]

The ENTRY record (code 2) contains a variable number of values describing a unique set of function parameter attributes. Each *attrgrp* value is used as a key with which to look up an entry in the attribute group table described in the PARAMATTR_GROUP_BLOCK block.

PARAMATTR_CODE_ENTRY_OLD Record

Note: This is a legacy encoding for attributes, produced by LLVM versions 3.2 and earlier. It is guaranteed to be understood by the current LLVM version, as specified in the *IR Backwards Compatibility* policy.

[ENTRY, paramidx0, attr0, paramidx1, attr1...]

The `ENTRY` record (code 1) contains an even number of values describing a unique set of function parameter attributes. Each *paramidx* value indicates which set of attributes is represented, with 0 representing the return value attributes, 0xFFFFFFFF representing function attributes, and other values representing 1-based function parameters. Each *attr* value is a bitmap with the following interpretation:

- bit 0: zeroext
- bit 1: signext
- bit 2: noreturn
- bit 3: inreg
- bit 4: sret
- bit 5: nounwind
- bit 6: noalias
- bit 7: byval
- bit 8: nest
- bit 9: readnone
- bit 10: readonly
- bit 11: noinline
- bit 12: alwaysinline
- bit 13: optsize
- bit 14: ssp
- bit 15: sspreq
- bits 16-31: align *n*
- bit 32: nocapture
- bit 33: noredzone
- bit 34: noimplicitfloat
- bit 35: naked
- bit 36: inlinehint
- bits 37-39: alignstack *n*, represented as the logarithm base 2 of the requested alignment, plus 1

PARAMATTR_GROUP_BLOCK Contents

The `PARAMATTR_GROUP_BLOCK` block (id 10) contains a table of entries describing the attribute groups present in the module. These entries can be referenced within `PARAMATTR_CODE_ENTRY` entries.

PARAMATTR_GRP_CODE_ENTRY Record

```
[ENTRY, grpid, paramidx, attr0, attr1, ...]
```

The `ENTRY` record (code 3) contains *grpid* and *paramidx* values, followed by a variable number of values describing a unique group of attributes. The *grpid* value is a unique key for the attribute group, which can be referenced within `PARAMATTR_CODE_ENTRY` entries. The *paramidx* value indicates which set of attributes is represented, with 0

representing the return value attributes, 0xFFFFFFFF representing function attributes, and other values representing 1-based function parameters.

Each *attr* is itself represented as a variable number of values:

```
kind, key [, ...], [value [, ...]]
```

Each attribute is either a well-known LLVM attribute (possibly with an integer value associated with it), or an arbitrary string (possibly with an arbitrary string value associated with it). The *kind* value is an integer code distinguishing between these possibilities:

- code 0: well-known attribute
- code 1: well-known attribute with an integer value
- code 3: string attribute
- code 4: string attribute with a string value

For well-known attributes (code 0 or 1), the *key* value is an integer code identifying the attribute. For attributes with an integer argument (code 1), the *value* value indicates the argument.

For string attributes (code 3 or 4), the *key* value is actually a variable number of values representing the bytes of a null-terminated string. For attributes with a string argument (code 4), the *value* value is similarly a variable number of values representing the bytes of a null-terminated string.

The integer codes are mapped to well-known attributes as follows.

- code 1: `align(<n>)`
- code 2: `alwaysinline`
- code 3: `byval`
- code 4: `inlinehint`
- code 5: `inreg`
- code 6: `minsize`
- code 7: `naked`
- code 8: `nest`
- code 9: `noalias`
- code 10: `nobuiltin`
- code 11: `nocapture`
- code 12: `noduplicates`
- code 13: `noimplicitfloat`
- code 14: `noinline`
- code 15: `nonlazybind`
- code 16: `noredzone`
- code 17: `noreturn`
- code 18: `nounwind`
- code 19: `optsize`
- code 20: `readnone`
- code 21: `readonly`

- code 22: returned
- code 23: returns_twice
- code 24: signext
- code 25: alignstack(<n>)
- code 26: ssp
- code 27: sspreq
- code 28: sspstrong
- code 29: sret
- code 30: sanitize_address
- code 31: sanitize_thread
- code 32: sanitize_memory
- code 33: uwtable
- code 34: zeroext
- code 35: builtin
- code 36: cold
- code 37: optnone
- code 38: inalloca
- code 39: nonnull
- code 40: jumpable
- code 41: dereferenceable(<n>)
- code 42: dereferenceable_or_null(<n>)
- code 43: convergent
- code 44: safestack
- code 45: argmemonly
- code 46: swiftself
- code 47: swifterror
- code 48: norecurse
- code 49: inaccessiblememonly
- code 50: inaccessiblememonly_or_argmemonly
- code 51: allocsize(<EltSizeParam>[, <NumEltsParam>])
- code 52: writeonly

Note: The `allocsize` attribute has a special encoding for its arguments. Its two arguments, which are 32-bit integers, are packed into one 64-bit integer value (i.e. `(EltSizeParam << 32) | NumEltsParam`), with `NumEltsParam` taking on the sentinel value -1 if it is not specified.

TYPE_BLOCK Contents

The `TYPE_BLOCK` block (id 17) contains records which constitute a table of type operator entries used to represent types referenced within an LLVM module. Each record (with the exception of *NUMENTRY*) generates a single type table entry, which may be referenced by 0-based index from instructions, constants, metadata, type symbol table entries, or other type operator records.

Entries within `TYPE_BLOCK` are constructed to ensure that each entry is unique (i.e., no two indices represent structurally equivalent types).

TYPE_CODE_NUMENTRY Record

[NUMENTRY, numentries]

The `NUMENTRY` record (code 1) contains a single value which indicates the total number of type code entries in the type table of the module. If present, `NUMENTRY` should be the first record in the block.

TYPE_CODE_VOID Record

[VOID]

The `VOID` record (code 2) adds a `void` type to the type table.

TYPE_CODE_HALF Record

[HALF]

The `HALF` record (code 10) adds a `half` (16-bit floating point) type to the type table.

TYPE_CODE_FLOAT Record

[FLOAT]

The `FLOAT` record (code 3) adds a `float` (32-bit floating point) type to the type table.

TYPE_CODE_DOUBLE Record

[DOUBLE]

The `DOUBLE` record (code 4) adds a `double` (64-bit floating point) type to the type table.

TYPE_CODE_LABEL Record

[LABEL]

The `LABEL` record (code 5) adds a `label` type to the type table.

TYPE_CODE_OPAQUE Record

[OPAQUE]

The OPAQUE record (code 6) adds an opaque type to the type table, with a name defined by a previously encountered STRUCT_NAME record. Note that distinct opaque types are not unified.

TYPE_CODE_INTEGER Record

[INTEGER, width]

The INTEGER record (code 7) adds an integer type to the type table. The single *width* field indicates the width of the integer type.

TYPE_CODE_POINTER Record

[POINTER, pointee type, address space]

The POINTER record (code 8) adds a pointer type to the type table. The operand fields are

- *pointee type*: The type index of the pointed-to type
- *address space*: If supplied, the target-specific numbered address space where the pointed-to object resides. Otherwise, the default address space is zero.

TYPE_CODE_FUNCTION_OLD Record

Note: This is a legacy encoding for functions, produced by LLVM versions 3.0 and earlier. It is guaranteed to be understood by the current LLVM version, as specified in the *IR Backwards Compatibility* policy.

[FUNCTION_OLD, vararg, ignored, retty, ...paramty...]

The FUNCTION_OLD record (code 9) adds a function type to the type table. The operand fields are

- *vararg*: Non-zero if the type represents a varargs function
- *ignored*: This value field is present for backward compatibility only, and is ignored
- *retty*: The type index of the function's return type
- *paramty*: Zero or more type indices representing the parameter types of the function

TYPE_CODE_ARRAY Record

[ARRAY, numelts, eltty]

The ARRAY record (code 11) adds an array type to the type table. The operand fields are

- *numelts*: The number of elements in arrays of this type
- *eltty*: The type index of the array element type

TYPE_CODE_VECTOR Record

[VECTOR, numelts, eltty]

The VECTOR record (code 12) adds a vector type to the type table. The operand fields are

- *numelts*: The number of elements in vectors of this type
- *eltty*: The type index of the vector element type

TYPE_CODE_X86_FP80 Record

[X86_FP80]

The X86_FP80 record (code 13) adds an x86_fp80 (80-bit floating point) type to the type table.

TYPE_CODE_FP128 Record

[FP128]

The FP128 record (code 14) adds an fp128 (128-bit floating point) type to the type table.

TYPE_CODE_PPC_FP128 Record

[PPC_FP128]

The PPC_FP128 record (code 15) adds a ppc_fp128 (128-bit floating point) type to the type table.

TYPE_CODE_METADATA Record

[METADATA]

The METADATA record (code 16) adds a metadata type to the type table.

TYPE_CODE_X86_MMX Record

[X86_MMX]

The X86_MMX record (code 17) adds an x86_mmx type to the type table.

TYPE_CODE_STRUCT_ANON Record

[STRUCT_ANON, ispacked, ...eltty...]

The STRUCT_ANON record (code 18) adds a literal struct type to the type table. The operand fields are

- *ispacked*: Non-zero if the type represents a packed structure
- *eltty*: Zero or more type indices representing the element types of the structure

TYPE_CODE_STRUCT_NAME Record

[STRUCT_NAME, ...string...]

The STRUCT_NAME record (code 19) contains a variable number of values representing the bytes of a struct name. The next OPAQUE or STRUCT_NAMED record will use this name.

TYPE_CODE_STRUCT_NAMED Record

[STRUCT_NAMED, ispacked, ...elty...]

The STRUCT_NAMED record (code 20) adds an identified struct type to the type table, with a name defined by a previously encountered STRUCT_NAME record. The operand fields are

- *ispacked*: Non-zero if the type represents a packed structure
- *elty*: Zero or more type indices representing the element types of the structure

TYPE_CODE_FUNCTION Record

[FUNCTION, vararg, retty, ...paramty...]

The FUNCTION record (code 21) adds a function type to the type table. The operand fields are

- *vararg*: Non-zero if the type represents a varargs function
- *retty*: The type index of the function's return type
- *paramty*: Zero or more type indices representing the parameter types of the function

CONSTANTS_BLOCK Contents

The CONSTANTS_BLOCK block (id 11) ...

FUNCTION_BLOCK Contents

The FUNCTION_BLOCK block (id 12) ...

In addition to the record types described below, a FUNCTION_BLOCK block may contain the following sub-blocks:

- *CONSTANTS_BLOCK*
- *VALUE_SYMTAB_BLOCK*
- *METADATA_ATTACHMENT*

VALUE_SYMTAB_BLOCK Contents

The VALUE_SYMTAB_BLOCK block (id 14) ...

METADATA_BLOCK Contents

The METADATA_BLOCK block (id 15) ...

METADATA_ATTACHMENT Contents

The METADATA_ATTACHMENT block (id 16) ...

STRTAB_BLOCK Contents

The STRTAB block (id 23) contains a single record (STRTAB_BLOB, id 1) with a single blob operand containing the bitcode file's string table.

Strings in the string table are not null terminated. A record's *strtab offset* and *strtab size* operands specify the byte offset and size of a string within the string table.

The string table is used by all preceding blocks in the bitcode file that are not succeeded by another intervening STRTAB block. Normally a bitcode file will have a single string table, but it may have more than one if it was created by binary concatenation of multiple bitcode files.

4.4 LLVM Block Frequency Terminology

- *Introduction*
- *Branch Probability*
- *Branch Weight*
- *Block Frequency*
- *Implementation: a series of DAGs*
- *Block Mass*
- *Loop Scale*
- *Implementation: Getting from mass and scale to frequency*
- *Block Bias*

4.4.1 Introduction

Block Frequency is a metric for estimating the relative frequency of different basic blocks. This document describes the terminology that the `BlockFrequencyInfo` and `MachineBlockFrequencyInfo` analysis passes use.

4.4.2 Branch Probability

Blocks with multiple successors have probabilities associated with each outgoing edge. These are called branch probabilities. For a given block, the sum of its outgoing branch probabilities should be 1.0.

4.4.3 Branch Weight

Rather than storing fractions on each edge, we store an integer weight. Weights are relative to the other edges of a given predecessor block. The branch probability associated with a given edge is its own weight divided by the sum of the weights on the predecessor's outgoing edges.

For example, consider this IR:

```
define void @foo() {  
    / ...  
    A:  
        br i1 %cond, label %B, label %C, !prof !0  
    / ...  
}  
!0 = metadata !{metadata !"branch_weights", i32 7, i32 8}
```

and this simple graph representation:

```
A -> B (edge-weight: 7)  
A -> C (edge-weight: 8)
```

The probability of branching from block A to block B is 7/15, and the probability of branching from block A to block C is 8/15.

See *LLVM Branch Weight Metadata* for details about the branch weight IR representation.

4.4.4 Block Frequency

Block frequency is a relative metric that represents the number of times a block executes. The ratio of a block frequency to the entry block frequency is the expected number of times the block will execute per entry to the function.

Block frequency is the main output of the `BlockFrequencyInfo` and `MachineBlockFrequencyInfo` analysis passes.

4.4.5 Implementation: a series of DAGs

The implementation of the block frequency calculation analyses each loop, bottom-up, ignoring backedges; i.e., as a DAG. After each loop is processed, it's packaged up to act as a pseudo-node in its parent loop's (or the function's) DAG analysis.

4.4.6 Block Mass

For each DAG, the entry node is assigned a mass of `UINT64_MAX` and mass is distributed to successors according to branch weights. Block Mass uses a fixed-point representation where `UINT64_MAX` represents 1.0 and 0 represents a number just above 0.0.

After mass is fully distributed, in any cut of the DAG that separates the exit nodes from the entry node, the sum of the block masses of the nodes succeeded by a cut edge should equal `UINT64_MAX`. In other words, mass is conserved as it “falls” through the DAG.

If a function's basic block graph is a DAG, then block masses are valid block frequencies. This works poorly in practise though, since downstream users rely on adding block frequencies together without hitting the maximum.

4.4.7 Loop Scale

Loop scale is a metric that indicates how many times a loop iterates per entry. As mass is distributed through the loop's DAG, the (otherwise ignored) backedge mass is collected. This backedge mass is used to compute the exit frequency, and thus the loop scale.

4.4.8 Implementation: Getting from mass and scale to frequency

After analysing the complete series of DAGs, each block has a mass (local to its containing loop, if any), and each loop pseudo-node has a loop scale and its own mass (from its parent's DAG).

We can get an initial frequency assignment (with entry frequency of 1.0) by multiplying these masses and loop scales together. A given block's frequency is the product of its mass, the mass of containing loops' pseudo nodes, and the containing loops' loop scales.

Since downstream users need integers (not floating point), this initial frequency assignment is shifted as necessary into the range of `uint64_t`.

4.4.9 Block Bias

Block bias is a proposed *absolute* metric to indicate a bias toward or away from a given block during a function's execution. The idea is that bias can be used in isolation to indicate whether a block is relatively hot or cold, or to compare two blocks to indicate whether one is hotter or colder than the other.

The proposed calculation involves calculating a *reference* block frequency, where:

- every branch weight is assumed to be 1 (i.e., every branch probability distribution is even) and
- loop scales are ignored.

This reference frequency represents what the block frequency would be in an unbiased graph.

The bias is the ratio of the block frequency to this reference block frequency.

4.5 LLVM Branch Weight Metadata

- *Introduction*
- *Supported Instructions*
 - *BranchInst*
 - *SwitchInst*
 - *IndirectBrInst*
 - *CallInst*
 - *Other*
- *Built-in expect Instructions*
 - *if statement*
 - *switch statement*
- *CFG Modifications*
- *Function Entry Counts*

4.5.1 Introduction

Branch Weight Metadata represents branch weights as its likeliness to be taken (see *LLVM Block Frequency Terminology*). Metadata is assigned to the `TerminatorInst` as a `MDNode` of the `MD_prof` kind. The first operator is always a `MDString` node with the string “branch_weights”. Number of operators depends on the terminator type.

Branch weights might be fetch from the profiling file, or generated based on `__builtin_expect` instruction.

All weights are represented as an unsigned 32-bit values, where higher value indicates greater chance to be taken.

4.5.2 Supported Instructions

BranchInst

Metadata is only assigned to the conditional branches. There are two extra operands for the true and the false branch.

```
!0 = metadata !{
  metadata !"branch_weights",
  i32 <TRUE_BRANCH_WEIGHT>,
  i32 <FALSE_BRANCH_WEIGHT>
}
```

SwitchInst

Branch weights are assigned to every case (including the default case which is always case #0).

```
!0 = metadata !{
  metadata !"branch_weights",
  i32 <DEFAULT_BRANCH_WEIGHT>
  [ , i32 <CASE_BRANCH_WEIGHT> ... ]
}
```

IndirectBrInst

Branch weights are assigned to every destination.

```
!0 = metadata !{
  metadata !"branch_weights",
  i32 <LABEL_BRANCH_WEIGHT>
  [ , i32 <LABEL_BRANCH_WEIGHT> ... ]
}
```

CallInst

Calls may have branch weight metadata, containing the execution count of the call. It is currently used in SamplePGO mode only, to augment the block and entry counts which may not be accurate with sampling.

```
!0 = metadata !{
  metadata !"branch_weights",
  i32 <CALL_BRANCH_WEIGHT>
}
```

Other

Other terminator instructions are not allowed to contain Branch Weight Metadata.

4.5.3 Built-in expect Instructions

`__builtin_expect(long exp, long c)` instruction provides branch prediction information. The return value is the value of `exp`.

It is especially useful in conditional statements. Currently Clang supports two conditional statements:

if statement

The `exp` parameter is the condition. The `c` parameter is the expected comparison value. If it is equal to 1 (true), the condition is likely to be true, in other case condition is likely to be false. For example:

```
if (__builtin_expect(x > 0, 1)) {
    // This block is likely to be taken.
}
```

switch statement

The `exp` parameter is the value. The `c` parameter is the expected value. If the expected value doesn't show on the cases list, the `default` case is assumed to be likely taken.

```
switch (__builtin_expect(x, 5)) {
default: break;
case 0: // ...
case 3: // ...
case 5: // This case is likely to be taken.
}
```

4.5.4 CFG Modifications

Branch Weight Metadata is not proof against CFG changes. If terminator operands' are changed some action should be taken. In other case some misoptimizations may occur due to incorrect branch prediction information.

4.5.5 Function Entry Counts

To allow comparing different functions during inter-procedural analysis and optimization, `MD_prof` nodes can also be assigned to a function definition. The first operand is a string indicating the name of the associated counter.

Currently, one counter is supported: "function_entry_count". The second operand is a 64-bit counter that indicates the number of times that this function was invoked (in the case of instrumentation-based profiles). In the case of sampling-based profiles, this operand is an approximation of how many times the function was invoked.

For example, in the code below, the instrumentation for function `foo()` indicates that it was called 2,590 times at runtime.

```
define i32 @foo() !prof !1 {  
    ret i32 0  
}  
!1 = !{"function_entry_count", i64 2590}
```

If “function_entry_count” has more than 2 operands, the later operands are the GUID of the functions that needs to be imported by ThinLTO. This is only set by sampling based profile. It is needed because the sampling based profile was collected on a binary that had already imported and inlined these functions, and we need to ensure the IR matches in the ThinLTO backends for profile annotation. The reason why we cannot annotate this on the callsite is that it can only go down 1 level in the call chain. For the cases where `foo_in_a_cc()->bar_in_b_cc()->baz_in_c_cc()`, we will need to go down 2 levels in the call chain to import both `bar_in_b_cc` and `baz_in_c_cc`.

4.6 LLVM bugpoint tool: design and usage

- *Description*
- *Design Philosophy*
 - *Automatic Debugger Selection*
 - *Crash debugger*
 - *Code generator debugger*
 - *Miscompilation debugger*
- *Advice for using bugpoint*
- *What to do when bugpoint isn't enough*

4.6.1 Description

`bugpoint` narrows down the source of problems in LLVM tools and passes. It can be used to debug three types of failures: optimizer crashes, miscompilations by optimizers, or bad native code generation (including problems in the static and JIT compilers). It aims to reduce large test cases to small, useful ones. For example, if `opt` crashes while optimizing a file, it will identify the optimization (or combination of optimizations) that causes the crash, and reduce the file down to a small example which triggers the crash.

For detailed case scenarios, such as debugging `opt`, or one of the LLVM code generators, see [How to submit an LLVM bug report](#).

4.6.2 Design Philosophy

`bugpoint` is designed to be a useful tool without requiring any hooks into the LLVM infrastructure at all. It works with any and all LLVM passes and code generators, and does not need to “know” how they work. Because of this, it may appear to do stupid things or miss obvious simplifications. `bugpoint` is also designed to trade off programmer time for computer time in the compiler-debugging process; consequently, it may take a long period of (unattended) time to reduce a test case, but we feel it is still worth it. Note that `bugpoint` is generally very quick unless debugging a miscompilation where each test of the program (which requires executing it) takes a long time.

Automatic Debugger Selection

`bugpoint` reads each `.bc` or `.ll` file specified on the command line and links them together into a single module, called the test program. If any LLVM passes are specified on the command line, it runs these passes on the test program. If any of the passes crash, or if they produce malformed output (which causes the verifier to abort), `bugpoint` starts the *crash debugger*.

Otherwise, if the `-output` option was not specified, `bugpoint` runs the test program with the “safe” backend (which is assumed to generate good code) to generate a reference output. Once `bugpoint` has a reference output for the test program, it tries executing it with the selected code generator. If the selected code generator crashes, `bugpoint` starts the *crash debugger* on the code generator. Otherwise, if the resulting output differs from the reference output, it assumes the difference resulted from a code generator failure, and starts the *code generator debugger*.

Finally, if the output of the selected code generator matches the reference output, `bugpoint` runs the test program after all of the LLVM passes have been applied to it. If its output differs from the reference output, it assumes the difference resulted from a failure in one of the LLVM passes, and enters the *miscompilation debugger*. Otherwise, there is no problem `bugpoint` can debug.

Crash debugger

If an optimizer or code generator crashes, `bugpoint` will try as hard as it can to reduce the list of passes (for optimizer crashes) and the size of the test program. First, `bugpoint` figures out which combination of optimizer passes triggers the bug. This is useful when debugging a problem exposed by `opt`, for example, because it runs over 38 passes.

Next, `bugpoint` tries removing functions from the test program, to reduce its size. Usually it is able to reduce a test program to a single function, when debugging intraprocedural optimizations. Once the number of functions has been reduced, it attempts to delete various edges in the control flow graph, to reduce the size of the function as much as possible. Finally, `bugpoint` deletes any individual LLVM instructions whose absence does not eliminate the failure. At the end, `bugpoint` should tell you what passes crash, give you a bitcode file, and give you instructions on how to reproduce the failure with `opt` or `llc`.

Code generator debugger

The code generator debugger attempts to narrow down the amount of code that is being miscompiled by the selected code generator. To do this, it takes the test program and partitions it into two pieces: one piece which it compiles with the “safe” backend (into a shared object), and one piece which it runs with either the JIT or the static LLC compiler. It uses several techniques to reduce the amount of code pushed through the LLVM code generator, to reduce the potential scope of the problem. After it is finished, it emits two bitcode files (called “test” [to be compiled with the code generator] and “safe” [to be compiled with the “safe” backend], respectively), and instructions for reproducing the problem. The code generator debugger assumes that the “safe” backend produces good code.

Miscompilation debugger

The miscompilation debugger works similarly to the code generator debugger. It works by splitting the test program into two pieces, running the optimizations specified on one piece, linking the two pieces back together, and then executing the result. It attempts to narrow down the list of passes to the one (or few) which are causing the miscompilation, then reduce the portion of the test program which is being miscompiled. The miscompilation debugger assumes that the selected code generator is working properly.

4.6.3 Advice for using bugpoint

`bugpoint` can be a remarkably useful tool, but it sometimes works in non-obvious ways. Here are some hints and tips:

- In the code generator and miscompilation debuggers, `bugpoint` only works with programs that have deterministic output. Thus, if the program outputs `argv[0]`, the date, time, or any other “random” data, `bugpoint` may misinterpret differences in these data, when output, as the result of a miscompilation. Programs should be temporarily modified to disable outputs that are likely to vary from run to run.
- In the code generator and miscompilation debuggers, debugging will go faster if you manually modify the program or its inputs to reduce the runtime, but still exhibit the problem.
- `bugpoint` is extremely useful when working on a new optimization: it helps track down regressions quickly. To avoid having to relink `bugpoint` every time you change your optimization however, have `bugpoint` dynamically load your optimization with the `-load` option.
- `bugpoint` can generate a lot of output and run for a long period of time. It is often useful to capture the output of the program to file. For example, in the C shell, you can run:

```
$ bugpoint ... |& tee bugpoint.log
```

to get a copy of `bugpoint`’s output in the file `bugpoint.log`, as well as on your terminal.

- `bugpoint` cannot debug problems with the LLVM linker. If `bugpoint` crashes before you see its “All input ok” message, you might try `llvm-link -v` on the same set of input files. If that also crashes, you may be experiencing a linker bug.
- `bugpoint` is useful for proactively finding bugs in LLVM. Invoking `bugpoint` with the `-find-bugs` option will cause the list of specified optimizations to be randomized and applied to the program. This process will repeat until a bug is found or the user kills `bugpoint`.

4.6.4 What to do when `bugpoint` isn’t enough

Sometimes, `bugpoint` is not enough. In particular, `InstCombine` and `TargetLowering` both have visitor structured code with lots of potential transformations. If the process of using `bugpoint` has left you with still too much code to figure out and the problem seems to be in `instcombine`, the following steps may help. These same techniques are useful with `TargetLowering` as well.

Turn on `-debug-only=instcombine` and see which transformations within `instcombine` are firing by selecting out lines with “IC” in them.

At this point, you have a decision to make. Is the number of transformations small enough to step through them using a debugger? If so, then try that.

If there are too many transformations, then a source modification approach may be helpful. In this approach, you can modify the source code of `instcombine` to disable just those transformations that are being performed on your test input and perform a binary search over the set of transformations. One set of places to modify are the “`visit*`” methods of `InstCombiner` (e.g. `visitICmpInst`) by adding a “`return false`” as the first line of the method.

If that still doesn’t remove enough, then change the caller of `InstCombiner::DoOneIteration`, `InstCombiner::runOnFunction` to limit the number of iterations.

You may also find it useful to use “`-stats`” now to see what parts of `instcombine` are firing. This can guide where to put additional reporting code.

At this point, if the amount of transformations is still too large, then inserting code to limit whether or not to execute the body of the code in the visit function can be helpful. Add a static counter which is incremented on every invocation of the function. Then add code which simply returns false on desired ranges. For example:

```
static int calledCount = 0;
calledCount++;
DEBUG(if (calledCount < 212) return false);
DEBUG(if (calledCount > 217) return false);
```



```

DEBUG(if (calledCount == 213) return false);
DEBUG(if (calledCount == 214) return false);
DEBUG(if (calledCount == 215) return false);
DEBUG(if (calledCount == 216) return false);
DEBUG(dbgs() << "visitXOR calledCount: " << calledCount << "\n");
DEBUG(dbgs() << "I: "; I->dump());

```

could be added to `visitXOR` to limit `visitXor` to being applied only to calls 212 and 217. This is from an actual test case and raises an important point—a simple binary search may not be sufficient, as transformations that interact may require isolating more than one call. In `TargetLowering`, use `return SDNode();` instead of `return false;`.

Now that the number of transformations is down to a manageable number, try examining the output to see if you can figure out which transformations are being done. If that can be figured out, then do the usual debugging. If which code corresponds to the transformation being performed isn't obvious, set a breakpoint after the call count based disabling and step through the code. Alternatively, you can use “printf” style debugging to report waypoints.

4.7 The LLVM Target-Independent Code Generator

- *Introduction*
 - *Required components in the code generator*
 - *The high-level design of the code generator*
 - *Using TableGen for target description*
- *Target description classes*
 - *The TargetMachine class*
 - *The DataLayout class*
 - *The TargetLowering class*
 - *The TargetRegisterInfo class*
 - *The TargetInstrInfo class*
 - *The TargetFrameLowering class*
 - *The TargetSubtarget class*
 - *The TargetJITInfo class*
- *Machine code description classes*
 - *The MachineInstr class*
 - * *Using the MachineInstrBuilder.h functions*
 - * *Fixed (preassigned) registers*
 - * *Call-clobbered registers*
 - * *Machine code in SSA form*
 - *The MachineBasicBlock class*
 - *The MachineFunction class*

- *MachineInstr Bundles*
- *The “MC” Layer*
 - *The MCStreamer API*
 - *The MCContext class*
 - *The MCSymbol class*
 - *The MCSection class*
 - *The MCInst class*
- *Target-independent code generation algorithms*
 - *Instruction Selection*
 - * *Introduction to SelectionDAGs*
 - * *SelectionDAG Instruction Selection Process*
 - * *Initial SelectionDAG Construction*
 - * *SelectionDAG LegalizeTypes Phase*
 - * *SelectionDAG Legalize Phase*
 - * *SelectionDAG Optimization Phase: the DAG Combiner*
 - * *SelectionDAG Select Phase*
 - * *SelectionDAG Scheduling and Formation Phase*
 - * *Future directions for the SelectionDAG*
 - *SSA-based Machine Code Optimizations*
 - *Live Intervals*
 - * *Live Variable Analysis*
 - * *Live Intervals Analysis*
 - *Register Allocation*
 - * *How registers are represented in LLVM*
 - * *Mapping virtual registers to physical registers*
 - * *Handling two address instructions*
 - * *The SSA deconstruction phase*
 - * *Instruction folding*
 - * *Built in register allocators*
 - *Prolog/Epilog Code Insertion*
 - *Late Machine Code Optimizations*
 - *Code Emission*
 - *VLIW Packetizer*
 - * *Mapping from instructions to functional units*
 - * *How the packetization tables are generated and used*

- *Implementing a Native Assembler*
 - *Instruction Parsing*
 - *Instruction Alias Processing*
 - * *Mnemonic Aliases*
 - * *Instruction Aliases*
 - *Instruction Matching*
- *Target-specific Implementation Notes*
 - *Target Feature Matrix*
 - * *Is Generally Reliable*
 - * *Assembly Parser*
 - * *Disassembler*
 - * *Inline Asm*
 - * *JIT Support*
 - * *.o File Writing*
 - * *Tail Calls*
 - * *Segmented Stacks*
 - *Tail call optimization*
 - *Sibling call optimization*
 - *The X86 backend*
 - * *X86 Target Triples supported*
 - * *X86 Calling Conventions supported*
 - * *Representing X86 addressing modes in MachineInstrs*
 - * *X86 address spaces supported*
 - * *Instruction naming*
 - *The PowerPC backend*
 - * *LLVM PowerPC ABI*
 - * *Frame Layout*
 - * *Prolog/Epilog*
 - * *Dynamic Allocation*
 - *The NVPTX backend*
 - *The extended Berkeley Packet Filter (eBPF) backend*
 - * *Instruction encoding (arithmetic and jump)*
 - * *Instruction encoding (load, store)*
 - * *Packet data access (BPF_ABS, BPF_IND)*
 - * *eBPF maps*

- * *Function calls*
- * *Program start*
- *The AMDGPU backend*

Warning: This is a work in progress.

4.7.1 Introduction

The LLVM target-independent code generator is a framework that provides a suite of reusable components for translating the LLVM internal representation to the machine code for a specified target—either in assembly form (suitable for a static compiler) or in binary machine code format (usable for a JIT compiler). The LLVM target-independent code generator consists of six main components:

1. *Abstract target description* interfaces which capture important properties about various aspects of the machine, independently of how they will be used. These interfaces are defined in `include/llvm/Target/`.
2. Classes used to represent the *code being generated* for a target. These classes are intended to be abstract enough to represent the machine code for *any* target machine. These classes are defined in `include/llvm/CodeGen/`. At this level, concepts like “constant pool entries” and “jump tables” are explicitly exposed.
3. Classes and algorithms used to represent code at the object file level, the *MC Layer*. These classes represent assembly level constructs like labels, sections, and instructions. At this level, concepts like “constant pool entries” and “jump tables” don’t exist.
4. *Target-independent algorithms* used to implement various phases of native code generation (register allocation, scheduling, stack frame representation, etc). This code lives in `lib/CodeGen/`.
5. *Implementations of the abstract target description interfaces* for particular targets. These machine descriptions make use of the components provided by LLVM, and can optionally provide custom target-specific passes, to build complete code generators for a specific target. Target descriptions live in `lib/Target/`.
6. The target-independent JIT components. The LLVM JIT is completely target independent (it uses the `TargetJITInfo` structure to interface for target-specific issues. The code for the target-independent JIT lives in `lib/ExecutionEngine/JIT`.

Depending on which part of the code generator you are interested in working on, different pieces of this will be useful to you. In any case, you should be familiar with the *target description* and *machine code representation* classes. If you want to add a backend for a new target, you will need to *implement the target description* classes for your new target and understand the *LLVM code representation*. If you are interested in implementing a new *code generation algorithm*, it should only depend on the target-description and machine code representation classes, ensuring that it is portable.

Required components in the code generator

The two pieces of the LLVM code generator are the high-level interface to the code generator and the set of reusable components that can be used to build target-specific backends. The two most important interfaces (*TargetMachine* and *DataLayout*) are the only ones that are required to be defined for a backend to fit into the LLVM system, but the others must be defined if the reusable code generator components are going to be used.

This design has two important implications. The first is that LLVM can support completely non-traditional code generation targets. For example, the C backend does not require register allocation, instruction selection, or any of the other standard components provided by the system. As such, it only implements these two interfaces, and does its own thing. Note that C backend was removed from the trunk since LLVM 3.1 release. Another example of a code

generator like this is a (purely hypothetical) backend that converts LLVM to the GCC RTL form and uses GCC to emit machine code for a target.

This design also implies that it is possible to design and implement radically different code generators in the LLVM system that do not make use of any of the built-in components. Doing so is not recommended at all, but could be required for radically different targets that do not fit into the LLVM machine description model: FPGAs for example.

The high-level design of the code generator

The LLVM target-independent code generator is designed to support efficient and quality code generation for standard register-based microprocessors. Code generation in this model is divided into the following stages:

1. *Instruction Selection* — This phase determines an efficient way to express the input LLVM code in the target instruction set. This stage produces the initial code for the program in the target instruction set, then makes use of virtual registers in SSA form and physical registers that represent any required register assignments due to target constraints or calling conventions. This step turns the LLVM code into a DAG of target instructions.
2. *Scheduling and Formation* — This phase takes the DAG of target instructions produced by the instruction selection phase, determines an ordering of the instructions, then emits the instructions as *MachineInstrs* with that ordering. Note that we describe this in the *instruction selection section* because it operates on a *SelectionDAG*.
3. *SSA-based Machine Code Optimizations* — This optional stage consists of a series of machine-code optimizations that operate on the SSA-form produced by the instruction selector. Optimizations like modulo-scheduling or peephole optimization work here.
4. *Register Allocation* — The target code is transformed from an infinite virtual register file in SSA form to the concrete register file used by the target. This phase introduces spill code and eliminates all virtual register references from the program.
5. *Prolog/Epilog Code Insertion* — Once the machine code has been generated for the function and the amount of stack space required is known (used for LLVM *alloca*'s and spill slots), the prolog and epilog code for the function can be inserted and “abstract stack location references” can be eliminated. This stage is responsible for implementing optimizations like frame-pointer elimination and stack packing.
6. *Late Machine Code Optimizations* — Optimizations that operate on “final” machine code can go here, such as spill code scheduling and peephole optimizations.
7. *Code Emission* — The final stage actually puts out the code for the current function, either in the target assembler format or in machine code.

The code generator is based on the assumption that the instruction selector will use an optimal pattern matching selector to create high-quality sequences of native instructions. Alternative code generator designs based on pattern expansion and aggressive iterative peephole optimization are much slower. This design permits efficient compilation (important for JIT environments) and aggressive optimization (used when generating code offline) by allowing components of varying levels of sophistication to be used for any step of compilation.

In addition to these stages, target implementations can insert arbitrary target-specific passes into the flow. For example, the X86 target uses a special pass to handle the 80x87 floating point stack architecture. Other targets with unusual requirements can be supported with custom passes as needed.

Using TableGen for target description

The target description classes require a detailed description of the target architecture. These target descriptions often have a large amount of common information (e.g., an `add` instruction is almost identical to a `sub` instruction). In order to allow the maximum amount of commonality to be factored out, the LLVM code generator uses the *TableGen* tool to describe big chunks of the target machine, which allows the use of domain-specific and target-specific abstractions to reduce the amount of repetition.

As LLVM continues to be developed and refined, we plan to move more and more of the target description to the `.td` form. Doing so gives us a number of advantages. The most important is that it makes it easier to port LLVM because it reduces the amount of C++ code that has to be written, and the surface area of the code generator that needs to be understood before someone can get something working. Second, it makes it easier to change things. In particular, if tables and other things are all emitted by `tblgen`, we only need a change in one place (`tblgen`) to update all of the targets to a new interface.

4.7.2 Target description classes

The LLVM target description classes (located in the `include/llvm/Target` directory) provide an abstract description of the target machine independent of any particular client. These classes are designed to capture the *abstract* properties of the target (such as the instructions and registers it has), and do not incorporate any particular pieces of code generation algorithms.

All of the target description classes (except the *DataLayout* class) are designed to be subclassed by the concrete target implementation, and have virtual methods implemented. To get to these implementations, the *TargetMachine* class provides accessors that should be implemented by the target.

The TargetMachine class

The *TargetMachine* class provides virtual methods that are used to access the target-specific implementations of the various target description classes via the `get*Info` methods (`getInstrInfo`, `getRegisterInfo`, `getFrameInfo`, etc.). This class is designed to be specialized by a concrete target implementation (e.g., *X86TargetMachine*) which implements the various virtual methods. The only required target description class is the *DataLayout* class, but if the code generator components are to be used, the other interfaces should be implemented as well.

The DataLayout class

The *DataLayout* class is the only required target description class, and it is the only class that is not extensible (you cannot derive a new class from it). *DataLayout* specifies information about how the target lays out memory for structures, the alignment requirements for various data types, the size of pointers in the target, and whether the target is little-endian or big-endian.

The TargetLowering class

The *TargetLowering* class is used by SelectionDAG based instruction selectors primarily to describe how LLVM code should be lowered to SelectionDAG operations. Among other things, this class indicates:

- an initial register class to use for various `ValueTypes`,
- which operations are natively supported by the target machine,
- the return type of `setcc` operations,
- the type to use for shift amounts, and
- various high-level characteristics, like whether it is profitable to turn division by a constant into a multiplication sequence.

The `TargetRegisterInfo` class

The `TargetRegisterInfo` class is used to describe the register file of the target and any interactions between the registers.

Registers are represented in the code generator by unsigned integers. Physical registers (those that actually exist in the target description) are unique small numbers, and virtual registers are generally large. Note that register #0 is reserved as a flag value.

Each register in the processor description has an associated `TargetRegisterDesc` entry, which provides a textual name for the register (used for assembly output and debugging dumps) and a set of aliases (used to indicate whether one register overlaps with another).

In addition to the per-register description, the `TargetRegisterInfo` class exposes a set of processor specific register classes (instances of the `TargetRegisterClass` class). Each register class contains sets of registers that have the same properties (for example, they are all 32-bit integer registers). Each SSA virtual register created by the instruction selector has an associated register class. When the register allocator runs, it replaces virtual registers with a physical register in the set.

The target-specific implementations of these classes is auto-generated from a *TableGen* description of the register file.

The `TargetInstrInfo` class

The `TargetInstrInfo` class is used to describe the machine instructions supported by the target. Descriptions define things like the mnemonic for the opcode, the number of operands, the list of implicit register uses and defs, whether the instruction has certain target-independent properties (accesses memory, is commutable, etc), and holds any target-specific flags.

The `TargetFrameLowering` class

The `TargetFrameLowering` class is used to provide information about the stack frame layout of the target. It holds the direction of stack growth, the known stack alignment on entry to each function, and the offset to the local area. The offset to the local area is the offset from the stack pointer on function entry to the first location where function data (local variables, spill locations) can be stored.

The `TargetSubtarget` class

The `TargetSubtarget` class is used to provide information about the specific chip set being targeted. A sub-target informs code generation of which instructions are supported, instruction latencies and instruction execution itinerary; i.e., which processing units are used, in what order, and for how long.

The `TargetJITInfo` class

The `TargetJITInfo` class exposes an abstract interface used by the Just-In-Time code generator to perform target-specific activities, such as emitting stubs. If a `TargetMachine` supports JIT code generation, it should provide one of these objects through the `getJITInfo` method.

4.7.3 Machine code description classes

At the high-level, LLVM code is translated to a machine specific representation formed out of *MachineFunction*, *MachineBasicBlock*, and *MachineInstr* instances (defined in `include/llvm/CodeGen`). This representation is completely target agnostic, representing instructions in their most abstract form: an opcode and a series of operands.

This representation is designed to support both an SSA representation for machine code, as well as a register allocated, non-SSA form.

The `MachineInstr` class

Target machine instructions are represented as instances of the `MachineInstr` class. This class is an extremely abstract way of representing machine instructions. In particular, it only keeps track of an opcode number and a set of operands.

The opcode number is a simple unsigned integer that only has meaning to a specific backend. All of the instructions for a target should be defined in the `*InstrInfo.td` file for the target. The opcode enum values are auto-generated from this description. The `MachineInstr` class does not have any information about how to interpret the instruction (i.e., what the semantics of the instruction are); for that you must refer to the [TargetInstrInfo](#) class.

The operands of a machine instruction can be of several different types: a register reference, a constant integer, a basic block reference, etc. In addition, a machine operand should be marked as a def or a use of the value (though only registers are allowed to be defs).

By convention, the LLVM code generator orders instruction operands so that all register definitions come before the register uses, even on architectures that are normally printed in other orders. For example, the SPARC add instruction: “add %i1, %i2, %i3” adds the “%i1”, and “%i2” registers and stores the result into the “%i3” register. In the LLVM code generator, the operands should be stored as “%i3, %i1, %i2”: with the destination first.

Keeping destination (definition) operands at the beginning of the operand list has several advantages. In particular, the debugging printer will print the instruction like this:

```
%r3 = add %i1, %i2
```

Also if the first operand is a def, it is easier to *create instructions* whose only def is the first operand.

Using the `MachineInstrBuilder.h` functions

Machine instructions are created by using the `BuildMI` functions, located in the `include/llvm/CodeGen/MachineInstrBuilder.h` file. The `BuildMI` functions make it easy to build arbitrary machine instructions. Usage of the `BuildMI` functions look like this:

```
// Create a 'DestReg = mov 42' (rendered in X86 assembly as 'mov DestReg, 42')
// instruction and insert it at the end of the given MachineBasicBlock.
const TargetInstrInfo &TII = ...
MachineBasicBlock &MBB = ...
DebugLoc DL;
MachineInstr *MI = BuildMI(MBB, DL, TII.get(X86::MOV32ri), DestReg).addImm(42);

// Create the same instr, but insert it before a specified iterator point.
MachineBasicBlock::iterator MBBI = ...
BuildMI(MBB, MBBI, DL, TII.get(X86::MOV32ri), DestReg).addImm(42);

// Create a 'cmp Reg, 0' instruction, no destination reg.
MI = BuildMI(MBB, DL, TII.get(X86::CMP32ri8)).addReg(Reg).addImm(42);

// Create an 'sahf' instruction which takes no operands and stores nothing.
MI = BuildMI(MBB, DL, TII.get(X86::SAHF));

// Create a self looping branch instruction.
BuildMI(MBB, DL, TII.get(X86::JNE)).addMBB(&MBB);
```


If you need to add a definition operand (other than the optional destination register), you must explicitly mark it as such:

```
MI.addReg(Reg, RegState::Define);
```

Fixed (preassigned) registers

One important issue that the code generator needs to be aware of is the presence of fixed registers. In particular, there are often places in the instruction stream where the register allocator *must* arrange for a particular value to be in a particular register. This can occur due to limitations of the instruction set (e.g., the X86 can only do a 32-bit divide with the EAX/EDX registers), or external factors like calling conventions. In any case, the instruction selector should emit code that copies a virtual register into or out of a physical register when needed.

For example, consider this simple LLVM example:

```
define i32 @test(i32 %X, i32 %Y) {
    %Z = sdiv i32 %X, %Y
    ret i32 %Z
}
```

The X86 instruction selector might produce this machine code for the `div` and `ret`:

```
;; Start of div
%EAX = mov %reg1024          ;; Copy X (in reg1024) into EAX
%reg1027 = sar %reg1024, 31
%EDX = mov %reg1027          ;; Sign extend X into EDX
idiv %reg1025                ;; Divide by Y (in reg1025)
%reg1026 = mov %EAX          ;; Read the result (Z) out of EAX

;; Start of ret
%EAX = mov %reg1026          ;; 32-bit return value goes in EAX
ret
```

By the end of code generation, the register allocator would coalesce the registers and delete the resultant identity moves producing the following code:

```
;; X is in EAX, Y is in ECX
mov %EAX, %EDX
sar %EDX, 31
idiv %ECX
ret
```

This approach is extremely general (if it can handle the X86 architecture, it can handle anything!) and allows all of the target specific knowledge about the instruction stream to be isolated in the instruction selector. Note that physical registers should have a short lifetime for good code generation, and all physical registers are assumed dead on entry to and exit from basic blocks (before register allocation). Thus, if you need a value to be live across basic block boundaries, it *must* live in a virtual register.

Call-clobbered registers

Some machine instructions, like calls, clobber a large number of physical registers. Rather than adding `<def, dead>` operands for all of them, it is possible to use an `MO_RegisterMask` operand instead. The register mask operand holds a bit mask of preserved registers, and everything else is considered to be clobbered by the instruction.

Machine code in SSA form

MachineInstr's are initially selected in SSA-form, and are maintained in SSA-form until register allocation happens. For the most part, this is trivially simple since LLVM is already in SSA form; LLVM PHI nodes become machine code PHI nodes, and virtual registers are only allowed to have a single definition.

After register allocation, machine code is no longer in SSA-form because there are no virtual registers left in the code.

The MachineBasicBlock class

The `MachineBasicBlock` class contains a list of machine instructions (*MachineInstr* instances). It roughly corresponds to the LLVM code input to the instruction selector, but there can be a one-to-many mapping (i.e. one LLVM basic block can map to multiple machine basic blocks). The `MachineBasicBlock` class has a “`getBasicBlock`” method, which returns the LLVM basic block that it comes from.

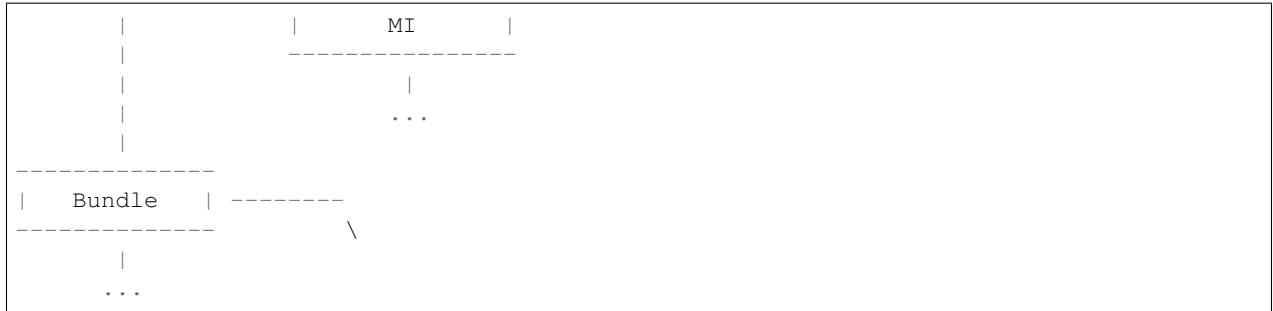
The MachineFunction class

The `MachineFunction` class contains a list of machine basic blocks (*`MachineBasicBlock`* instances). It corresponds one-to-one with the LLVM function input to the instruction selector. In addition to a list of basic blocks, the `MachineFunction` contains a `MachineConstantPool`, a `MachineFrameInfo`, a `MachineFunctionInfo`, and a `MachineRegisterInfo`. See `include/llvm/CodeGen/MachineFunction.h` for more information.

MachineInstr Bundles

LLVM code generator can model sequences of instructions as MachineInstr bundles. A MI bundle can model a VLIW group / pack which contains an arbitrary number of parallel instructions. It can also be used to model a sequential list of instructions (potentially with data dependencies) that cannot be legally separated (e.g. ARM Thumb2 IT blocks).

Conceptually a MI bundle is a MI with a number of other MIs nested within:



MI bundle support does not change the physical representations of `MachineBasicBlock` and `MachineInstr`. All the MIs (including top level and nested ones) are stored as sequential list of MIs. The “bundled” MIs are marked with the ‘`InsideBundle`’ flag. A top level MI with the special `BUNDLE` opcode is used to represent the start of a bundle. It’s legal to mix `BUNDLE` MIs with individual MIs that are not inside bundles nor represent bundles.

`MachineInstr` passes should operate on a MI bundle as a single unit. Member methods have been taught to correctly handle bundles and MIs inside bundles. The `MachineBasicBlock` iterator has been modified to skip over bundled MIs to enforce the bundle-as-a-single-unit concept. An alternative iterator `instr_iterator` has been added to `MachineBasicBlock` to allow passes to iterate over all of the MIs in a `MachineBasicBlock`, including those which are nested inside bundles. The top level `BUNDLE` instruction must have the correct set of register `MachineOperand`’s that represent the cumulative inputs and outputs of the bundled MIs.

Packing / bundling of `MachineInstr`’s should be done as part of the register allocation super-pass. More specifically, the pass which determines what MIs should be bundled together must be done after code generator exits SSA form (i.e. after two-address pass, PHI elimination, and copy coalescing). Bundles should only be finalized (i.e. adding `BUNDLE` MIs and input and output register `MachineOperands`) after virtual registers have been rewritten into physical registers. This requirement eliminates the need to add virtual register operands to `BUNDLE` instructions which would effectively double the virtual register def and use lists.

4.7.4 The “MC” Layer

The MC Layer is used to represent and process code at the raw machine code level, devoid of “high level” information like “constant pools”, “jump tables”, “global variables” or anything like that. At this level, LLVM handles things like label names, machine instructions, and sections in the object file. The code in this layer is used for a number of important purposes: the tail end of the code generator uses it to write a `.s` or `.o` file, and it is also used by the `llvm-mc` tool to implement standalone machine code assemblers and disassemblers.

This section describes some of the important classes. There are also a number of important subsystems that interact at this layer, they are described later in this manual.

The `MCStreamer` API

`MCStreamer` is best thought of as an assembler API. It is an abstract API which is *implemented* in different ways (e.g. to output a `.s` file, output an ELF `.o` file, etc) but whose API correspond directly to what you see in a `.s` file. `MCStreamer` has one method per directive, such as `EmitLabel`, `EmitSymbolAttribute`, `SwitchSection`, `EmitValue` (for `.byte`, `.word`), etc, which directly correspond to assembly level directives. It also has an `EmitInstruction` method, which is used to output an `MCInst` to the streamer.

This API is most important for two clients: the `llvm-mc` stand-alone assembler is effectively a parser that parses a line, then invokes a method on `MCStreamer`. In the code generator, the *Code Emission* phase of the code generator lowers higher level LLVM IR and `Machine*` constructs down to the MC layer, emitting directives through `MCStreamer`.

On the implementation side of `MCStreamer`, there are two major implementations: one for writing out a `.s` file (`MCAsmStreamer`), and one for writing out a `.o` file (`MCOjectStreamer`). `MCAsmStreamer` is a straightforward

implementation that prints out a directive for each method (e.g. `EmitValue -> .byte`), but `MCOjectStreamer` implements a full assembler.

For target specific directives, the `MCStreamer` has a `MCTargetStreamer` instance. Each target that needs it defines a class that inherits from it and is a lot like `MCStreamer` itself: It has one method per directive and two classes that inherit from it, a target object streamer and a target asm streamer. The target asm streamer just prints it (`emitFnStart -> .fnstart`), and the object streamer implement the assembler logic for it.

To make `llvm` use these classes, the target initialization must call `TargetRegistry::RegisterAsmStreamer` and `TargetRegistry::RegisterMCOjectStreamer` passing callbacks that allocate the corresponding target streamer and pass it to `createAsmStreamer` or to the appropriate object streamer constructor.

The `MContext` class

The `MContext` class is the owner of a variety of unique data structures at the MC layer, including symbols, sections, etc. As such, this is the class that you interact with to create symbols and sections. This class can not be subclassed.

The `MCSymbol` class

The `MCSymbol` class represents a symbol (aka label) in the assembly file. There are two interesting kinds of symbols: assembler temporary symbols, and normal symbols. Assembler temporary symbols are used and processed by the assembler but are discarded when the object file is produced. The distinction is usually represented by adding a prefix to the label, for example “L” labels are assembler temporary labels in MachO.

`MCSymbols` are created by `MContext` and unique there. This means that `MCSymbols` can be compared for pointer equivalence to find out if they are the same symbol. Note that pointer inequality does not guarantee the labels will end up at different addresses though. It’s perfectly legal to output something like this to the `.s` file:

```
foo:
bar:
    .byte 4
```

In this case, both the `foo` and `bar` symbols will have the same address.

The `MCSection` class

The `MCSection` class represents an object-file specific section. It is subclassed by object file specific implementations (e.g. `MCSectionMachO`, `MCSectionCOFF`, `MCSectionELF`) and these are created and unique by `MContext`. The `MCStreamer` has a notion of the current section, which can be changed with the `SwitchToSection` method (which corresponds to a “.section” directive in a `.s` file).

The `MCInst` class

The `MCInst` class is a target-independent representation of an instruction. It is a simple class (much more so than *MachineInstr*) that holds a target-specific opcode and a vector of `MCOperands`. `MCOperand`, in turn, is a simple discriminated union of three cases: 1) a simple immediate, 2) a target register ID, 3) a symbolic expression (e.g. “`Lfoo-Lbar+42`”) as an `MCEExpr`.

`MCInst` is the common currency used to represent machine instructions at the MC layer. It is the type used by the instruction encoder, the instruction printer, and the type generated by the assembly parser and disassembler.

4.7.5 Target-independent code generation algorithms

This section documents the phases described in the *high-level design of the code generator*. It explains how they work and some of the rationale behind their design.

Instruction Selection

Instruction Selection is the process of translating LLVM code presented to the code generator into target-specific machine instructions. There are several well-known ways to do this in the literature. LLVM uses a SelectionDAG based instruction selector.

Portions of the DAG instruction selector are generated from the target description (`*.td`) files. Our goal is for the entire instruction selector to be generated from these `.td` files, though currently there are still things that require custom C++ code.

Introduction to SelectionDAGs

The SelectionDAG provides an abstraction for code representation in a way that is amenable to instruction selection using automatic techniques (e.g. dynamic-programming based optimal pattern matching selectors). It is also well-suited to other phases of code generation; in particular, instruction scheduling (SelectionDAG's are very close to scheduling DAGs post-selection). Additionally, the SelectionDAG provides a host representation where a large variety of very-low-level (but target-independent) *optimizations* may be performed; ones which require extensive information about the instructions efficiently supported by the target.

The SelectionDAG is a Directed-Acyclic-Graph whose nodes are instances of the `SDNode` class. The primary payload of the `SDNode` is its operation code (Opcode) that indicates what operation the node performs and the operands to the operation. The various operation node types are described at the top of the `include/llvm/CodeGen/ISDOpcodes.h` file.

Although most operations define a single value, each node in the graph may define multiple values. For example, a combined div/rem operation will define both the dividend and the remainder. Many other situations require multiple values as well. Each node also has some number of operands, which are edges to the node defining the used value. Because nodes may define multiple values, edges are represented by instances of the `SDValue` class, which is a `<SDNode, unsigned>` pair, indicating the node and result value being used, respectively. Each value produced by an `SDNode` has an associated MVT (Machine Value Type) indicating what the type of the value is.

SelectionDAGs contain two different kinds of values: those that represent data flow and those that represent control flow dependencies. Data values are simple edges with an integer or floating point value type. Control edges are represented as “chain” edges which are of type `MVT::Other`. These edges provide an ordering between nodes that have side effects (such as loads, stores, calls, returns, etc). All nodes that have side effects should take a token chain as input and produce a new one as output. By convention, token chain inputs are always operand #0, and chain results are always the last value produced by an operation. However, after instruction selection, the machine nodes have their chain after the instruction's operands, and may be followed by glue nodes.

A SelectionDAG has designated “Entry” and “Root” nodes. The Entry node is always a marker node with an Opcode of `ISD::EntryToken`. The Root node is the final side-effecting node in the token chain. For example, in a single basic block function it would be the return node.

One important concept for SelectionDAGs is the notion of a “legal” vs. “illegal” DAG. A legal DAG for a target is one that only uses supported operations and supported types. On a 32-bit PowerPC, for example, a DAG with a value of type `i1`, `i8`, `i16`, or `i64` would be illegal, as would a DAG that uses a `SREM` or `UREM` operation. The *legalize types* and *legalize operations* phases are responsible for turning an illegal DAG into a legal DAG.

SelectionDAG Instruction Selection Process

SelectionDAG-based instruction selection consists of the following steps:

1. *Build initial DAG* — This stage performs a simple translation from the input LLVM code to an illegal SelectionDAG.
2. *Optimize SelectionDAG* — This stage performs simple optimizations on the SelectionDAG to simplify it, and recognize meta instructions (like rotates and `div/rem` pairs) for targets that support these meta operations. This makes the resultant code more efficient and the *select instructions from DAG* phase (below) simpler.
3. *Legalize SelectionDAG Types* — This stage transforms SelectionDAG nodes to eliminate any types that are unsupported on the target.
4. *Optimize SelectionDAG* — The SelectionDAG optimizer is run to clean up redundancies exposed by type legalization.
5. *Legalize SelectionDAG Ops* — This stage transforms SelectionDAG nodes to eliminate any operations that are unsupported on the target.
6. *Optimize SelectionDAG* — The SelectionDAG optimizer is run to eliminate inefficiencies introduced by operation legalization.
7. *Select instructions from DAG* — Finally, the target instruction selector matches the DAG operations to target instructions. This process translates the target-independent input DAG into another DAG of target instructions.
8. *SelectionDAG Scheduling and Formation* — The last phase assigns a linear order to the instructions in the target-instruction DAG and emits them into the MachineFunction being compiled. This step uses traditional prepass scheduling techniques.

After all of these steps are complete, the SelectionDAG is destroyed and the rest of the code generation passes are run.

One great way to visualize what is going on here is to take advantage of a few LLC command line options. The following options pop up a window displaying the SelectionDAG at specific times (if you only get errors printed to the console while using this, you probably [need to configure your system](#) to add support for it).

- `-view-dag-combine1-dags` displays the DAG after being built, before the first optimization pass.
- `-view-legalize-dags` displays the DAG before Legalization.
- `-view-dag-combine2-dags` displays the DAG before the second optimization pass.
- `-view-isel-dags` displays the DAG before the Select phase.
- `-view-sched-dags` displays the DAG before Scheduling.

The `-view-sunit-dags` displays the Scheduler's dependency graph. This graph is based on the final SelectionDAG, with nodes that must be scheduled together bundled into a single scheduling-unit node, and with immediate operands and other nodes that aren't relevant for scheduling omitted.

The option `-filter-view-dags` allows to select the name of the basic block that you are interested to visualize and filters all the previous `view-*-dags` options.

Initial SelectionDAG Construction

The initial SelectionDAG is navelly peephole expanded from the LLVM input by the `SelectionDAGBuilder` class. The intent of this pass is to expose as much low-level, target-specific details to the SelectionDAG as possible. This pass is mostly hard-coded (e.g. an LLVM `add` turns into an `SDNode add` while a `getelementptr` is expanded into the obvious arithmetic). This pass requires target-specific hooks to lower calls, returns, varargs, etc. For these features, the *TargetLowering* interface is used.

SelectionDAG LegalizeTypes Phase

The Legalize phase is in charge of converting a DAG to only use the types that are natively supported by the target.

There are two main ways of converting values of unsupported scalar types to values of supported types: converting small types to larger types (“promoting”), and breaking up large integer types into smaller ones (“expanding”). For example, a target might require that all f32 values are promoted to f64 and that all i1/i8/i16 values are promoted to i32. The same target might require that all i64 values be expanded into pairs of i32 values. These changes can insert sign and zero extensions as needed to make sure that the final code has the same behavior as the input.

There are two main ways of converting values of unsupported vector types to value of supported types: splitting vector types, multiple times if necessary, until a legal type is found, and extending vector types by adding elements to the end to round them out to legal types (“widening”). If a vector gets split all the way down to single-element parts with no supported vector type being found, the elements are converted to scalars (“scalarizing”).

A target implementation tells the legalizer which types are supported (and which register class to use for them) by calling the `addRegisterClass` method in its `TargetLowering` constructor.

SelectionDAG Legalize Phase

The Legalize phase is in charge of converting a DAG to only use the operations that are natively supported by the target.

Targets often have weird constraints, such as not supporting every operation on every supported datatype (e.g. X86 does not support byte conditional moves and PowerPC does not support sign-extending loads from a 16-bit memory location). Legalize takes care of this by open-coding another sequence of operations to emulate the operation (“expansion”), by promoting one type to a larger type that supports the operation (“promotion”), or by using a target-specific hook to implement the legalization (“custom”).

A target implementation tells the legalizer which operations are not supported (and which of the above three actions to take) by calling the `setOperationAction` method in its `TargetLowering` constructor.

Prior to the existence of the Legalize passes, we required that every target *selector* supported and handled every operator and type even if they are not natively supported. The introduction of the Legalize phases allows all of the canonicalization patterns to be shared across targets, and makes it very easy to optimize the canonicalized code because it is still in the form of a DAG.

SelectionDAG Optimization Phase: the DAG Combiner

The SelectionDAG optimization phase is run multiple times for code generation, immediately after the DAG is built and once after each legalization. The first run of the pass allows the initial code to be cleaned up (e.g. performing optimizations that depend on knowing that the operators have restricted type inputs). Subsequent runs of the pass clean up the messy code generated by the Legalize passes, which allows Legalize to be very simple (it can focus on making code legal instead of focusing on generating *good* and legal code).

One important class of optimizations performed is optimizing inserted sign and zero extension instructions. We currently use ad-hoc techniques, but could move to more rigorous techniques in the future. Here are some good papers on the subject:

“[Widening integer arithmetic](#)” Kevin Redwine and Norman Ramsey International Conference on Compiler Construction (CC) 2004

“[Effective sign extension elimination](#)” Motohiro Kawahito, Hideaki Komatsu, and Toshio Nakatani Proceedings of the ACM SIGPLAN 2002 Conference on Programming Language Design and Implementation.

SelectionDAG Select Phase

The Select phase is the bulk of the target-specific code for instruction selection. This phase takes a legal SelectionDAG as input, pattern matches the instructions supported by the target to this DAG, and produces a new DAG of target code. For example, consider the following LLVM fragment:

```
%t1 = fadd float %W, %X
%t2 = fmul float %t1, %Y
%t3 = fadd float %t2, %Z
```

This LLVM code corresponds to a SelectionDAG that looks basically like this:

```
(fadd:f32 (fmul:f32 (fadd:f32 W, X), Y), Z)
```

If a target supports floating point multiply-and-add (FMA) operations, one of the adds can be merged with the multiply. On the PowerPC, for example, the output of the instruction selector might look like this DAG:

```
(FMADDS (FADDS W, X), Y, Z)
```

The FMADDS instruction is a ternary instruction that multiplies its first two operands and adds the third (as single-precision floating-point numbers). The FADDS instruction is a simple binary single-precision add instruction. To perform this pattern match, the PowerPC backend includes the following instruction definitions:

```
def FMADDS : AForm_1<59, 29,
    (ops F4RC:$FRT, F4RC:$FRA, F4RC:$FRC, F4RC:$FRB),
    "fmadds $FRT, $FRA, $FRC, $FRB",
    [(set F4RC:$FRT, (fadd (fmul F4RC:$FRA, F4RC:$FRC),
        F4RC:$FRB))]>;
def FADDS : AForm_2<59, 21,
    (ops F4RC:$FRT, F4RC:$FRA, F4RC:$FRB),
    "fadds $FRT, $FRA, $FRB",
    [(set F4RC:$FRT, (fadd F4RC:$FRA, F4RC:$FRB))]>;
```

The highlighted portion of the instruction definitions indicates the pattern used to match the instructions. The DAG operators (like `fmul`/`fadd`) are defined in the `include/llvm/Target/TargetSelectionDAG.td` file. “F4RC” is the register class of the input and result values.

The TableGen DAG instruction selector generator reads the instruction patterns in the `.td` file and automatically builds parts of the pattern matching code for your target. It has the following strengths:

- At compiler-compile time, it analyzes your instruction patterns and tells you if your patterns make sense or not.
- It can handle arbitrary constraints on operands for the pattern match. In particular, it is straight-forward to say things like “match any immediate that is a 13-bit sign-extended value”. For examples, see the `immSExt16` and related `tblgen` classes in the PowerPC backend.
- It knows several important identities for the patterns defined. For example, it knows that addition is commutative, so it allows the FMADDS pattern above to match “(fadd X, (fmul Y, Z))” as well as “(fadd (fmul X, Y), Z)”, without the target author having to specially handle this case.
- It has a full-featured type-inferencing system. In particular, you should rarely have to explicitly tell the system what type parts of your patterns are. In the FMADDS case above, we didn’t have to tell `tblgen` that all of the nodes in the pattern are of type ‘f32’. It was able to infer and propagate this knowledge from the fact that F4RC has type ‘f32’.
- Targets can define their own (and rely on built-in) “pattern fragments”. Pattern fragments are chunks of reusable patterns that get inlined into your patterns during compiler-compile time. For example, the integer “(not x)” operation is actually defined as a pattern fragment that expands as “(xor x, -1)”, since the SelectionDAG

does not have a native ‘not’ operation. Targets can define their own short-hand fragments as they see fit. See the definition of ‘not’ and ‘ineg’ for examples.

- In addition to instructions, targets can specify arbitrary patterns that map to one or more instructions using the ‘Pat’ class. For example, the PowerPC has no way to load an arbitrary integer immediate into a register in one instruction. To tell tblgen how to do this, it defines:

```
// Arbitrary immediate support.  Implement in terms of LIS/ORI.
def : Pat<(i32 imm:$imm),
      (ORI (LIS (HI16 imm:$imm)), (LO16 imm:$imm))>;
```

If none of the single-instruction patterns for loading an immediate into a register match, this will be used. This rule says “match an arbitrary i32 immediate, turning it into an ORI (‘or a 16-bit immediate’) and an LIS (‘load 16-bit immediate, where the immediate is shifted to the left 16 bits’) instruction”. To make this work, the LO16/HI16 node transformations are used to manipulate the input immediate (in this case, take the high or low 16-bits of the immediate).

- When using the ‘Pat’ class to map a pattern to an instruction that has one or more complex operands (like e.g. [X86 addressing mode](#)), the pattern may either specify the operand as a whole using a ComplexPattern, or else it may specify the components of the complex operand separately. The latter is done e.g. for pre-increment instructions by the PowerPC back end:

```
def STWU : DForm_1<37, (outs ptr_rc:$ea_res), (ins GPRC:$rS, memri:$dst),
      "stwu $rS, $dst", LdStStoreUpd, []>,
      RegConstraint<"$dst.reg = $ea_res">, NoEncode<"$ea_res">;

def : Pat<(pre_store GPRC:$rS, ptr_rc:$ptrreg, iaddroff:$ptroff),
      (STWU GPRC:$rS, iaddroff:$ptroff, ptr_rc:$ptrreg)>;
```

Here, the pair of `ptroff` and `ptrreg` operands is matched onto the complex operand `dst` of class `memri` in the `STWU` instruction.

- While the system does automate a lot, it still allows you to write custom C++ code to match special cases if there is something that is hard to express.

While it has many strengths, the system currently has some limitations, primarily because it is a work in progress and is not yet finished:

- Overall, there is no way to define or match SelectionDAG nodes that define multiple values (e.g. `SMUL_LOHI`, `LOAD`, `CALL`, etc). This is the biggest reason that you currently still *have to* write custom C++ code for your instruction selector.
- There is no great way to support matching complex addressing modes yet. In the future, we will extend pattern fragments to allow them to define multiple values (e.g. the four operands of the [X86 addressing mode](#), which are currently matched with custom C++ code). In addition, we’ll extend fragments so that a fragment can match multiple different patterns.
- We don’t automatically infer flags like `isStore/isLoad` yet.
- We don’t automatically generate the set of supported registers and operations for the [Legalizer](#) yet.
- We don’t have a way of tying in custom legalized nodes yet.

Despite these limitations, the instruction selector generator is still quite useful for most of the binary and logical operations in typical instruction sets. If you run into any problems or can’t figure out how to do something, please let Chris know!

SelectionDAG Scheduling and Formation Phase

The scheduling phase takes the DAG of target instructions from the selection phase and assigns an order. The scheduler can pick an order depending on various constraints of the machines (i.e. order for minimal register pressure or try to cover instruction latencies). Once an order is established, the DAG is converted to a list of *MachineInstrs* and the SelectionDAG is destroyed.

Note that this phase is logically separate from the instruction selection phase, but is tied to it closely in the code because it operates on SelectionDAGs.

Future directions for the SelectionDAG

1. Optional function-at-a-time selection.
2. Auto-generate entire selector from `.td` file.

SSA-based Machine Code Optimizations

To Be Written

Live Intervals

Live Intervals are the ranges (intervals) where a variable is *live*. They are used by some *register allocator* passes to determine if two or more virtual registers which require the same physical register are live at the same point in the program (i.e., they conflict). When this situation occurs, one virtual register must be *spilled*.

Live Variable Analysis

The first step in determining the live intervals of variables is to calculate the set of registers that are immediately dead after the instruction (i.e., the instruction calculates the value, but it is never used) and the set of registers that are used by the instruction, but are never used after the instruction (i.e., they are killed). Live variable information is computed for each *virtual* register and *register allocatable* physical register in the function. This is done in a very efficient manner because it uses SSA to sparsely compute lifetime information for virtual registers (which are in SSA form) and only has to track physical registers within a block. Before register allocation, LLVM can assume that physical registers are only live within a single basic block. This allows it to do a single, local analysis to resolve physical register lifetimes within each basic block. If a physical register is not register allocatable (e.g., a stack pointer or condition codes), it is not tracked.

Physical registers may be live in to or out of a function. Live in values are typically arguments in registers. Live out values are typically return values in registers. Live in values are marked as such, and are given a dummy “defining” instruction during live intervals analysis. If the last basic block of a function is a `return`, then it’s marked as using all live out values in the function.

PHI nodes need to be handled specially, because the calculation of the live variable information from a depth first traversal of the CFG of the function won’t guarantee that a virtual register used by the PHI node is defined before it’s used. When a PHI node is encountered, only the definition is handled, because the uses will be handled in other basic blocks.

For each PHI node of the current basic block, we simulate an assignment at the end of the current basic block and traverse the successor basic blocks. If a successor basic block has a PHI node and one of the PHI node’s operands is coming from the current basic block, then the variable is marked as *alive* within the current basic block and all of its predecessor basic blocks, until the basic block with the defining instruction is encountered.

Live Intervals Analysis

We now have the information available to perform the live intervals analysis and build the live intervals themselves. We start off by numbering the basic blocks and machine instructions. We then handle the “live-in” values. These are in physical registers, so the physical register is assumed to be killed by the end of the basic block. Live intervals for virtual registers are computed for some ordering of the machine instructions $[1, N]$. A live interval is an interval $[i, j)$, where $1 \leq i \leq j \leq N$, for which a variable is live.

Note: More to come...

Register Allocation

The *Register Allocation problem* consists in mapping a program P_v , that can use an unbounded number of virtual registers, to a program P_p that contains a finite (possibly small) number of physical registers. Each target architecture has a different number of physical registers. If the number of physical registers is not enough to accommodate all the virtual registers, some of them will have to be mapped into memory. These virtuals are called *spilled virtuals*.

How registers are represented in LLVM

In LLVM, physical registers are denoted by integer numbers that normally range from 1 to 1023. To see how this numbering is defined for a particular architecture, you can read the `GenRegisterNames.inc` file for that architecture. For instance, by inspecting `lib/Target/X86/X86GenRegisterInfo.inc` we see that the 32-bit register EAX is denoted by 43, and the MMX register MM0 is mapped to 65.

Some architectures contain registers that share the same physical location. A notable example is the X86 platform. For instance, in the X86 architecture, the registers EAX, AX and AL share the first eight bits. These physical registers are marked as *aliased* in LLVM. Given a particular architecture, you can check which registers are aliased by inspecting its `RegisterInfo.td` file. Moreover, the class `MCRRegAliasIterator` enumerates all the physical registers aliased to a register.

Physical registers, in LLVM, are grouped in *Register Classes*. Elements in the same register class are functionally equivalent, and can be interchangeably used. Each virtual register can only be mapped to physical registers of a particular class. For instance, in the X86 architecture, some virtuals can only be allocated to 8 bit registers. A register class is described by `TargetRegisterClass` objects. To discover if a virtual register is compatible with a given physical, this code can be used:

```
bool RegMapping_Fer::compatible_class(MachineFunction &mf,
                                     unsigned v_reg,
                                     unsigned p_reg) {
    assert(TargetRegisterInfo::isPhysicalRegister(p_reg) &&
           "Target register must be physical");
    const TargetRegisterClass *trc = mf.getRegInfo().getRegClass(v_reg);
    return trc->contains(p_reg);
}
```

Sometimes, mostly for debugging purposes, it is useful to change the number of physical registers available in the target architecture. This must be done statically, inside the `TargetRegsterInfo.td` file. Just `grep` for `RegisterClass`, the last parameter of which is a list of registers. Just commenting some out is one simple way to avoid them being used. A more polite way is to explicitly exclude some registers from the *allocation order*. See the definition of the GR8 register class in `lib/Target/X86/X86RegisterInfo.td` for an example of this.

Virtual registers are also denoted by integer numbers. Contrary to physical registers, different virtual registers never share the same number. Whereas physical registers are statically defined in a `TargetRegisterInfo`.

td file and cannot be created by the application developer, that is not the case with virtual registers. In order to create new virtual registers, use the method `MachineRegisterInfo::createVirtualRegister()`. This method will return a new virtual register. Use an `IndexedMap<Foo, VirtReg2IndexFunctor>` to hold information per virtual register. If you need to enumerate all virtual registers, use the function `TargetRegisterInfo::index2VirtReg()` to find the virtual register numbers:

```
for (unsigned i = 0, e = MRI->getNumVirtRegs(); i != e; ++i) {
    unsigned VirtReg = TargetRegisterInfo::index2VirtReg(i);
    stuff(VirtReg);
}
```

Before register allocation, the operands of an instruction are mostly virtual registers, although physical registers may also be used. In order to check if a given machine operand is a register, use the boolean function `MachineOperand::isRegister()`. To obtain the integer code of a register, use `MachineOperand::getReg()`. An instruction may define or use a register. For instance, `ADD reg:1026 := reg:1025 reg:1024` defines the registers 1024, and uses registers 1025 and 1026. Given a register operand, the method `MachineOperand::isUse()` informs if that register is being used by the instruction. The method `MachineOperand::isDef()` informs if that registers is being defined.

We will call physical registers present in the LLVM bitcode before register allocation *pre-colored registers*. Pre-colored registers are used in many different situations, for instance, to pass parameters of functions calls, and to store results of particular instructions. There are two types of pre-colored registers: the ones *implicitly* defined, and those *explicitly* defined. Explicitly defined registers are normal operands, and can be accessed with `MachineInstr::getOperand(int)::getReg()`. In order to check which registers are implicitly defined by an instruction, use the `TargetInstrInfo::get(opcode)::ImplicitDefs`, where `opcode` is the opcode of the target instruction. One important difference between explicit and implicit physical registers is that the latter are defined statically for each instruction, whereas the former may vary depending on the program being compiled. For example, an instruction that represents a function call will always implicitly define or use the same set of physical registers. To read the registers implicitly used by an instruction, use `TargetInstrInfo::get(opcode)::ImplicitUses`. Pre-colored registers impose constraints on any register allocation algorithm. The register allocator must make sure that none of them are overwritten by the values of virtual registers while still alive.

Mapping virtual registers to physical registers

There are two ways to map virtual registers to physical registers (or to memory slots). The first way, that we will call *direct mapping*, is based on the use of methods of the classes `TargetRegisterInfo`, and `MachineOperand`. The second way, that we will call *indirect mapping*, relies on the `VirtRegMap` class in order to insert loads and stores sending and getting values to and from memory.

The direct mapping provides more flexibility to the developer of the register allocator; however, it is more error prone, and demands more implementation work. Basically, the programmer will have to specify where load and store instructions should be inserted in the target function being compiled in order to get and store values in memory. To assign a physical register to a virtual register present in a given operand, use `MachineOperand::setReg(p_reg)`. To insert a store instruction, use `TargetInstrInfo::storeRegToStackSlot(...)`, and to insert a load instruction, use `TargetInstrInfo::loadRegFromStackSlot`.

The indirect mapping shields the application developer from the complexities of inserting load and store instructions. In order to map a virtual register to a physical one, use `VirtRegMap::assignVirt2Phys(vreg, preg)`. In order to map a certain virtual register to memory, use `VirtRegMap::assignVirt2StackSlot(vreg)`. This method will return the stack slot where `vreg`'s value will be located. If it is necessary to map another virtual register to the same stack slot, use `VirtRegMap::assignVirt2StackSlot(vreg, stack_location)`. One important point to consider when using the indirect mapping, is that even if a virtual register is mapped to memory, it still needs to be mapped to a physical register. This physical register is the location where the virtual register is supposed to be found before being stored or after being reloaded.

If the indirect strategy is used, after all the virtual registers have been mapped to physical registers or stack slots, it is necessary to use a spiller object to place load and store instructions in the code. Every virtual that has been mapped to a stack slot will be stored to memory after being defined and will be loaded before being used. The implementation of the spiller tries to recycle load/store instructions, avoiding unnecessary instructions. For an example of how to invoke the spiller, see `RegAllocLinearScan::runOnMachineFunction` in `lib/CodeGen/RegAllocLinearScan.cpp`.

Handling two address instructions

With very rare exceptions (e.g., function calls), the LLVM machine code instructions are three address instructions. That is, each instruction is expected to define at most one register, and to use at most two registers. However, some architectures use two address instructions. In this case, the defined register is also one of the used registers. For instance, an instruction such as `ADD %EAX, %EBX`, in X86 is actually equivalent to `%EAX = %EAX + %EBX`.

In order to produce correct code, LLVM must convert three address instructions that represent two address instructions into true two address instructions. LLVM provides the pass `TwoAddressInstructionPass` for this specific purpose. It must be run before register allocation takes place. After its execution, the resulting code may no longer be in SSA form. This happens, for instance, in situations where an instruction such as `%a = ADD %b %c` is converted to two instructions such as:

```
%a = MOVE %b
%a = ADD %a %c
```

Notice that, internally, the second instruction is represented as `ADD %a[def/use] %c`. I.e., the register operand `%a` is both used and defined by the instruction.

The SSA deconstruction phase

An important transformation that happens during register allocation is called the *SSA Deconstruction Phase*. The SSA form simplifies many analyses that are performed on the control flow graph of programs. However, traditional instruction sets do not implement PHI instructions. Thus, in order to generate executable code, compilers must replace PHI instructions with other instructions that preserve their semantics.

There are many ways in which PHI instructions can safely be removed from the target code. The most traditional PHI deconstruction algorithm replaces PHI instructions with copy instructions. That is the strategy adopted by LLVM. The SSA deconstruction algorithm is implemented in `lib/CodeGen/PHIElimination.cpp`. In order to invoke this pass, the identifier `PHIEliminationID` must be marked as required in the code of the register allocator.

Instruction folding

Instruction folding is an optimization performed during register allocation that removes unnecessary copy instructions. For instance, a sequence of instructions such as:

```
%EBX = LOAD %mem_address
%EAX = COPY %EBX
```

can be safely substituted by the single instruction:

```
%EAX = LOAD %mem_address
```

Instructions can be folded with the `TargetRegisterInfo::foldMemoryOperand(...)` method. Care must be taken when folding instructions; a folded instruction can be quite different from the original instruction. See

`LiveIntervals::addIntervalsForSpills` in `lib/CodeGen/LiveIntervalAnalysis.cpp` for an example of its use.

Built in register allocators

The LLVM infrastructure provides the application developer with three different register allocators:

- *Fast* — This register allocator is the default for debug builds. It allocates registers on a basic block level, attempting to keep values in registers and reusing registers as appropriate.
- *Basic* — This is an incremental approach to register allocation. Live ranges are assigned to registers one at a time in an order that is driven by heuristics. Since code can be rewritten on-the-fly during allocation, this framework allows interesting allocators to be developed as extensions. It is not itself a production register allocator but is a potentially useful stand-alone mode for triaging bugs and as a performance baseline.
- *Greedy* — *The default allocator.* This is a highly tuned implementation of the *Basic* allocator that incorporates global live range splitting. This allocator works hard to minimize the cost of spill code.
- *PBQP* — A Partitioned Boolean Quadratic Programming (PBQP) based register allocator. This allocator works by constructing a PBQP problem representing the register allocation problem under consideration, solving this using a PBQP solver, and mapping the solution back to a register assignment.

The type of register allocator used in `llc` can be chosen with the command line option `-regalloc=...`:

```
$ llc -regalloc=linearscan file.bc -o ln.s
$ llc -regalloc=fast file.bc -o fa.s
$ llc -regalloc=pbqp file.bc -o pbqp.s
```

Prolog/Epilog Code Insertion

Compact Unwind

Throwing an exception requires *unwinding* out of a function. The information on how to unwind a given function is traditionally expressed in DWARF unwind (a.k.a. frame) info. But that format was originally developed for debuggers to backtrace, and each Frame Description Entry (FDE) requires ~20-30 bytes per function. There is also the cost of mapping from an address in a function to the corresponding FDE at runtime. An alternative unwind encoding is called *compact unwind* and requires just 4-bytes per function.

The compact unwind encoding is a 32-bit value, which is encoded in an architecture-specific way. It specifies which registers to restore and from where, and how to unwind out of the function. When the linker creates a final linked image, it will create a `__TEXT,__unwind_info` section. This section is a small and fast way for the runtime to access unwind info for any given function. If we emit compact unwind info for the function, that compact unwind info will be encoded in the `__TEXT,__unwind_info` section. If we emit DWARF unwind info, the `__TEXT,__unwind_info` section will contain the offset of the FDE in the `__TEXT,__eh_frame` section in the final linked image.

For X86, there are three modes for the compact unwind encoding:

Function with a Frame Pointer (“EBP” or “RBP”) EBP/RBP-based frame, where EBP/RBP is pushed onto the stack immediately after the return address, then ESP/RSP is moved to EBP/RBP. Thus to unwind, ESP/RSP is restored with the current EBP/RBP value, then EBP/RBP is restored by popping the stack, and the return is done by popping the stack once more into the PC. All non-volatile registers that need to be restored must have been saved in a small range on the stack that starts EBP-4 to EBP-1020 (RBP-8 to RBP-1020). The offset (divided by 4 in 32-bit mode and 8 in 64-bit mode) is encoded in bits 16-23 (mask: 0x00FF0000). The registers saved are encoded in bits 0-14 (mask: 0x00007FFF) as five 3-bit entries from the following table:

Compact Number	i386 Register	x86-64 Register
1	EBX	RBX
2	ECX	R12
3	EDX	R13
4	EDI	R14
5	ESI	R15
6	EBP	RBP

Frameless with a Small Constant Stack Size (“EBP” or “RBP” is not used as a frame pointer) To return, a constant (encoded in the compact unwind encoding) is added to the ESP/RSP. Then the return is done by popping the stack into the PC. All non-volatile registers that need to be restored must have been saved on the stack immediately after the return address. The stack size (divided by 4 in 32-bit mode and 8 in 64-bit mode) is encoded in bits 16-23 (mask: 0x00FF0000). There is a maximum stack size of 1024 bytes in 32-bit mode and 2048 in 64-bit mode. The number of registers saved is encoded in bits 9-12 (mask: 0x00001C00). Bits 0-9 (mask: 0x000003FF) contain which registers were saved and their order. (See the `encodeCompactUnwindRegistersWithoutFrame()` function in `lib/Target/X86FrameLowering.cpp` for the encoding algorithm.)

Frameless with a Large Constant Stack Size (“EBP” or “RBP” is not used as a frame pointer) This case is like the “Frameless with a Small Constant Stack Size” case, but the stack size is too large to encode in the compact unwind encoding. Instead it requires that the function contains “`subl $nnnnnn, %esp`” in its prolog. The compact encoding contains the offset to the \$nnnnnn value in the function in bits 9-12 (mask: 0x00001C00).

Late Machine Code Optimizations

Note: To Be Written

Code Emission

The code emission step of code generation is responsible for lowering from the code generator abstractions (like *MachineFunction*, *MachineInstr*, etc) down to the abstractions used by the MC layer (*MCInst*, *MCStreamer*, etc). This is done with a combination of several different classes: the (misnamed) target-independent *AsmPrinter* class, target-specific subclasses of *AsmPrinter* (such as *SparcAsmPrinter*), and the *TargetLoweringObjectFile* class.

Since the MC layer works at the level of abstraction of object files, it doesn’t have a notion of functions, global variables etc. Instead, it thinks about labels, directives, and instructions. A key class used at this time is the *MCStreamer* class. This is an abstract API that is implemented in different ways (e.g. to output a .s file, output an ELF .o file, etc) that is effectively an “assembler API”. *MCStreamer* has one method per directive, such as *EmitLabel*, *EmitSymbolAttribute*, *SwitchSection*, etc, which directly correspond to assembly level directives.

If you are interested in implementing a code generator for a target, there are three important things that you have to implement for your target:

1. First, you need a subclass of *AsmPrinter* for your target. This class implements the general lowering process converting *MachineFunction*’s into MC label constructs. The *AsmPrinter* base class provides a number of useful methods and routines, and also allows you to override the lowering process in some important ways. You should get much of the lowering for free if you are implementing an ELF, COFF, or MachO target, because the *TargetLoweringObjectFile* class implements much of the common logic.
2. Second, you need to implement an instruction printer for your target. The instruction printer takes an *MCInst* and renders it to a `raw_ostream` as text. Most of this is automatically generated from the .td file (when you specify

something like “add \$dst, \$src1, \$src2” in the instructions), but you need to implement routines to print operands.

3. Third, you need to implement code that lowers a *MachineInstr* to an *MCInst*, usually implemented in “<target>MCInstLower.cpp”. This lowering process is often target specific, and is responsible for turning jump table entries, constant pool indices, global variable addresses, etc into *MCLabels* as appropriate. This translation layer is also responsible for expanding pseudo ops used by the code generator into the actual machine instructions they correspond to. The *MCInsts* that are generated by this are fed into the instruction printer or the encoder.

Finally, at your choosing, you can also implement a subclass of *MCCodeEmitter* which lowers *MCInst*’s into machine code bytes and relocations. This is important if you want to support direct .o file emission, or would like to implement an assembler for your target.

VLIW Packetizer

In a Very Long Instruction Word (VLIW) architecture, the compiler is responsible for mapping instructions to functional-units available on the architecture. To that end, the compiler creates groups of instructions called *packets* or *bundles*. The VLIW packetizer in LLVM is a target-independent mechanism to enable the packetization of machine instructions.

Mapping from instructions to functional units

Instructions in a VLIW target can typically be mapped to multiple functional units. During the process of packetizing, the compiler must be able to reason about whether an instruction can be added to a packet. This decision can be complex since the compiler has to examine all possible mappings of instructions to functional units. Therefore to alleviate compilation-time complexity, the VLIW packetizer parses the instruction classes of a target and generates tables at compiler build time. These tables can then be queried by the provided machine-independent API to determine if an instruction can be accommodated in a packet.

How the packetization tables are generated and used

The packetizer reads instruction classes from a target’s itineraries and creates a deterministic finite automaton (DFA) to represent the state of a packet. A DFA consists of three major elements: inputs, states, and transitions. The set of inputs for the generated DFA represents the instruction being added to a packet. The states represent the possible consumption of functional units by instructions in a packet. In the DFA, transitions from one state to another occur on the addition of an instruction to an existing packet. If there is a legal mapping of functional units to instructions, then the DFA contains a corresponding transition. The absence of a transition indicates that a legal mapping does not exist and that the instruction cannot be added to the packet.

To generate tables for a VLIW target, add *TargetGenDFAPacketizer.inc* as a target to the Makefile in the target directory. The exported API provides three functions: *DFAPacketizer::clearResources()*, *DFAPacketizer::reserveResources(MachineInstr *MI)*, and *DFAPacketizer::canReserveResources(MachineInstr *MI)*. These functions allow a target packetizer to add an instruction to an existing packet and to check whether an instruction can be added to a packet. See *llvm/CodeGen/DFAPacketizer.h* for more information.

4.7.6 Implementing a Native Assembler

Though you’re probably reading this because you want to write or maintain a compiler backend, LLVM also fully supports building a native assembler. We’ve tried hard to automate the generation of the assembler from the .td files (in particular the instruction syntax and encodings), which means that a large part of the manual and repetitive data entry can be factored and shared with the compiler.

Instruction Parsing

Note: To Be Written

Instruction Alias Processing

Once the instruction is parsed, it enters the `MatchInstructionImpl` function. The `MatchInstructionImpl` function performs alias processing and then does actual matching.

Alias processing is the phase that canonicalizes different lexical forms of the same instructions down to one representation. There are several different kinds of alias that are possible to implement and they are listed below in the order that they are processed (which is in order from simplest/weakest to most complex/powerful). Generally you want to use the first alias mechanism that meets the needs of your instruction, because it will allow a more concise description.

Mnemonic Aliases

The first phase of alias processing is simple instruction mnemonic remapping for classes of instructions which are allowed with two different mnemonics. This phase is a simple and unconditionally remapping from one input mnemonic to one output mnemonic. It isn't possible for this form of alias to look at the operands at all, so the remapping must apply for all forms of a given mnemonic. Mnemonic aliases are defined simply, for example X86 has:

```
def : MnemonicAlias<"cbw",      "cbtw">;
def : MnemonicAlias<"smovq",    "movsq">;
def : MnemonicAlias<"fldcw",    "fldcw">;
def : MnemonicAlias<"fucompi",  "fucomip">;
def : MnemonicAlias<"ud2a",     "ud2">;
```

... and many others. With a `MnemonicAlias` definition, the mnemonic is remapped simply and directly. Though `MnemonicAlias`'s can't look at any aspect of the instruction (such as the operands) they can depend on global modes (the same ones supported by the matcher), through a `Requires` clause:

```
def : MnemonicAlias<"pushf", "pushfq">, Requires<[In64BitMode]>;
def : MnemonicAlias<"pushf", "pushfl">, Requires<[In32BitMode]>;
```

In this example, the mnemonic gets mapped into a different one depending on the current instruction set.

Instruction Aliases

The most general phase of alias processing occurs while matching is happening: it provides new forms for the matcher to match along with a specific instruction to generate. An instruction alias has two parts: the string to match and the instruction to generate. For example:

```
def : InstAlias<"movsx $src, $dst", (MOVSX16rr8W GR16:$dst, GR8 :$src)>;
def : InstAlias<"movsx $src, $dst", (MOVSX16rm8W GR16:$dst, i8mem:$src)>;
def : InstAlias<"movsx $src, $dst", (MOVSX32rr8 GR32:$dst, GR8 :$src)>;
def : InstAlias<"movsx $src, $dst", (MOVSX32rr16 GR32:$dst, GR16 :$src)>;
def : InstAlias<"movsx $src, $dst", (MOVSX64rr8 GR64:$dst, GR8 :$src)>;
def : InstAlias<"movsx $src, $dst", (MOVSX64rr16 GR64:$dst, GR16 :$src)>;
def : InstAlias<"movsx $src, $dst", (MOVSX64rr32 GR64:$dst, GR32 :$src)>;
```

This shows a powerful example of the instruction aliases, matching the same mnemonic in multiple different ways depending on what operands are present in the assembly. The result of instruction aliases can include operands in a different order than the destination instruction, and can use an input multiple times, for example:

```
def : InstAlias<"clrb $reg", (XOR8rr GR8 :$reg, GR8 :$reg)>;
def : InstAlias<"clrw $reg", (XOR16rr GR16:$reg, GR16:$reg)>;
def : InstAlias<"clrl $reg", (XOR32rr GR32:$reg, GR32:$reg)>;
def : InstAlias<"clrq $reg", (XOR64rr GR64:$reg, GR64:$reg)>;
```

This example also shows that tied operands are only listed once. In the X86 backend, XOR8rr has two input GR8's and one output GR8 (where an input is tied to the output). InstAliases take a flattened operand list without duplicates for tied operands. The result of an instruction alias can also use immediates and fixed physical registers which are added as simple immediate operands in the result, for example:

```
// Fixed Immediate operand.
def : InstAlias<"aad", (AAD8i8 10)>;

// Fixed register operand.
def : InstAlias<"fcomi", (COM_FIr ST1)>;

// Simple alias.
def : InstAlias<"fcomi $reg", (COM_FIr RST:$reg)>;
```

Instruction aliases can also have a Requires clause to make them subtarget specific.

If the back-end supports it, the instruction printer can automatically emit the alias rather than what's being aliased. It typically leads to better, more readable code. If it's better to print out what's being aliased, then pass a '0' as the third parameter to the InstAlias definition.

Instruction Matching

Note: To Be Written

4.7.7 Target-specific Implementation Notes

This section of the document explains features or design decisions that are specific to the code generator for a particular target. First we start with a table that summarizes what features are supported by each target.

Target Feature Matrix

Note that this table does not list features that are not supported fully by any target yet. It considers a feature to be supported if at least one subtarget supports it. A feature being supported means that it is useful and works for most cases, it does not indicate that there are zero known bugs in the implementation. Here is the key:

Here is the table:

Is Generally Reliable

This box indicates whether the target is considered to be production quality. This indicates that the target has been used as a static compiler to compile large amounts of code by a variety of different people and is in continuous use.

Assembly Parser

This box indicates whether the target supports parsing target specific `.s` files by implementing the `MCAsmParser` interface. This is required for `llvm-mc` to be able to act as a native assembler and is required for inline assembly support in the native `.o` file writer.

Disassembler

This box indicates whether the target supports the `MCDisassembler` API for disassembling machine opcode bytes into `MCInst`'s.

Inline Asm

This box indicates whether the target supports most popular inline assembly constraints and modifiers.

JIT Support

This box indicates whether the target supports the JIT compiler through the `ExecutionEngine` interface. The ARM backend has basic support for integer code in ARM codegen mode, but lacks NEON and full Thumb support.

.o File Writing

This box indicates whether the target supports writing `.o` files (e.g. MachO, ELF, and/or COFF) files directly from the target. Note that the target also must include an assembly parser and general inline assembly support for full inline assembly support in the `.o` writer.

Targets that don't support this feature can obviously still write out `.o` files, they just rely on having an external assembler to translate from a `.s` file to a `.o` file (as is the case for many C compilers).

Tail Calls

This box indicates whether the target supports guaranteed tail calls. These are calls marked “[tail](#)” and use the `fastcc` calling convention. Please see the [tail call section](#) for more details.

Segmented Stacks

This box indicates whether the target supports segmented stacks. This replaces the traditional large C stack with many linked segments. It is compatible with the [gcc implementation](#) used by the Go front end. Basic support exists on the X86 backend. Currently `vararg` doesn't work and the object files are not marked the way the gold linker expects, but simple Go programs can be built by `dragonegg`.

Tail call optimization

Tail call optimization, callee reusing the stack of the caller, is currently supported on x86/x86-64 and PowerPC. It is performed if:

- Caller and callee have the calling convention `fastcc`, `cc 10` (GHC calling convention) or `cc 11` (HiPE calling convention).

- The call is a tail call - in tail position (ret immediately follows call and ret uses value of call or is void).
- Option `-tailcallopt` is enabled.
- Platform-specific constraints are met.

x86/x86-64 constraints:

- No variable argument lists are used.
- On x86-64 when generating GOT/PIC code only module-local calls (visibility = hidden or protected) are supported.

PowerPC constraints:

- No variable argument lists are used.
- No byval parameters are used.
- On ppc32/64 GOT/PIC only module-local calls (visibility = hidden or protected) are supported.

Example:

Call as `llc -tailcallopt test.ll`.

```
declare fastcc i32 @tailcallee(i32 inreg %a1, i32 inreg %a2, i32 %a3, i32 %a4)

define fastcc i32 @tailcaller(i32 %in1, i32 %in2) {
    %l1 = add i32 %in1, %in2
    %tmp = tail call fastcc i32 @tailcallee(i32 %in1 inreg, i32 %in2 inreg, i32 %in1,
    ↪ i32 %l1)
    ret i32 %tmp
}
```

Implications of `-tailcallopt`:

To support tail call optimization in situations where the callee has more arguments than the caller a ‘callee pops arguments’ convention is used. This currently causes each `fastcc` call that is not tail call optimized (because one or more of above constraints are not met) to be followed by a readjustment of the stack. So performance might be worse in such cases.

Sibling call optimization

Sibling call optimization is a restricted form of tail call optimization. Unlike tail call optimization described in the previous section, it can be performed automatically on any tail calls when `-tailcallopt` option is not specified.

Sibling call optimization is currently performed on x86/x86-64 when the following constraints are met:

- Caller and callee have the same calling convention. It can be either `c` or `fastcc`.
- The call is a tail call - in tail position (ret immediately follows call and ret uses value of call or is void).
- Caller and callee have matching return type or the callee result is not used.
- If any of the callee arguments are being passed in stack, they must be available in caller’s own incoming argument stack and the frame offsets must be the same.

Example:

```
declare i32 @bar(i32, i32)

define i32 @foo(i32 %a, i32 %b, i32 %c) {
entry:
```

```

%0 = tail call i32 @bar(i32 %a, i32 %b)
ret i32 %0
}

```

The X86 backend

The X86 code generator lives in the `lib/Target/X86` directory. This code generator is capable of targeting a variety of x86-32 and x86-64 processors, and includes support for ISA extensions such as MMX and SSE.

X86 Target Triples supported

The following are the known target triples that are supported by the X86 backend. This is not an exhaustive list, and it would be useful to add those that people test.

- **i686-pc-linux-gnu** — Linux
- **i386-unknown-freebsd5.3** — FreeBSD 5.3
- **i686-pc-cygwin** — Cygwin on Win32
- **i686-pc-mingw32** — MingW on Win32
- **i386-pc-mingw32msvc** — MingW crosscompiler on Linux
- **i686-apple-darwin*** — Apple Darwin on X86
- **x86_64-unknown-linux-gnu** — Linux

X86 Calling Conventions supported

The following target-specific calling conventions are known to backend:

- **x86_StdCall** — stdcall calling convention seen on Microsoft Windows platform (CC ID = 64).
- **x86_FastCall** — fastcall calling convention seen on Microsoft Windows platform (CC ID = 65).
- **x86_ThisCall** — Similar to X86_StdCall. Passes first argument in ECX, others via stack. Callee is responsible for stack cleaning. This convention is used by MSVC by default for methods in its ABI (CC ID = 70).

Representing X86 addressing modes in MachineInstrs

The x86 has a very flexible way of accessing memory. It is capable of forming memory addresses of the following expression directly in integer instructions (which use ModR/M addressing):

```
SegmentReg: Base + [1,2,4,8] * IndexReg + Disp32
```

In order to represent this, LLVM tracks no less than 5 operands for each memory operand of this form. This means that the “load” form of ‘mov’ has the following MachineOperands in this order:

Index:	0	1	2	3	4	5
Meaning:	DestReg,	BaseReg,	Scale,	IndexReg,	Displacement	Segment
OperandTy:	VirtReg,	VirtReg,	UnsImm,	VirtReg,	SignExtImm	PhysReg

Stores, and all other instructions, treat the four memory operands in the same way and in the same order. If the segment register is unspecified (regno = 0), then no segment override is generated. “Lea” operations do not have a segment register specified, so they only have 4 operands for their memory reference.

X86 address spaces supported

x86 has a feature which provides the ability to perform loads and stores to different address spaces via the x86 segment registers. A segment override prefix byte on an instruction causes the instruction's memory access to go to the specified segment. LLVM address space 0 is the default address space, which includes the stack, and any unqualified memory accesses in a program. Address spaces 1-255 are currently reserved for user-defined code. The GS-segment is represented by address space 256, the FS-segment is represented by address space 257, and the SS-segment is represented by address space 258. Other x86 segments have yet to be allocated address space numbers.

While these address spaces may seem similar to TLS via the `thread_local` keyword, and often use the same underlying hardware, there are some fundamental differences.

The `thread_local` keyword applies to global variables and specifies that they are to be allocated in thread-local memory. There are no type qualifiers involved, and these variables can be pointed to with normal pointers and accessed with normal loads and stores. The `thread_local` keyword is target-independent at the LLVM IR level (though LLVM doesn't yet have implementations of it for some configurations)

Special address spaces, in contrast, apply to static types. Every load and store has a particular address space in its address operand type, and this is what determines which address space is accessed. LLVM ignores these special address space qualifiers on global variables, and does not provide a way to directly allocate storage in them. At the LLVM IR level, the behavior of these special address spaces depends in part on the underlying OS or runtime environment, and they are specific to x86 (and LLVM doesn't yet handle them correctly in some cases).

Some operating systems and runtime environments use (or may in the future use) the FS/GS-segment registers for various low-level purposes, so care should be taken when considering them.

Instruction naming

An instruction name consists of the base name, a default operand size, and a character per operand with an optional special size. For example:

```
ADD8rr      -> add, 8-bit register, 8-bit register
IMUL16rmi   -> imul, 16-bit register, 16-bit memory, 16-bit immediate
IMUL16rmi8  -> imul, 16-bit register, 16-bit memory, 8-bit immediate
MOVSX32rm16 -> movsx, 32-bit register, 16-bit memory
```

The PowerPC backend

The PowerPC code generator lives in the `lib/Target/PowerPC` directory. The code generation is retargetable to several variations or *subtargets* of the PowerPC ISA; including `ppc32`, `ppc64` and `altivec`.

LLVM PowerPC ABI

LLVM follows the AIX PowerPC ABI, with two deviations. LLVM uses a PC relative (PIC) or static addressing for accessing global values, so no TOC (r2) is used. Second, r31 is used as a frame pointer to allow dynamic growth of a stack frame. LLVM takes advantage of having no TOC to provide space to save the frame pointer in the PowerPC linkage area of the caller frame. Other details of PowerPC ABI can be found at [PowerPC ABI](#). Note: This link describes the 32 bit ABI. The 64 bit ABI is similar except space for GPRs are 8 bytes wide (not 4) and r13 is reserved for system use.

Frame Layout

The size of a PowerPC frame is usually fixed for the duration of a function's invocation. Since the frame is fixed size, all references into the frame can be accessed via fixed offsets from the stack pointer. The exception to this is when dynamic alloca or variable sized arrays are present, then a base pointer (r31) is used as a proxy for the stack pointer and stack pointer is free to grow or shrink. A base pointer is also used if llvm-gcc is not passed the -fomit-frame-pointer flag. The stack pointer is always aligned to 16 bytes, so that space allocated for altivec vectors will be properly aligned.

An invocation frame is laid out as follows (low memory at top):

The *linkage* area is used by a callee to save special registers prior to allocating its own frame. Only three entries are relevant to LLVM. The first entry is the previous stack pointer (sp), aka link. This allows probing tools like gdb or exception handlers to quickly scan the frames in the stack. A function epilog can also use the link to pop the frame from the stack. The third entry in the linkage area is used to save the return address from the lr register. Finally, as mentioned above, the last entry is used to save the previous frame pointer (r31.) The entries in the linkage area are the size of a GPR, thus the linkage area is 24 bytes long in 32 bit mode and 48 bytes in 64 bit mode.

32 bit linkage area:

64 bit linkage area:

The *parameter area* is used to store arguments being passed to a callee function. Following the PowerPC ABI, the first few arguments are actually passed in registers, with the space in the parameter area unused. However, if there are not enough registers or the callee is a thunk or vararg function, these register arguments can be spilled into the parameter area. Thus, the parameter area must be large enough to store all the parameters for the largest call sequence made by the caller. The size must also be minimally large enough to spill registers r3-r10. This allows callees blind to the call signature, such as thunks and vararg functions, enough space to cache the argument registers. Therefore, the parameter area is minimally 32 bytes (64 bytes in 64 bit mode.) Also note that since the parameter area is a fixed offset from the top of the frame, that a callee can access its spilt arguments using fixed offsets from the stack pointer (or base pointer.)

Combining the information about the linkage, parameter areas and alignment. A stack frame is minimally 64 bytes in 32 bit mode and 128 bytes in 64 bit mode.

The *dynamic area* starts out as size zero. If a function uses dynamic alloca then space is added to the stack, the linkage and parameter areas are shifted to top of stack, and the new space is available immediately below the linkage and parameter areas. The cost of shifting the linkage and parameter areas is minor since only the link value needs to be copied. The link value can be easily fetched by adding the original frame size to the base pointer. Note that allocations in the dynamic space need to observe 16 byte alignment.

The *locals area* is where the llvm compiler reserves space for local variables.

The *saved registers area* is where the llvm compiler spills callee saved registers on entry to the callee.

Prolog/Epilog

The llvm prolog and epilog are the same as described in the PowerPC ABI, with the following exceptions. Callee saved registers are spilled after the frame is created. This allows the llvm epilog/prolog support to be common with other targets. The base pointer callee saved register r31 is saved in the TOC slot of linkage area. This simplifies allocation of space for the base pointer and makes it convenient to locate programmatically and during debugging.

Dynamic Allocation

Note: TODO - More to come.

The NVPTX backend

The NVPTX code generator under `lib/Target/NVPTX` is an open-source version of the NVIDIA NVPTX code generator for LLVM. It is contributed by NVIDIA and is a port of the code generator used in the CUDA compiler (`nvcc`). It targets the PTX 3.0/3.1 ISA and can target any compute capability greater than or equal to 2.0 (Fermi).

This target is of production quality and should be completely compatible with the official NVIDIA toolchain.

Code Generator Options:

The extended Berkeley Packet Filter (eBPF) backend

Extended BPF (or eBPF) is similar to the original (“classic”) BPF (cBPF) used to filter network packets. The `bpf()` system call performs a range of operations related to eBPF. For both cBPF and eBPF programs, the Linux kernel statically analyzes the programs before loading them, in order to ensure that they cannot harm the running system. eBPF is a 64-bit RISC instruction set designed for one to one mapping to 64-bit CPUs. Opcodes are 8-bit encoded, and 87 instructions are defined. There are 10 registers, grouped by function as outlined below.

R0	<code>return</code> value <code>from</code> <code>in</code> -kernel functions; exit value <code>for</code> eBPF program
R1 - R5	function call arguments to <code>in</code> -kernel functions
R6 - R9	callee-saved registers preserved by <code>in</code> -kernel functions
R10	stack frame pointer (read only)

Instruction encoding (arithmetic and jump)

eBPF is reusing most of the opcode encoding from classic to simplify conversion of classic BPF to eBPF. For arithmetic and jump instructions the 8-bit ‘code’ field is divided into three parts:

+-----+-----+-----+			
4 bits	1 bit	3 bits	
operation code	source	instruction class	
+-----+-----+-----+			
(MSB)			(LSB)

Three LSB bits store instruction class which is one of:

BPF_LD	0x0
BPF_LDX	0x1
BPF_ST	0x2
BPF_STX	0x3
BPF_ALU	0x4
BPF_JMP	0x5
(unused)	0x6
BPF_ALU64	0x7

When `BPF_CLASS(code) == BPF_ALU` or `BPF_ALU64` or `BPF_JMP`, 4th bit encodes source operand

BPF_X	0x0	use <code>src_reg</code> register <code>as</code> source operand
BPF_K	0x1	use 32 bit immediate <code>as</code> source operand

and four MSB bits store operation code

BPF_ADD	0x0	add
BPF_SUB	0x1	subtract
BPF_MUL	0x2	multiply

BPF_DIV	0x3	divide
BPF_OR	0x4	bitwise logical OR
BPF_AND	0x5	bitwise logical AND
BPF_LSH	0x6	left shift
BPF_RSH	0x7	right shift (zero extended)
BPF_NEG	0x8	arithmetic negation
BPF_MOD	0x9	modulo
BPF_XOR	0xa	bitwise logical XOR
BPF_MOV	0xb	move register to register
BPF_ARSH	0xc	right shift (sign extended)
BPF_END	0xd	endianness conversion

If `BPF_CLASS(code) == BPF_JMP`, `BPF_OP(code)` is one of

BPF_JA	0x0	unconditional jump
BPF_JEQ	0x1	jump ==
BPF_JGT	0x2	jump >
BPF_JGE	0x3	jump >=
BPF_JSET	0x4	jump if (DST & SRC)
BPF_JNE	0x5	jump !=
BPF_JSGT	0x6	jump signed >
BPF_JSGE	0x7	jump signed >=
BPF_CALL	0x8	function call
BPF_EXIT	0x9	function return

Instruction encoding (load, store)

For load and store instructions the 8-bit ‘code’ field is divided as:

+-----+-----+-----+-----+			
3 bits	2 bits	3 bits	
mode	size	instruction class	
+-----+-----+-----+-----+			
(MSB)		(LSB)	

Size modifier is one of

BPF_W	0x0	word
BPF_H	0x1	half word
BPF_B	0x2	byte
BPF_DW	0x3	double word

Mode modifier is one of

BPF_IMM	0x0	immediate
BPF_ABS	0x1	used to access packet data
BPF_IND	0x2	used to access packet data
BPF_MEM	0x3	memory
(reserved)	0x4	
(reserved)	0x5	
BPF_XADD	0x6	exclusive add

Packet data access (BPF_ABS, BPF_IND)

Two non-generic instructions: (BPF_ABS | <size> | BPF_LD) and (BPF_IND | <size> | BPF_LD) which are used to access packet data. Register R6 is an implicit input that must contain pointer to sk_buff. Register R0 is an implicit output which contains the data fetched from the packet. Registers R1-R5 are scratch registers and must not be used to store the data across BPF_ABS | BPF_LD or BPF_IND | BPF_LD instructions. These instructions have implicit program exit condition as well. When eBPF program is trying to access the data beyond the packet boundary, the interpreter will abort the execution of the program.

BPF_IND | BPF_W | BPF_LD is equivalent to: `R0 = ntohs(*(u32 *) (((struct sk_buff *) R6)->data + src_reg + imm32))`

eBPF maps

eBPF maps are provided for sharing data between kernel and user-space. Currently implemented types are hash and array, with potential extension to support bloom filters, radix trees, etc. A map is defined by its type, maximum number of elements, key size and value size in bytes. eBPF syscall supports create, update, find and delete functions on maps.

Function calls

Function call arguments are passed using up to five registers (R1 - R5). The return value is passed in a dedicated register (R0). Four additional registers (R6 - R9) are callee-saved, and the values in these registers are preserved within kernel functions. R0 - R5 are scratch registers within kernel functions, and eBPF programs must therefor store/restore values in these registers if needed across function calls. The stack can be accessed using the read-only frame pointer R10. eBPF registers map 1:1 to hardware registers on x86_64 and other 64-bit architectures. For example, x86_64 in-kernel JIT maps them as

R0	-	rax
R1	-	rdi
R2	-	rsi
R3	-	rdx
R4	-	rcx
R5	-	r8
R6	-	rbx
R7	-	r13
R8	-	r14
R9	-	r15
R10	-	rbp

since x86_64 ABI mandates rdi, rsi, rdx, rcx, r8, r9 for argument passing and rbx, r12 - r15 are callee saved.

Program start

An eBPF program receives a single argument and contains a single eBPF main routine; the program does not contain eBPF functions. Function calls are limited to a predefined set of kernel functions. The size of a program is limited to 4K instructions: this ensures fast termination and a limited number of kernel function calls. Prior to running an eBPF program, a verifier performs static analysis to prevent loops in the code and to ensure valid register usage and operand types.

The AMDGPU backend

The AMDGPU code generator lives in the `lib/Target/AMDGPU` directory. This code generator is capable of targeting a variety of AMD GPU processors. Refer to *User Guide for AMDGPU Backend* for more information.

4.8 Exception Handling in LLVM

- *Introduction*
 - *Itanium ABI Zero-cost Exception Handling*
 - *Setjmp/Longjmp Exception Handling*
 - *Windows Runtime Exception Handling*
 - *Overview*
- *LLVM Code Generation*
 - *Throw*
 - *Try/Catch*
 - *Cleanups*
 - *Throw Filters*
 - *Restrictions*
- *Exception Handling Intrinsics*
 - `llvm.eh.typeid.for`
 - `llvm.eh.begincatch`
 - `llvm.eh.endcatch`
 - `llvm.eh.exceptionpointer`
 - *SJLJ Intrinsics*
 - * `llvm.eh.sjlj.setjmp`
 - * `llvm.eh.sjlj.longjmp`
 - * `llvm.eh.sjlj.lsd`
 - * `llvm.eh.sjlj.callsite`
- *Asm Table Formats*
 - *Exception Handling Frame*
 - *Exception Tables*
- *Exception Handling using the Windows Runtime*
 - *Background on Windows exceptions*
 - *SEH filter expressions*
 - *New exception handling instructions*
 - *Funclet parent tokens*

4.8.1 Introduction

This document is the central repository for all information pertaining to exception handling in LLVM. It describes the format that LLVM exception handling information takes, which is useful for those interested in creating front-ends or dealing directly with the information. Further, this document provides specific examples of what exception handling information is used for in C and C++.

Itanium ABI Zero-cost Exception Handling

Exception handling for most programming languages is designed to recover from conditions that rarely occur during general use of an application. To that end, exception handling should not interfere with the main flow of an application's algorithm by performing checkpointing tasks, such as saving the current pc or register state.

The Itanium ABI Exception Handling Specification defines a methodology for providing outlying data in the form of exception tables without inlining speculative exception handling code in the flow of an application's main algorithm. Thus, the specification is said to add “zero-cost” to the normal execution of an application.

A more complete description of the Itanium ABI exception handling runtime support of can be found at [Itanium C++ ABI: Exception Handling](#). A description of the exception frame format can be found at [Exception Frames](#), with details of the DWARF 4 specification at [DWARF 4 Standard](#). A description for the C++ exception table formats can be found at [Exception Handling Tables](#).

Setjmp/Longjmp Exception Handling

Setjmp/Longjmp (SJLJ) based exception handling uses LLVM intrinsics *llvm.eh.sjlj.setjmp* and *llvm.eh.sjlj.longjmp* to handle control flow for exception handling.

For each function which does exception processing — be it `try/catch` blocks or cleanups — that function registers itself on a global frame list. When exceptions are unwinding, the runtime uses this list to identify which functions need processing.

Landing pad selection is encoded in the call site entry of the function context. The runtime returns to the function via *llvm.eh.sjlj.longjmp*, where a switch table transfers control to the appropriate landing pad based on the index stored in the function context.

In contrast to DWARF exception handling, which encodes exception regions and frame information in out-of-line tables, SJLJ exception handling builds and removes the unwind frame context at runtime. This results in faster exception handling at the expense of slower execution when no exceptions are thrown. As exceptions are, by their nature, intended for uncommon code paths, DWARF exception handling is generally preferred to SJLJ.

Windows Runtime Exception Handling

LLVM supports handling exceptions produced by the Windows runtime, but it requires a very different intermediate representation. It is not based on the “*landingpad*” instruction like the other two models, and is described later in this document under [Exception Handling using the Windows Runtime](#).

Overview

When an exception is thrown in LLVM code, the runtime does its best to find a handler suited to processing the circumstance.

The runtime first attempts to find an *exception frame* corresponding to the function where the exception was thrown. If the programming language supports exception handling (e.g. C++), the exception frame contains a reference to an exception table describing how to process the exception. If the language does not support exception handling (e.g. C), or if the exception needs to be forwarded to a prior activation, the exception frame contains information about how to unwind the current activation and restore the state of the prior activation. This process is repeated until the exception is handled. If the exception is not handled and no activations remain, then the application is terminated with an appropriate error message.

Because different programming languages have different behaviors when handling exceptions, the exception handling ABI provides a mechanism for supplying *personalities*. An exception handling personality is defined by way of a *personality function* (e.g. `__gxx_personality_v0` in C++), which receives the context of the exception, an *exception structure* containing the exception object type and value, and a reference to the exception table for the current function. The personality function for the current compile unit is specified in a *common exception frame*.

The organization of an exception table is language dependent. For C++, an exception table is organized as a series of code ranges defining what to do if an exception occurs in that range. Typically, the information associated with a range defines which types of exception objects (using C++ *type info*) that are handled in that range, and an associated action that should take place. Actions typically pass control to a *landing pad*.

A landing pad corresponds roughly to the code found in the `catch` portion of a `try/catch` sequence. When execution resumes at a landing pad, it receives an *exception structure* and a *selector value* corresponding to the *type* of exception thrown. The selector is then used to determine which *catch* should actually process the exception.

4.8.2 LLVM Code Generation

From a C++ developer's perspective, exceptions are defined in terms of the `throw` and `try/catch` statements. In this section we will describe the implementation of LLVM exception handling in terms of C++ examples.

Throw

Languages that support exception handling typically provide a `throw` operation to initiate the exception process. Internally, a `throw` operation breaks down into two steps.

1. A request is made to allocate exception space for an exception structure. This structure needs to survive beyond the current activation. This structure will contain the type and value of the object being thrown.
2. A call is made to the runtime to raise the exception, passing the exception structure as an argument.

In C++, the allocation of the exception structure is done by the `__cxa_allocate_exception` runtime function. The exception raising is handled by `__cxa_throw`. The type of the exception is represented using a C++ RTTI structure.

Try/Catch

A call within the scope of a `try` statement can potentially raise an exception. In those circumstances, the LLVM C++ front-end replaces the call with an `invoke` instruction. Unlike a call, the `invoke` has two potential continuation points:

1. where to continue when the call succeeds as per normal, and
2. where to continue if the call raises an exception, either by a throw or the unwinding of a throw

The term used to define the place where an `invoke` continues after an exception is called a *landing pad*. LLVM landing pads are conceptually alternative function entry points where an exception structure reference and a type info index are passed in as arguments. The landing pad saves the exception structure reference and then proceeds to select the catch block that corresponds to the type info of the exception object.

The LLVM *‘landingpad’ Instruction* is used to convey information about the landing pad to the back end. For C++, the `landingpad` instruction returns a pointer and integer pair corresponding to the pointer to the *exception structure* and the *selector value* respectively.

The `landingpad` instruction looks for a reference to the personality function to be used for this `try/catch` sequence in the parent function’s attribute list. The instruction contains a list of *cleanup*, *catch*, and *filter* clauses. The exception is tested against the clauses sequentially from first to last. The clauses have the following meanings:

- `catch <type> @ExcType`
 - This clause means that the landingpad block should be entered if the exception being thrown is of type `@ExcType` or a subtype of `@ExcType`. For C++, `@ExcType` is a pointer to the `std::type_info` object (an RTTI object) representing the C++ exception type.
 - If `@ExcType` is null, any exception matches, so the landingpad should always be entered. This is used for C++ catch-all blocks (“`catch (...)`”).
 - When this clause is matched, the selector value will be equal to the value returned by “`@llvm.eh.typeid.for(i8* @ExcType)`”. This will always be a positive value.
- `filter <type> [<type> @ExcType1, ..., <type> @ExcTypeN]`
 - This clause means that the landingpad should be entered if the exception being thrown does *not* match any of the types in the list (which, for C++, are again specified as `std::type_info` pointers).
 - C++ front-ends use this to implement C++ exception specifications, such as “`void foo() throw (ExcType1, ..., ExcTypeN) { ... }`”.
 - When this clause is matched, the selector value will be negative.
 - The array argument to `filter` may be empty; for example, “`[0 x i8**] undef`”. This means that the landingpad should always be entered. (Note that such a `filter` would not be equivalent to “`catch i8* null`”, because `filter` and `catch` produce negative and positive selector values respectively.)
- `cleanup`
 - This clause means that the landingpad should always be entered.
 - C++ front-ends use this for calling objects’ destructors.
 - When this clause is matched, the selector value will be zero.
 - The runtime may treat “`cleanup`” differently from “`catch <type> null`”.

In C++, if an unhandled exception occurs, the language runtime will call `std::terminate()`, but it is implementation-defined whether the runtime unwinds the stack and calls object destructors first. For example, the GNU C++ unwinder does not call object destructors when an unhandled exception occurs. The reason for this is to improve debuggability: it ensures that `std::terminate()` is called from the context of the `throw`, so that this context is not lost by unwinding the stack. A runtime will typically implement this by searching for a matching non-`cleanup` clause, and aborting if it does not find one, before entering any landingpad blocks.

Once the landing pad has the type info selector, the code branches to the code for the first catch. The catch then checks the value of the type info selector against the index of type info for that catch. Since the type info index is not known until all the type infos have been gathered in the backend, the catch code must call the *llvm.eh.typeid.for* intrinsic to determine the index for a given type info. If the catch fails to match the selector then control is passed on to the next catch.

Finally, the entry and exit of catch code is bracketed with calls to `__cxa_begin_catch` and `__cxa_end_catch`.

- `__cxa_begin_catch` takes an exception structure reference as an argument and returns the value of the exception object.
- `__cxa_end_catch` takes no arguments. This function:

1. Locates the most recently caught exception and decrements its handler count,
2. Removes the exception from the *caught* stack if the handler count goes to zero, and
3. Destroys the exception if the handler count goes to zero and the exception was not re-thrown by throw.

Note: a rethrow from within the catch may replace this call with a `__cxa_rethrow`.

Cleanups

A cleanup is extra code which needs to be run as part of unwinding a scope. C++ destructors are a typical example, but other languages and language extensions provide a variety of different kinds of cleanups. In general, a landing pad may need to run arbitrary amounts of cleanup code before actually entering a catch block. To indicate the presence of cleanups, a *'landingpad' Instruction* should have a *cleanup* clause. Otherwise, the unwinder will not stop at the landing pad if there are no catches or filters that require it to.

Note: Do not allow a new exception to propagate out of the execution of a cleanup. This can corrupt the internal state of the unwinder. Different languages describe different high-level semantics for these situations: for example, C++ requires that the process be terminated, whereas Ada cancels both exceptions and throws a third.

When all cleanups are finished, if the exception is not handled by the current function, resume unwinding by calling the *resume instruction*, passing in the result of the `landingpad` instruction for the original landing pad.

Throw Filters

C++ allows the specification of which exception types may be thrown from a function. To represent this, a top level landing pad may exist to filter out invalid types. To express this in LLVM code the *'landingpad' Instruction* will have a filter clause. The clause consists of an array of type infos. `landingpad` will return a negative value if the exception does not match any of the type infos. If no match is found then a call to `__cxa_call_unexpected` should be made, otherwise `_Unwind_Resume`. Each of these functions requires a reference to the exception structure. Note that the most general form of a `landingpad` instruction can have any number of catch, cleanup, and filter clauses (though having more than one cleanup is pointless). The LLVM C++ front-end can generate such `landingpad` instructions due to inlining creating nested exception handling scopes.

Restrictions

The unwinder delegates the decision of whether to stop in a call frame to that call frame's language-specific personality function. Not all unwinders guarantee that they will stop to perform cleanups. For example, the GNU C++ unwinder doesn't do so unless the exception is actually caught somewhere further up the stack.

In order for inlining to behave correctly, landing pads must be prepared to handle selector results that they did not originally advertise. Suppose that a function catches exceptions of type A, and it's inlined into a function that catches exceptions of type B. The inliner will update the `landingpad` instruction for the inlined landing pad to include the fact that B is also caught. If that landing pad assumes that it will only be entered to catch an A, it's in for a rude awakening. Consequently, landing pads must test for the selector results they understand and then resume exception propagation with the *resume instruction* if none of the conditions match.

4.8.3 Exception Handling Intrinsics

In addition to the `landingpad` and `resume` instructions, LLVM uses several intrinsic functions (name prefixed with `llvm.eh`) to provide exception handling information at various points in generated code.

`llvm.eh.typeid.for`

```
i32 @llvm.eh.typeid.for(i8* %type_info)
```

This intrinsic returns the type info index in the exception table of the current function. This value can be used to compare against the result of `landingpad` instruction. The single argument is a reference to a type info.

Uses of this intrinsic are generated by the C++ front-end.

`llvm.eh.begincatch`

```
void @llvm.eh.begincatch(i8* %ehptr, i8* %ehobj)
```

This intrinsic marks the beginning of catch handling code within the blocks following a `landingpad` instruction. The exact behavior of this function depends on the compilation target and the personality function associated with the `landingpad` instruction.

The first argument to this intrinsic is a pointer that was previously extracted from the aggregate return value of the `landingpad` instruction. The second argument to the intrinsic is a pointer to stack space where the exception object should be stored. The runtime handles the details of copying the exception object into the slot. If the second parameter is null, no copy occurs.

Uses of this intrinsic are generated by the C++ front-end. Many targets will use implementation-specific functions (such as `__cxa_begin_catch`) instead of this intrinsic. The intrinsic is provided for targets that require a more abstract interface.

When used in the native Windows C++ exception handling implementation, this intrinsic serves as a placeholder to delimit code before a catch handler is outlined. When the handler is outlined, this intrinsic will be replaced by instructions that retrieve the exception object pointer from the frame allocation block.

`llvm.eh.endcatch`

```
void @llvm.eh.endcatch()
```

This intrinsic marks the end of catch handling code within the current block, which will be a successor of a block which called `llvm.eh.begincatch`'. The exact behavior of this function depends on the compilation target and the personality function associated with the corresponding `landingpad` instruction.

There may be more than one call to `llvm.eh.endcatch` for any given call to `llvm.eh.begincatch` with each `llvm.eh.endcatch` call corresponding to the end of a different control path. All control paths following a call to `llvm.eh.begincatch` must reach a call to `llvm.eh.endcatch`.

Uses of this intrinsic are generated by the C++ front-end. Many targets will use implementation-specific functions (such as `__cxa_begin_catch`) instead of this intrinsic. The intrinsic is provided for targets that require a more abstract interface.

When used in the native Windows C++ exception handling implementation, this intrinsic serves as a placeholder to delimit code before a catch handler is outlined. After the handler is outlined, this intrinsic is simply removed.

llvm.eh.exceptionpointer

```
i8 addrspace(N)* @llvm.eh.padparam.pNi8(token %catchpad)
```

This intrinsic retrieves a pointer to the exception caught by the given `catchpad`.

SJLJ Intrinsics

The `llvm.eh.sjlj` intrinsics are used internally within LLVM's backend. Uses of them are generated by the backend's `SJLJEHPrepare` pass.

llvm.eh.sjlj.setjmp

```
i32 @llvm.eh.sjlj.setjmp(i8* %setjmp_buf)
```

For SJLJ based exception handling, this intrinsic forces register saving for the current function and stores the address of the following instruction for use as a destination address by [llvm.eh.sjlj.longjmp](#). The buffer format and the overall functioning of this intrinsic is compatible with the GCC `__builtin_setjmp` implementation allowing code built with the clang and GCC to interoperate.

The single parameter is a pointer to a five word buffer in which the calling context is saved. The front end places the frame pointer in the first word, and the target implementation of this intrinsic should place the destination address for a [llvm.eh.sjlj.longjmp](#) in the second word. The following three words are available for use in a target-specific manner.

llvm.eh.sjlj.longjmp

```
void @llvm.eh.sjlj.longjmp(i8* %setjmp_buf)
```

For SJLJ based exception handling, the `llvm.eh.sjlj.longjmp` intrinsic is used to implement `__builtin_longjmp()`. The single parameter is a pointer to a buffer populated by [llvm.eh.sjlj.setjmp](#). The frame pointer and stack pointer are restored from the buffer, then control is transferred to the destination address.

llvm.eh.sjlj.lsdas

```
i8* @llvm.eh.sjlj.lsdas()
```

For SJLJ based exception handling, the `llvm.eh.sjlj.lsdas` intrinsic returns the address of the Language Specific Data Area (LSDA) for the current function. The SJLJ front-end code stores this address in the exception handling function context for use by the runtime.

llvm.eh.sjlj.callsite

```
void @llvm.eh.sjlj.callsite(i32 %call_site_num)
```

For SJLJ based exception handling, the `llvm.eh.sjlj.callsite` intrinsic identifies the callsite value associated with the following `invoke` instruction. This is used to ensure that landing pad entries in the LSDA are generated in matching order.

4.8.4 Asm Table Formats

There are two tables that are used by the exception handling runtime to determine which actions should be taken when an exception is thrown.

Exception Handling Frame

An exception handling frame `eh_frame` is very similar to the unwind frame used by DWARF debug info. The frame contains all the information necessary to tear down the current frame and restore the state of the prior frame. There is an exception handling frame for each function in a compile unit, plus a common exception handling frame that defines information common to all functions in the unit.

The format of this call frame information (CFI) is often platform-dependent, however. ARM, for example, defines their own format. Apple has their own compact unwind info format. On Windows, another format is used for all architectures since 32-bit x86. LLVM will emit whatever information is required by the target.

Exception Tables

An exception table contains information about what actions to take when an exception is thrown in a particular part of a function's code. This is typically referred to as the language-specific data area (LSDA). The format of the LSDA table is specific to the personality function, but the majority of personalities out there use a variation of the tables consumed by `__gxx_personality_v0`. There is one exception table per function, except leaf functions and functions that have calls only to non-throwing functions. They do not need an exception table.

4.8.5 Exception Handling using the Windows Runtime

Background on Windows exceptions

Interacting with exceptions on Windows is significantly more complicated than on Itanium C++ ABI platforms. The fundamental difference between the two models is that Itanium EH is designed around the idea of “successive unwinding,” while Windows EH is not.

Under Itanium, throwing an exception typically involves allocating thread local memory to hold the exception, and calling into the EH runtime. The runtime identifies frames with appropriate exception handling actions, and successively resets the register context of the current thread to the most recently active frame with actions to run. In LLVM, execution resumes at a `landingpad` instruction, which produces register values provided by the runtime. If a function is only cleaning up allocated resources, the function is responsible for calling `_Unwind_Resume` to transition to the next most recently active frame after it is finished cleaning up. Eventually, the frame responsible for handling the exception calls `__cxa_end_catch` to destroy the exception, release its memory, and resume normal control flow.

The Windows EH model does not use these successive register context resets. Instead, the active exception is typically described by a frame on the stack. In the case of C++ exceptions, the exception object is allocated in stack memory and its address is passed to `__CxxThrowException`. General purpose structured exceptions (SEH) are more analogous to Linux signals, and they are dispatched by userspace DLLs provided with Windows. Each frame on the stack has an assigned EH personality routine, which decides what actions to take to handle the exception. There are a few major personalities for C and C++ code: the C++ personality (`__CxxFrameHandler3`) and the SEH personalities (`_except_handler3`, `_except_handler4`, and `__C_specific_handler`). All of them implement cleanups by calling back into a “funclet” contained in the parent function.

Funclets, in this context, are regions of the parent function that can be called as though they were a function pointer with a very special calling convention. The frame pointer of the parent frame is passed into the funclet either using the standard EBP register or as the first parameter register, depending on the architecture. The funclet implements the

EH action by accessing local variables in memory through the frame pointer, and returning some appropriate value, continuing the EH process. No variables live in to or out of the funclet can be allocated in registers.

The C++ personality also uses funclets to contain the code for catch blocks (i.e. all user code between the braces in `catch (Type obj) { ... }`). The runtime must use funclets for catch bodies because the C++ exception object is allocated in a child stack frame of the function handling the exception. If the runtime rewound the stack back to frame of the catch, the memory holding the exception would be overwritten quickly by subsequent function calls. The use of funclets also allows `__CxxFrameHandler3` to implement rethrow without resorting to TLS. Instead, the runtime throws a special exception, and then uses SEH (`__try` / `__except`) to resume execution with new information in the child frame.

In other words, the successive unwinding approach is incompatible with Visual C++ exceptions and general purpose Windows exception handling. Because the C++ exception object lives in stack memory, LLVM cannot provide a custom personality function that uses landingpads. Similarly, SEH does not provide any mechanism to rethrow an exception or continue unwinding. Therefore, LLVM must use the IR constructs described later in this document to implement compatible exception handling.

SEH filter expressions

The SEH personality functions also use funclets to implement filter expressions, which allow executing arbitrary user code to decide which exceptions to catch. Filter expressions should not be confused with the `filter` clause of the LLVM `landingpad` instruction. Typically filter expressions are used to determine if the exception came from a particular DLL or code region, or if code faulted while accessing a particular memory address range. LLVM does not currently have IR to represent filter expressions because it is difficult to represent their control dependencies. Filter expressions run during the first phase of EH, before cleanups run, making it very difficult to build a faithful control flow graph. For now, the new EH instructions cannot represent SEH filter expressions, and frontends must outline them ahead of time. Local variables of the parent function can be escaped and accessed using the `llvm.localescape` and `llvm.localrecover` intrinsics.

New exception handling instructions

The primary design goal of the new EH instructions is to support funclet generation while preserving information about the CFG so that SSA formation still works. As a secondary goal, they are designed to be generic across MSVC and Itanium C++ exceptions. They make very few assumptions about the data required by the personality, so long as it uses the familiar core EH actions: `catch`, `cleanup`, and `terminate`. However, the new instructions are hard to modify without knowing details of the EH personality. While they can be used to represent Itanium EH, the `landingpad` model is strictly better for optimization purposes.

The following new instructions are considered “exception handling pads”, in that they must be the first non-phi instruction of a basic block that may be the unwind destination of an EH flow edge: `catchswitch`, `catchpad`, and `cleanuppad`. As with landingpads, when entering a try scope, if the frontend encounters a call site that may throw an exception, it should emit an `invoke` that unwinds to a `catchswitch` block. Similarly, inside the scope of a C++ object with a destructor, `invokes` should unwind to a `cleanuppad`.

New instructions are also used to mark the points where control is transferred out of a catch/cleanup handler (which will correspond to exits from the generated funclet). A catch handler which reaches its end by normal execution executes a `catchret` instruction, which is a terminator indicating where in the function control is returned to. A cleanup handler which reaches its end by normal execution executes a `cleanupret` instruction, which is a terminator indicating where the active exception will unwind to next.

Each of these new EH pad instructions has a way to identify which action should be considered after this action. The `catchswitch` instruction is a terminator and has an unwind destination operand analogous to the unwind destination of an `invoke`. The `cleanuppad` instruction is not a terminator, so the unwind destination is stored on the `cleanupret` instruction instead. Successfully executing a catch handler should resume normal control flow, so neither `catchpad` nor `catchret` instructions can unwind. All of these “unwind edges” may refer to a basic block

that contains an EH pad instruction, or they may unwind to the caller. Unwinding to the caller has roughly the same semantics as the `resume` instruction in the landingpad model. When inlining through an `invoke`, instructions that unwind to the caller are hooked up to unwind to the unwind destination of the call site.

Putting things together, here is a hypothetical lowering of some C++ that uses all of the new IR instructions:

```
struct Cleanup {
    Cleanup();
    ~Cleanup();
    int m;
};
void may_throw();
int f() noexcept {
    try {
        Cleanup obj;
        may_throw();
    } catch (int e) {
        may_throw();
        return e;
    }
    return 0;
}
```

```
define i32 @f() nounwind personality i32 (...)@__CxxFrameHandler3 {
entry:
    %obj = alloca %struct.Cleanup, align 4
    %e = alloca i32, align 4
    %call = invoke %struct.Cleanup* @"\01??0Cleanup@@QEAA@XZ"(%struct.Cleanup* nonnull
↳%obj)
        to label %invoke.cont unwind label %lpad.catch

invoke.cont:
    ; preds = %entry
    invoke void @"\01?may_throw@@YAXXZ"()
        to label %invoke.cont.2 unwind label %lpad.cleanup

invoke.cont.2:
    ; preds = %invoke.cont
    call void @"\01??_DCleanup@@QEAA@XZ"(%struct.Cleanup* nonnull %obj) nounwind
    br label %return

return:
    ; preds = %invoke.cont.3, %invoke.
↳cont.2
    %retval.0 = phi i32 [ 0, %invoke.cont.2 ], [ %3, %invoke.cont.3 ]
    ret i32 %retval.0

lpad.cleanup:
    ; preds = %invoke.cont.2
    %0 = cleanuppadd within none []
    call void @"\01??1Cleanup@@QEAA@XZ"(%struct.Cleanup* nonnull %obj) nounwind
    cleanupret %0 unwind label %lpad.catch

lpad.catch:
    ; preds = %lpad.cleanup, %entry
    %1 = catchswitch within none [label %catch.body] unwind label %lpad.terminate

catch.body:
    ; preds = %lpad.catch
    %catch = catchpadd within %1 [%rtti.TypeDescriptor2* @"\01??_R0H@8", i32 0, i32* %e]
    invoke void @"\01?may_throw@@YAXXZ"()
        to label %invoke.cont.3 unwind label %lpad.terminate

invoke.cont.3:
    ; preds = %catch.body
```

```

    %3 = load i32, i32* %e, align 4
    catchret from %catch to label %return

lpad.terminate:                                ; preds = %catch.body, %lpad.catch
    cleanuppad within none []
    call void @"\01?terminate@@YAXXZ"
    unreachable
}

```

Funclet parent tokens

In order to produce tables for EH personalities that use funclets, it is necessary to recover the nesting that was present in the source. This funclet parent relationship is encoded in the IR using tokens produced by the new “pad” instructions. The token operand of a “pad” or “ret” instruction indicates which funclet it is in, or “none” if it is not nested within another funclet.

The `catchpad` and `cleanuppad` instructions establish new funclets, and their tokens are consumed by other “pad” instructions to establish membership. The `catchswitch` instruction does not create a funclet, but it produces a token that is always consumed by its immediate successor `catchpad` instructions. This ensures that every catch handler modelled by a `catchpad` belongs to exactly one `catchswitch`, which models the dispatch point after a C++ try.

Here is an example of what this nesting looks like using some hypothetical C++ code:

```

void f() {
    try {
        throw;
    } catch (...) {
        try {
            throw;
        } catch (...) {
        }
    }
}

```

```

define void @f() #0 personality i8* bitcast (i32 (...)* @__CxxFrameHandler3 to i8*) {
entry:
    invoke void @_CxxThrowException(i8* null, %eh.ThrowInfo* null) #1
        to label %unreachable unwind label %catch.dispatch

catch.dispatch:                                ; preds = %entry
    %0 = catchswitch within none [label %catch] unwind to caller

catch:                                          ; preds = %catch.dispatch
    %1 = catchpad within %0 [i8* null, i32 64, i8* null]
    invoke void @_CxxThrowException(i8* null, %eh.ThrowInfo* null) #1
        to label %unreachable unwind label %catch.dispatch2

catch.dispatch2:                              ; preds = %catch
    %2 = catchswitch within %1 [label %catch3] unwind to caller

catch3:                                       ; preds = %catch.dispatch2
    %3 = catchpad within %2 [i8* null, i32 64, i8* null]
    catchret from %3 to label %try.cont

try.cont:                                    ; preds = %catch3
    catchret from %1 to label %try.cont6
}

```

```
try.cont6:                                ; preds = %try.cont
    ret void

unreachable:                              ; preds = %catch, %entry
    unreachable
}
```

The “inner” `catchswitch` consumes `%1` which is produced by the outer `catchswitch`.

Funclet transitions

The EH tables for personalities that use funclets make implicit use of the funclet nesting relationship to encode unwind destinations, and so are constrained in the set of funclet transitions they can represent. The related LLVM IR instructions accordingly have constraints that ensure encodability of the EH edges in the flow graph.

A `catchswitch`, `catchpad`, or `cleanuppad` is said to be “entered” when it executes. It may subsequently be “exited” by any of the following means:

- A `catchswitch` is immediately exited when none of its constituent `catchpads` are appropriate for the in-flight exception and it unwinds to its unwind destination or the caller.
- A `catchpad` and its parent `catchswitch` are both exited when a `catchret` from the `catchpad` is executed.
- A `cleanuppad` is exited when a `cleanupret` from it is executed.
- Any of these pads is exited when control unwinds to the function’s caller, either by a `call` which unwinds all the way to the function’s caller, a nested `catchswitch` marked “unwinds to caller”, or a nested `cleanuppad`’s `cleanupret` marked “unwinds to caller”.
- Any of these pads is exited when an unwind edge (from an `invoke`, nested `catchswitch`, or nested `cleanuppad`’s `cleanupret`) unwinds to a destination pad that is not a descendant of the given pad.

Note that the `ret` instruction is *not* a valid way to exit a funclet pad; it is undefined behavior to execute a `ret` when a pad has been entered but not exited.

A single unwind edge may exit any number of pads (with the restrictions that the edge from a `catchswitch` must exit at least itself, and the edge from a `cleanupret` must exit at least its `cleanuppad`), and then must enter exactly one pad, which must be distinct from all the exited pads. The parent of the pad that an unwind edge enters must be the most-recently-entered not-yet-exited pad (after exiting from any pads that the unwind edge exits), or “none” if there is no such pad. This ensures that the stack of executing funclets at run-time always corresponds to some path in the funclet pad tree that the parent tokens encode.

All unwind edges which exit any given funclet pad (including `cleanupret` edges exiting their `cleanuppad` and `catchswitch` edges exiting their `catchswitch`) must share the same unwind destination. Similarly, any funclet pad which may be exited by unwind to caller must not be exited by any exception edges which unwind anywhere other than the caller. This ensures that each funclet as a whole has only one unwind destination, which EH tables for funclet personalities may require. Note that any unwind edge which exits a `catchpad` also exits its parent `catchswitch`, so this implies that for any given `catchswitch`, its unwind destination must also be the unwind destination of any unwind edge that exits any of its constituent `catchpads`. Because `catchswitch` has no `nounwind` variant, and because IR producers are not *required* to annotate calls which will not unwind as `nounwind`, it is legal to nest a `call` or an “unwind to caller” `catchswitch` within a funclet pad that has an unwind destination other than caller; it is undefined behavior for such a `call` or `catchswitch` to unwind.

Finally, the funclet pads’ unwind destinations cannot form a cycle. This ensures that EH lowering can construct “try regions” with a tree-like structure, which funclet-based personalities may require.

4.9 LLVM Link Time Optimization: Design and Implementation

- *Description*
- *Design Philosophy*
 - *Example of link time optimization*
 - *Alternative Approaches*
- *Multi-phase communication between libLTO and linker*
 - *Phase 1 : Read LLVM Bitcode Files*
 - *Phase 2 : Symbol Resolution*
 - *Phase 3 : Optimize Bitcode Files*
 - *Phase 4 : Symbol Resolution after optimization*
- *libLTO*
 - *lto_module_t*
 - *lto_code_gen_t*

4.9.1 Description

LLVM features powerful intermodular optimizations which can be used at link time. Link Time Optimization (LTO) is another name for intermodular optimization when performed during the link stage. This document describes the interface and design between the LTO optimizer and the linker.

4.9.2 Design Philosophy

The LLVM Link Time Optimizer provides complete transparency, while doing intermodular optimization, in the compiler tool chain. Its main goal is to let the developer take advantage of intermodular optimizations without making any significant changes to the developer's makefiles or build system. This is achieved through tight integration with the linker. In this model, the linker treats LLVM bitcode files like native object files and allows mixing and matching among them. The linker uses *libLTO*, a shared object, to handle LLVM bitcode files. This tight integration between the linker and LLVM optimizer helps to do optimizations that are not possible in other models. The linker input allows the optimizer to avoid relying on conservative escape analysis.

Example of link time optimization

The following example illustrates the advantages of LTO's integrated approach and clean interface. This example requires a system linker which supports LTO through the interface described in this document. Here, clang transparently invokes system linker.

- Input source file `a.c` is compiled into LLVM bitcode form.
- Input source file `main.c` is compiled into native object code.

```
--- a.h ---
extern int foo1(void);
extern void foo2(void);
```

```
extern void foo4(void);

--- a.c ---
#include "a.h"

static signed int i = 0;

void foo2(void) {
    i = -1;
}

static int foo3() {
    foo4();
    return 10;
}

int foo1(void) {
    int data = 0;

    if (i < 0)
        data = foo3();

    data = data + 42;
    return data;
}

--- main.c ---
#include <stdio.h>
#include "a.h"

void foo4(void) {
    printf("Hi\n");
}

int main() {
    return foo1();
}
```

To compile, run:

```
% clang -flto -c a.c -o a.o          # <-- a.o is LLVM bytecode file
% clang -c main.c -o main.o         # <-- main.o is native object file
% clang -flto a.o main.o -o main    # <-- standard link command with -flto
```

- In this example, the linker recognizes that `foo2()` is an externally visible symbol defined in LLVM bytecode file. The linker completes its usual symbol resolution pass and finds that `foo2()` is not used anywhere. This information is used by the LLVM optimizer and it removes `foo2()`.
- As soon as `foo2()` is removed, the optimizer recognizes that condition `i < 0` is always false, which means `foo3()` is never used. Hence, the optimizer also removes `foo3()`.
- And this in turn, enables linker to remove `foo4()`.

This example illustrates the advantage of tight integration with the linker. Here, the optimizer can not remove `foo3()` without the linker's input.

Alternative Approaches

Compiler driver invokes link time optimizer separately. In this model the link time optimizer is not able to take advantage of information collected during the linker's normal symbol resolution phase. In the above example, the optimizer can not remove `f002()` without the linker's input because it is externally visible. This in turn prohibits the optimizer from removing `f003()`.

Use separate tool to collect symbol information from all object files. In this model, a new, separate, tool or library replicates the linker's capability to collect information for link time optimization. Not only is this code duplication difficult to justify, but it also has several other disadvantages. For example, the linking semantics and the features provided by the linker on various platform are not unique. This means, this new tool needs to support all such features and platforms in one super tool or a separate tool per platform is required. This increases maintenance cost for link time optimizer significantly, which is not necessary. This approach also requires staying synchronized with linker developments on various platforms, which is not the main focus of the link time optimizer. Finally, this approach increases end user's build time due to the duplication of work done by this separate tool and the linker itself.

4.9.3 Multi-phase communication between libLTO and linker

The linker collects information about symbol definitions and uses in various link objects which is more accurate than any information collected by other tools during typical build cycles. The linker collects this information by looking at the definitions and uses of symbols in native `.o` files and using symbol visibility information. The linker also uses user-supplied information, such as a list of exported symbols. LLVM optimizer collects control flow information, data flow information and knows much more about program structure from the optimizer's point of view. Our goal is to take advantage of tight integration between the linker and the optimizer by sharing this information during various linking phases.

Phase 1 : Read LLVM Bitcode Files

The linker first reads all object files in natural order and collects symbol information. This includes native object files as well as LLVM bitcode files. To minimize the cost to the linker in the case that all `.o` files are native object files, the linker only calls `lto_module_create()` when a supplied object file is found to not be a native object file. If `lto_module_create()` returns that the file is an LLVM bitcode file, the linker then iterates over the module using `lto_module_get_symbol_name()` and `lto_module_get_symbol_attribute()` to get all symbols defined and referenced. This information is added to the linker's global symbol table.

The `lto*` functions are all implemented in a shared object `libLTO`. This allows the LLVM LTO code to be updated independently of the linker tool. On platforms that support it, the shared object is lazily loaded.

Phase 2 : Symbol Resolution

In this stage, the linker resolves symbols using global symbol table. It may report undefined symbol errors, read archive members, replace weak symbols, etc. The linker is able to do this seamlessly even though it does not know the exact content of input LLVM bitcode files. If dead code stripping is enabled then the linker collects the list of live symbols.

Phase 3 : Optimize Bitcode Files

After symbol resolution, the linker tells the LTO shared object which symbols are needed by native object files. In the example above, the linker reports that only `f001()` is used by native object files using

`lto_codegen_add_must_preserve_symbol()`. Next the linker invokes the LLVM optimizer and code generators using `lto_codegen_compile()` which returns a native object file creating by merging the LLVM bitcode files and applying various optimization passes.

Phase 4 : Symbol Resolution after optimization

In this phase, the linker reads optimized a native object file and updates the internal global symbol table to reflect any changes. The linker also collects information about any changes in use of external symbols by LLVM bitcode files. In the example above, the linker notes that `foo4()` is not used any more. If dead code stripping is enabled then the linker refreshes the live symbol information appropriately and performs dead code stripping.

After this phase, the linker continues linking as if it never saw LLVM bitcode files.

4.9.4 libLTO

libLTO is a shared object that is part of the LLVM tools, and is intended for use by a linker. libLTO provides an abstract C interface to use the LLVM interprocedural optimizer without exposing details of LLVM's internals. The intention is to keep the interface as stable as possible even when the LLVM optimizer continues to evolve. It should even be possible for a completely different compilation technology to provide a different libLTO that works with their object files and the standard linker tool.

`lto_module_t`

A non-native object file is handled via an `lto_module_t`. The following functions allow the linker to check if a file (on disk or in a memory buffer) is a file which libLTO can process:

```
lto_module_is_object_file(const char*)
lto_module_is_object_file_for_target(const char*, const char*)
lto_module_is_object_file_in_memory(const void*, size_t)
lto_module_is_object_file_in_memory_for_target(const void*, size_t, const char*)
```

If the object file can be processed by libLTO, the linker creates a `lto_module_t` by using one of:

```
lto_module_create(const char*)
lto_module_create_from_memory(const void*, size_t)
```

and when done, the handle is released via

```
lto_module_dispose(lto_module_t)
```

The linker can introspect the non-native object file by getting the number of symbols and getting the name and attributes of each symbol via:

```
lto_module_get_num_symbols(lto_module_t)
lto_module_get_symbol_name(lto_module_t, unsigned int)
lto_module_get_symbol_attribute(lto_module_t, unsigned int)
```

The attributes of a symbol include the alignment, visibility, and kind.

`lto_code_gen_t`

Once the linker has loaded each non-native object files into an `lto_module_t`, it can request libLTO to process them all and generate a native object file. This is done in a couple of steps. First, a code generator is created with:

```
lto_codegen_create()
```

Then, each non-native object file is added to the code generator with:

```
lto_codegen_add_module(lto_code_gen_t, lto_module_t)
```

The linker then has the option of setting some codegen options. Whether or not to generate DWARF debug info is set with:

```
lto_codegen_set_debug_model(lto_code_gen_t)
```

which kind of position independence is set with:

```
lto_codegen_set_pic_model(lto_code_gen_t)
```

And each symbol that is referenced by a native object file or otherwise must not be optimized away is set with:

```
lto_codegen_add_must_preserve_symbol(lto_code_gen_t, const char*)
```

After all these settings are done, the linker requests that a native object file be created from the modules with the settings using:

```
lto_codegen_compile(lto_code_gen_t, size*)
```

which returns a pointer to a buffer containing the generated native object file. The linker then parses that and links it with the rest of the native object files.

4.10 Segmented Stacks in LLVM

- *Introduction*
- *Implementation Details*
 - *Allocating Stacklets*
 - *Variable Sized Allocas*

4.10.1 Introduction

Segmented stack allows stack space to be allocated incrementally than as a monolithic chunk (of some worst case size) at thread initialization. This is done by allocating stack blocks (henceforth called *stacklets*) and linking them into a doubly linked list. The function prologue is responsible for checking if the current stacklet has enough space for the function to execute; and if not, call into the libgcc runtime to allocate more stack space. Segmented stacks are enabled with the "split-stack" attribute on LLVM functions.

The runtime functionality is [already there in libgcc](#).

4.10.2 Implementation Details

Allocating Stacklets

As mentioned above, the function prologue checks if the current stacklet has enough space. The current approach is to use a slot in the TCB to store the current stack limit (minus the amount of space needed to allocate a new block) - this slot's offset is again dictated by `libgcc`. The generated assembly looks like this on x86-64:

```
leaq    -8(%rsp), %r10
cmpq    %fs:112,  %r10
jg      .LBB0_2

# More stack space needs to be allocated
movabsq $8, %r10    # The amount of space needed
movabsq $0, %r11    # The total size of arguments passed on stack
callq   __morestack
ret     # The reason for this extra return is explained below
.LBB0_2:
# Usual prologue continues here
```

The size of function arguments on the stack needs to be passed to `__morestack` (this function is implemented in `libgcc`) since that number of bytes has to be copied from the previous stacklet to the current one. This is so that SP (and FP) relative addressing of function arguments work as expected.

The unusual `ret` is needed to have the function which made a call to `__morestack` return correctly. `__morestack`, instead of returning, calls into `.LBB0_2`. This is possible since both, the size of the `ret` instruction and the PC of call to `__morestack` are known. When the function body returns, control is transferred back to `__morestack`. `__morestack` then de-allocates the new stacklet, restores the correct SP value, and does a second return, which returns control to the correct caller.

Variable Sized Allocas

The section on *allocating stacklets* automatically assumes that every stack frame will be of fixed size. However, LLVM allows the use of the `llvm.alloca` intrinsic to allocate dynamically sized blocks of memory on the stack. When faced with such a variable-sized `alloca`, code is generated to:

- Check if the current stacklet has enough space. If yes, just bump the SP, like in the normal case.
- If not, generate a call to `libgcc`, which allocates the memory from the heap.

The memory allocated from the heap is linked into a list in the current stacklet, and freed along with the same. This prevents a memory leak.

4.11 TableGen Fundamentals

4.11.1 Moved

The TableGen fundamentals documentation has moved to a directory on its own and is now available at [TableGen](#). Please, change your links to that page.

4.12 TableGen

- *Introduction*
- *The TableGen program*
 - *Running TableGen*
 - *Example*
- *Syntax*
 - *Basic concepts*
- *TableGen backends*
- *TableGen Deficiencies*

4.12.1 TableGen BackEnds

- *Introduction*
- *LLVM BackEnds*
 - *CodeEmitter*
 - *RegisterInfo*
 - *InstrInfo*
 - *AsmWriter*
 - *AsmMatcher*
 - *Disassembler*
 - *PseudoLowering*
 - *CallingConv*
 - *DAGISel*
 - *DFAPacketizer*
 - *FastISel*
 - *Subtarget*
 - *Intrinsic*
 - *OptParserDefs*
 - *CTags*
 - *X86EVEX2VEX*
- *Clang BackEnds*
 - *ClangAttrClasses*
 - *ClangAttrParserStringSwitches*
 - *ClangAttrImpl*
 - *ClangAttrList*

- *ClangAttrPCHRead*
 - *ClangAttrPCHWrite*
 - *ClangAttrSpellings*
 - *ClangAttrSpellingListIndex*
 - *ClangAttrVisitor*
 - *ClangAttrTemplateInstantiate*
 - *ClangAttrParsedAttrList*
 - *ClangAttrParsedAttrImpl*
 - *ClangAttrParsedAttrKinds*
 - *ClangAttrDump*
 - *ClangDiagsDefs*
 - *ClangDiagGroups*
 - *ClangDiagsIndexName*
 - *ClangCommentNodes*
 - *ClangDeclNodes*
 - *ClangStmtNodes*
 - *ClangSACheckers*
 - *ClangCommentHTMLTags*
 - *ClangCommentHTMLTagsProperties*
 - *ClangCommentHTMLNamedCharacterReferences*
 - *ClangCommentCommandInfo*
 - *ClangCommentCommandList*
 - *ArmNeon*
 - *ArmNeonSema*
 - *ArmNeonTest*
 - *AttrDocs*
- *How to write a back-end*

Introduction

TableGen backends are at the core of TableGen’s functionality. The source files provide the semantics to a generated (in memory) structure, but it’s up to the backend to print this out in a way that is meaningful to the user (normally a C program including a file or a textual list of warnings, options and error messages).

TableGen is used by both LLVM and Clang with very different goals. LLVM uses it as a way to automate the generation of massive amounts of information regarding instructions, schedules, cores and architecture features. Some backends generate output that is consumed by more than one source file, so they need to be created in a way that is easy to use pre-processor tricks. Some backends can also print C code structures, so that they can be directly included as-is.

Clang, on the other hand, uses it mainly for diagnostic messages (errors, warnings, tips) and attributes, so more on the

textual end of the scale.

LLVM BackEnds

Warning: This document is raw. Each section below needs three sub-sections: description of its purpose with a list of users, output generated from generic input, and finally why it needed a new backend (in case there's something similar).

Overall, each backend will take the same TableGen file type and transform into similar output for different targets/uses. There is an implicit contract between the TableGen files, the back-ends and their users.

For instance, a global contract is that each back-end produces macro-guarded sections. Based on whether the file is included by a header or a source file, or even in which context of each file the include is being used, you have to define a macro just before including it, to get the right output:

```
#define GET_REGINFO_TARGET_DESC
#include "ARMGenRegisterInfo.inc"
```

And just part of the generated file would be included. This is useful if you need the same information in multiple formats (instantiation, initialization, getter/setter functions, etc) from the same source TableGen file without having to re-compile the TableGen file multiple times.

Sometimes, multiple macros might be defined before the same include file to output multiple blocks:

```
#define GET_REGISTER_MATCHER
#define GET_SUBTARGET_FEATURE_NAME
#define GET_MATCHER_IMPLEMENTATION
#include "ARMGenAsmMatcher.inc"
```

The macros will be undef'd automatically as they're used, in the include file.

On all LLVM back-ends, the `llvm-tblgen` binary will be executed on the root TableGen file `<Target>.td`, which should include all others. This guarantees that all information needed is accessible, and that no duplication is needed in the TableGen files.

CodeEmitter

Purpose: CodeEmitterGen uses the descriptions of instructions and their fields to construct an automated code emitter: a function that, given a `MachineInstr`, returns the (currently, 32-bit unsigned) value of the instruction.

Output: C++ code, implementing the target's CodeEmitter class by overriding the virtual functions as `<Target>CodeEmitter::function()`.

Usage: Used to include directly at the end of `<Target>MCCodeEmitter.cpp`.

RegisterInfo

Purpose: This tablegen backend is responsible for emitting a description of a target register file for a code generator. It uses instances of the `Register`, `RegisterAliases`, and `RegisterClass` classes to gather this information.

Output: C++ code with enums and structures representing the register mappings, properties, masks, etc.

Usage: Both on `<Target>BaseRegisterInfo` and `<Target>MCTargetDesc` (headers and source files) with macros defining in which they are for declaration vs. initialization issues.

InstrInfo

Purpose: This tablegen backend is responsible for emitting a description of the target instruction set for the code generator. (what are the differences from CodeEmitter?)

Output: C++ code with enums and structures representing the instruction mappings, properties, masks, etc.

Usage: Both on `<Target>BaseInstrInfo` and `<Target>MCTargetDesc` (headers and source files) with macros defining in which they are for declaration vs. initialization issues.

AsmWriter

Purpose: Emits an assembly printer for the current target.

Output: Implementation of `<Target>InstPrinter::printInstruction()`, among other things.

Usage: Included directly into `InstPrinter/<Target>InstPrinter.cpp`.

AsmMatcher

Purpose: Emits a target specifier matcher for converting parsed assembly operands in the `MCIInst` structures. It also emits a matcher for custom operand parsing. Extensive documentation is written on the `AsmMatcherEmitter.cpp` file.

Output: Assembler parsers' matcher functions, declarations, etc.

Usage: Used in back-ends' `AsmParser/<Target>AsmParser.cpp` for building the `AsmParser` class.

Disassembler

Purpose: Contains disassembler table emitters for various architectures. Extensive documentation is written on the `DisassemblerEmitter.cpp` file.

Output: Decoding tables, static decoding functions, etc.

Usage: Directly included in `Disassembler/<Target>Disassembler.cpp` to cater for all default decodings, after all hand-made ones.

PseudoLowering

Purpose: Generate pseudo instruction lowering.

Output: Implements `<Target>AsmPrinter::emitPseudoExpansionLowering()`.

Usage: Included directly into `<Target>AsmPrinter.cpp`.

CallingConv

Purpose: Responsible for emitting descriptions of the calling conventions supported by this target.

Output: Implement static functions to deal with calling conventions chained by matching styles, returning false on no match.

Usage: Used in `ISelLowering` and `FastISel` as function pointers to implementation returned by a CC selection function.

DAGISel

Purpose: Generate a DAG instruction selector.

Output: Creates huge functions for automating DAG selection.

Usage: Included in `<Target>ISelDAGToDAG.cpp` inside the target's implementation of `SelectionDAGISel`.

DFAPacketizer

Purpose: This class parses the `Schedule.td` file and produces an API that can be used to reason about whether an instruction can be added to a packet on a VLIW architecture. The class internally generates a deterministic finite automaton (DFA) that models all possible mappings of machine instructions to functional units as instructions are added to a packet.

Output: Scheduling tables for GPU back-ends (Hexagon, AMD).

Usage: Included directly on `<Target>InstrInfo.cpp`.

FastISel

Purpose: This tablegen backend emits code for use by the “fast” instruction selection algorithm. See the comments at the top of `lib/CodeGen/SelectionDAG/FastISel.cpp` for background. This file scans through the target's tablegen instruction-info files and extracts instructions with obvious-looking patterns, and it emits code to look up these instructions by type and operator.

Output: Generates `Predicate` and `FastEmit` methods.

Usage: Implements private methods of the targets' implementation of `FastISel` class.

Subtarget

Purpose: Generate subtarget enumerations.

Output: Enums, globals, local tables for sub-target information.

Usage: Populates `<Target>Subtarget` and `MCTargetDesc/<Target>MCTargetDesc` files (both headers and source).

Intrinsic

Purpose: Generate (target) intrinsic information.

OptParserDefs

Purpose: Print enum values for a class.

CTags

Purpose: This tablegen backend emits an index of definitions in `ctags(1)` format. A helper script, `utils/TableGen/tdtags`, provides an easier-to-use interface; run `'tdtags -H'` for documentation.

X86EVEX2VEX

Purpose: This X86 specific tablegen backend emits tables that map EVEX encoded instructions to their VEX encoded identical instruction.

Clang BackEnds

ClangAttrClasses

Purpose: Creates `Attrs.inc`, which contains semantic attribute class declarations for any attribute in `Attr.td` that has not set `ASTNode = 0`. This file is included as part of `Attr.h`.

ClangAttrParserStringSwitches

Purpose: Creates `AttrParserStringSwitches.inc`, which contains `StringSwitch::Case` statements for parser-related string switches. Each switch is given its own macro (such as `CLANG_ATTR_ARG_CONTEXT_LIST`, or `CLANG_ATTR_IDENTIFIER_ARG_LIST`), which is expected to be defined before including `AttrParserStringSwitches.inc`, and undefined after.

ClangAttrImpl

Purpose: Creates `AttrImpl.inc`, which contains semantic attribute class definitions for any attribute in `Attr.td` that has not set `ASTNode = 0`. This file is included as part of `AttrImpl.cpp`.

ClangAttrList

Purpose: Creates `AttrList.inc`, which is used when a list of semantic attribute identifiers is required. For instance, `AttrKinds.h` includes this file to generate the list of `attr::Kind` enumeration values. This list is separated out into multiple categories: attributes, inheritable attributes, and inheritable parameter attributes. This categorization happens automatically based on information in `Attr.td` and is used to implement the `classof` functionality required for `dyn_cast` and similar APIs.

ClangAttrPCHRead

Purpose: Creates `AttrPCHRead.inc`, which is used to deserialize attributes in the `ASTReader::ReadAttributes` function.

ClangAttrPCHWrite

Purpose: Creates `AttrPCHWrite.inc`, which is used to serialize attributes in the `ASTWriter::WriteAttributes` function.

ClangAttrSpellings

Purpose: Creates `AttrSpellings.inc`, which is used to implement the `__has_attribute` feature test macro.

ClangAttrSpellingListIndex

Purpose: Creates `AttrSpellingListIndex.inc`, which is used to map parsed attribute spellings (including which syntax or scope was used) to an attribute spelling list index. These spelling list index values are internal implementation details exposed via `AttributeList::getAttributeSpellingListIndex`.

ClangAttrVisitor

Purpose: Creates `AttrVisitor.inc`, which is used when implementing recursive AST visitors.

ClangAttrTemplateInstantiate

Purpose: Creates `AttrTemplateInstantiate.inc`, which implements the `instantiateTemplateAttribute` function, used when instantiating a template that requires an attribute to be cloned.

ClangAttrParsedAttrList

Purpose: Creates `AttrParsedAttrList.inc`, which is used to generate the `AttributeList::Kind` parsed attribute enumeration.

ClangAttrParsedAttrImpl

Purpose: Creates `AttrParsedAttrImpl.inc`, which is used by `AttributeList.cpp` to implement several functions on the `AttributeList` class. This functionality is implemented via the `AttrInfoMap ParsedAttrInfo` array, which contains one element per parsed attribute object.

ClangAttrParsedAttrKinds

Purpose: Creates `AttrParsedAttrKinds.inc`, which is used to implement the `AttributeList::getKind` function, mapping a string (and syntax) to a parsed attribute `AttributeList::Kind` enumeration.

ClangAttrDump

Purpose: Creates `AttrDump.inc`, which dumps information about an attribute. It is used to implement `ASTDumper::dumpAttr`.

ClangDiagsDefs

Generate Clang diagnostics definitions.

ClangDiagGroups

Generate Clang diagnostic groups.

ClangDiagsIndexName

Generate Clang diagnostic name index.

ClangCommentNodes

Generate Clang AST comment nodes.

ClangDeclNodes

Generate Clang AST declaration nodes.

ClangStmtNodes

Generate Clang AST statement nodes.

ClangSACheckers

Generate Clang Static Analyzer checkers.

ClangCommentHTMLTags

Generate efficient matchers for HTML tag names that are used in documentation comments.

ClangCommentHTMLTagsProperties

Generate efficient matchers for HTML tag properties.

ClangCommentHTMLNamedCharacterReferences

Generate function to translate named character references to UTF-8 sequences.

ClangCommentCommandInfo

Generate command properties for commands that are used in documentation comments.

ClangCommentCommandList

Generate list of commands that are used in documentation comments.

ArmNeon

Generate arm_neon.h for clang.

ArmNeonSema

Generate ARM NEON sema support for clang.

ArmNeonTest

Generate ARM NEON tests for clang.

AttrDocs

Purpose: Creates `AttributeReference.rst` from `AttrDocs.td`, and is used for documenting user-facing attributes.

How to write a back-end

TODO.

Until we get a step-by-step HowTo for writing TableGen backends, you can at least grab the boilerplate (build system, new files, etc.) from Clang's r173931.

TODO: How they work, how to write one. This section should not contain details about any particular backend, except maybe `-print-enums` as an example. This should highlight the APIs in `TableGen/Record.h`.

4.12.2 TableGen Language Reference

- *Introduction*
- *Notation*
- *Lexical Analysis*
- *Syntax*
 - *classes*
 - *Declarations*
 - *Types*
 - *Values*
 - *Bodies*
 - *def*
 - *defm*
 - *foreach*
 - *Top-Level let*
 - *multiclass*

Warning: This document is extremely rough. If you find something lacking, please fix it, file a documentation bug, or ask about it on [llvm-dev](#).

Introduction

This document is meant to be a normative spec about the TableGen language in and of itself (i.e. how to understand a given construct in terms of how it affects the final set of records represented by the TableGen file). If you are unsure if this document is really what you are looking for, please read the *[introduction to TableGen](#)* first.

Notation

The lexical and syntax notation used here is intended to imitate [Python's](#). In particular, for lexical definitions, the productions operate at the character level and there is no implied whitespace between elements. The syntax definitions operate at the token level, so there is implied whitespace between tokens.

Lexical Analysis

TableGen supports BCPL (`// ...`) and nestable C-style (`/* ... */`) comments.

The following is a listing of the basic punctuation tokens:

`- + [] { } () < > : ; . = ? #`

Numeric literals take one of the following forms:

```
TokInteger      ::=  DecimalInteger | HexInteger | BinInteger
DecimalInteger  ::=  ["+" | "-"] ("0"..."9")+
HexInteger      ::=  "0x" ("0"..."9" | "a"..."f" | "A"..."F")+
BinInteger      ::=  "0b" ("0" | "1")+
```

One aspect to note is that the *DecimalInteger* token *includes* the + or -, as opposed to having + and - be unary operators as most languages do.

Also note that *BinInteger* creates a value of type `bits<n>` (where `n` is the number of bits). This will implicitly convert to integers when needed.

TableGen has identifier-like tokens:

```
ualpha          ::=  "a"..."z" | "A"..."Z" | "_"
TokIdentifier    ::=  ("0"..."9")* ualpha (ualpha | "0"..."9")*
TokVarName       ::=  "$" ualpha (ualpha | "0"..."9")*
```

Note that unlike most languages, TableGen allows *TokIdentifier* to begin with a number. In case of ambiguity, a token will be interpreted as a numeric literal rather than an identifier.

TableGen also has two string-like literals:

```
TokString        ::=  ''' <non-''' characters and C-like escapes> '''
TokCodeFragment  ::=  "[{<shortest text not containing ">"> "}]"
```

TokCodeFragment is essentially a multiline string literal delimited by [`{` and `}`].

Note: The current implementation accepts the following C-like escapes:

```
\\ \' \" \t \n
```

TableGen also has the following keywords:

<code>bit</code>	<code>bits</code>	<code>class</code>	<code>code</code>	<code>dag</code>
<code>def</code>	<code>foreach</code>	<code>defm</code>	<code>field</code>	<code>in</code>
<code>int</code>	<code>let</code>	<code>list</code>	<code>multiclass</code>	<code>string</code>

TableGen also has “bang operators” which have a wide variety of meanings:

```
BangOperator ::= one of
                !eq !if !head !tail !con
                !add !shl !sra !srl !and
                !or !empty !subst !foreach !strconcat
                !cast !listconcat
```

Syntax

TableGen has an `include` mechanism. It does not play a role in the syntax per se, since it is lexically replaced with the contents of the included file.

```
IncludeDirective ::= "include" TokString
```

TableGen’s top-level production consists of “objects”.

```
TableGenFile ::= Object*
Object       ::= Class | Def | Defm | Let | MultiClass | Foreach
```

classes

```
Class ::= "class" TokIdentifier [TemplateArgList] ObjectBody
```

A `class` declaration creates a record which other records can inherit from. A class can be parametrized by a list of “template arguments”, whose values can be used in the class body.

A given class can only be defined once. A `class` declaration is considered to define the class if any of the following is true:

1. The *TemplateArgList* is present.
2. The *Body* in the *ObjectBody* is present and is not empty.
3. The *BaseClassList* in the *ObjectBody* is present.

You can declare an empty class by giving an empty *TemplateArgList* and an empty *ObjectBody*. This can serve as a restricted form of forward declaration: note that records deriving from the forward-declared class will inherit no fields from it since the record expansion is done when the record is parsed.

```
TemplateArgList ::= "<" Declaration ("," Declaration) * ">"
```

Declarations

The declaration syntax is pretty much what you would expect as a C++ programmer.

```
Declaration ::= Type TokIdentifier ["=" Value]
```

It assigns the value to the identifier.

Types

```
Type ::= "string" | "code" | "bit" | "int" | "dag"  
      | "bits" "<" TokInteger ">"  
      | "list" "<" Type ">"  
      | ClassID  
ClassID ::= TokIdentifier
```

Both *string* and *code* correspond to the string type; the difference is purely to indicate programmer intention.

The *ClassID* must identify a class that has been previously declared or defined.

Values

```
Value ::= SimpleValue ValueSuffix*  
ValueSuffix ::= "{" RangeList "}"  
              | "[" RangeList "]"  
              | "." TokIdentifier  
RangeList ::= RangePiece ("," RangePiece) *  
RangePiece ::= TokInteger  
              | TokInteger "-" TokInteger  
              | TokInteger TokInteger
```

The peculiar last form of *RangePiece* is due to the fact that the “-” is included in the *TokInteger*, hence 1-5 gets lexed as two consecutive *TokInteger*’s, with values 1 and -5, instead of “1”, “-”, and “5”. The *RangeList* can be thought of as specifying “list slice” in some contexts.

SimpleValue has a number of forms:

```
SimpleValue ::= TokIdentifier
```


The value will be the variable referenced by the identifier. It can be one of:

- name of a def, such as the use of Bar in:

```
def Bar : SomeClass {
    int X = 5;
}

def Foo {
    SomeClass Baz = Bar;
}
```

- value local to a def, such as the use of Bar in:

```
def Foo {
    int Bar = 5;
    int Baz = Bar;
}
```

- a template arg of a class, such as the use of Bar in:

```
class Foo<int Bar> {
    int Baz = Bar;
}
```

- value local to a multiclass, such as the use of Bar in:

```
multiclass Foo {
    int Bar = 5;
    int Baz = Bar;
}
```

- a template arg to a multiclass, such as the use of Bar in:

```
multiclass Foo<int Bar> {
    int Baz = Bar;
}
```

SimpleValue ::= *TokInteger*

This represents the numeric value of the integer.

SimpleValue ::= *TokString+*

Multiple adjacent string literals are concatenated like in C/C++. The value is the concatenation of the strings.

SimpleValue ::= *TokCodeFragment*

The value is the string value of the code fragment.

SimpleValue ::= "?"

? represents an “unset” initializer.

```
SimpleValue ::= "{ " ValueList " } "  
ValueList  ::= [ ValueListNE ]  
ValueListNE ::= Value ( " , " Value ) *
```

This represents a sequence of bits, as would be used to initialize a `bits<n>` field (where `n` is the number of bits).

```
SimpleValue ::= ClassID "<" ValueListNE ">"
```

This generates a new anonymous record definition (as would be created by an unnamed `def` inheriting from the given class with the given template arguments) and the value is the value of that record definition.

```
SimpleValue ::= "[ " ValueList " ] " [ "<" Type ">" ]
```

A list initializer. The optional *Type* can be used to indicate a specific element type, otherwise the element type will be deduced from the given values.

```
SimpleValue ::= "( " DagArg DagArgList " ) "  
DagArgList  ::= DagArg ( " , " DagArg ) *  
DagArg      ::= Value [ ":" TokVarName ] | TokVarName
```

The initial *DagArg* is called the “operator” of the dag.

```
SimpleValue ::= BangOperator [ "<" Type ">" ] "( " ValueListNE " ) "
```

Bodies

```
ObjectBody      ::= BaseClassList Body  
BaseClassList   ::= [ ":" BaseClassListNE ]  
BaseClassListNE ::= SubClassRef ( " , " SubClassRef ) *  
SubClassRef     ::= ( ClassID | MultiClassID ) [ "<" ValueList ">" ]  
DefmID          ::= TokIdentifier
```

The version with the *MultiClassID* is only valid in the *BaseClassList* of a *defm*. The *MultiClassID* should be the name of a multiclass.

It is after parsing the base class list that the “let stack” is applied.

```
Body      ::= ";" | "{ " BodyList " } "  
BodyList  ::= BodyItem *  
BodyItem  ::= Declaration ";"  
           | "let" TokIdentifier [RangeList] "=" Value ";"
```

The `let` form allows overriding the value of an inherited field.

def

```
Def ::= "def" TokIdentifier ObjectBody
```

Defines a record whose name is given by the *TokIdentifier*. The fields of the record are inherited from the base classes and defined in the body.

Special handling occurs if this `def` appears inside a `multiclass` or a `foreach`.

defm

```
Defm ::= "defm" TokIdentifier ":" BaseClassListNE ";"
```

Note that in the *BaseClassList*, all of the `multiclass`'s must precede any `class`'s that appear.

foreach

```
Foreach ::= "foreach" Declaration "in" "{" Object* "}"
          | "foreach" Declaration "in" Object
```

The value assigned to the variable in the declaration is iterated over and the object or object list is reevaluated with the variable set at each iterated value.

Top-Level let

```
Let      ::= "let" LetList "in" "{" Object* "}"
          | "let" LetList "in" Object
LetList ::= LetItem ("," LetItem)*
LetItem ::= TokIdentifier [RangeList] "=" Value
```

This is effectively equivalent to `let` inside the body of a record except that it applies to multiple records at a time. The bindings are applied at the end of parsing the base classes of a record.

multiclass

```
MultiClass      ::= "multiclass" TokIdentifier [TemplateArgList]
                  [":" BaseMultiClassList] "{" MultiClassObject+ "}"
BaseMultiClassList ::= MultiClassID ("," MultiClassID)*
MultiClassID    ::= TokIdentifier
MultiClassObject ::= Def | Defm | Let | Foreach
```

4.12.3 TableGen Language Introduction

- *Introduction*
- *TableGen syntax*
 - *TableGen primitives*
 - * *TableGen comments*
 - * *The TableGen type system*
 - * *TableGen values and expressions*
 - *Classes and definitions*
 - * *Value definitions*
 - * *‘let’ expressions*
 - * *Class template arguments*
 - * *Multiclass definitions and instances*
 - *File scope entities*
 - * *File inclusion*
 - * *‘let’ expressions*
 - * *Looping*
- *Code Generator backend info*

Warning: This document is extremely rough. If you find something lacking, please fix it, file a documentation bug, or ask about it on llvm-dev.

Introduction

This document is not meant to be a normative spec about the TableGen language in and of itself (i.e. how to understand a given construct in terms of how it affects the final set of records represented by the TableGen file). For the formal language specification, see [TableGen Language Reference](#).

TableGen syntax

TableGen doesn’t care about the meaning of data (that is up to the backend to define), but it does care about syntax, and it enforces a simple type system. This section describes the syntax and the constructs allowed in a TableGen file.

TableGen primitives

TableGen comments

TableGen supports C++ style “//” comments, which run to the end of the line, and it also supports **nestable** “/* */” comments.

The TableGen type system

TableGen files are strongly typed, in a simple (but complete) type-system. These types are used to perform automatic conversions, check for errors, and to help interface designers constrain the input that they allow. Every *value definition* is required to have an associated type.

TableGen supports a mixture of very low-level types (such as `bit`) and very high-level types (such as `dag`). This flexibility is what allows it to describe a wide range of information conveniently and compactly. The TableGen types are:

bit A ‘bit’ is a boolean value that can hold either 0 or 1.

int The ‘int’ type represents a simple 32-bit integer value, such as 5.

string The ‘string’ type represents an ordered sequence of characters of arbitrary length.

code The *code* type represents a code fragment, which can be single/multi-line string literal.

bits<n> A ‘bits’ type is an arbitrary, but fixed, size integer that is broken up into individual bits. This type is useful because it can handle some bits being defined while others are undefined.

list<ty> This type represents a list whose elements are some other type. The contained type is arbitrary: it can even be another list type.

Class type Specifying a class name in a type context means that the defined value must be a subclass of the specified class. This is useful in conjunction with the `list` type, for example, to constrain the elements of the list to a common base class (e.g., a `list<Register>` can only contain definitions derived from the “Register” class).

dag This type represents a nestable directed graph of elements.

To date, these types have been sufficient for describing things that TableGen has been used for, but it is straight-forward to extend this list if needed.

TableGen values and expressions

TableGen allows for a pretty reasonable number of different expression forms when building up values. These forms allow the TableGen file to be written in a natural syntax and flavor for the application. The current expression forms supported include:

? uninitialized field

0b1001011 binary integer value. Note that this is sized by the number of bits given and will not be silently extended/truncated.

7 decimal integer value

0x7F hexadecimal integer value

"foo" a single-line string value, can be assigned to `string` or `code` variable.

[{ ... }] usually called a “code fragment”, but is just a multiline string literal

[**X**, **Y**, **Z**]<**type**> list value. <type> is the type of the list element and is usually optional. In rare cases, TableGen is unable to deduce the element type in which case the user must specify it explicitly.

{ **a**, **b**, **0b10** } initializer for a “bits<4>” value. 1-bit from “a”, 1-bit from “b”, 2-bits from 0b10.

value value reference

value{17} access to one bit of a value

value{15-17} access to an ordered sequence of bits of a value, in particular **value{15-17}** produces an order that is the reverse of **value{17-15}**.

DEF reference to a record definition

CLASS<val list> reference to a new anonymous definition of CLASS with the specified template arguments.

X.Y reference to the subfield of a value

list[4-7,17,2-3] A slice of the 'list' list, including elements 4,5,6,7,17,2, and 3 from it. Elements may be included multiple times.

foreach <var> = [<list>] in { <body> }

foreach <var> = [<list>] in <def> Replicate <body> or <def>, replacing instances of <var> with each value in <list>. <var> is scoped at the level of the **foreach** loop and must not conflict with any other object introduced in <body> or <def>. Currently only **defs** are expanded within <body>.

foreach <var> = 0-15 in ...

foreach <var> = {0-15,32-47} in ... Loop over ranges of integers. The braces are required for multiple ranges.

(DEF a, b) a dag value. The first element is required to be a record definition, the remaining elements in the list may be arbitrary other values, including nested 'dag' values.

!listconcat(a, b, ...) A list value that is the result of concatenating the 'a' and 'b' lists. The lists must have the same element type. More than two arguments are accepted with the result being the concatenation of all the lists given.

!strconcat(a, b, ...) A string value that is the result of concatenating the 'a' and 'b' strings. More than two arguments are accepted with the result being the concatenation of all the strings given.

str1#str2 '#' (paste) is a shorthand for **!strconcat**. It may concatenate things that are not quoted strings, in which case an implicit **!cast<string>** is done on the operand of the paste.

!cast<type>(a) A symbol of type *type* obtained by looking up the string 'a' in the symbol table. If the type of 'a' does not match *type*, TableGen aborts with an error. **!cast<string>** is a special case in that the argument must be an object defined by a 'def' construct.

!subst(a, b, c) If 'a' and 'b' are of string type or are symbol references, substitute 'b' for 'a' in 'c.' This operation is analogous to \$(subst) in GNU make.

!foreach(a, b, c) For each member of dag or list 'b' apply operator 'c.' 'a' is a dummy variable that should be declared as a member variable of an instantiated class. This operation is analogous to \$(foreach) in GNU make.

!head(a) The first element of list 'a.'

!tail(a) The 2nd-N elements of list 'a.'

!empty(a) An integer {0,1} indicating whether list 'a' is empty.

!if(a,b,c) 'b' if the result of 'int' or 'bit' operator 'a' is nonzero, 'c' otherwise.

!eq(a,b) 'bit 1' if string a is equal to string b, 0 otherwise. This only operates on string, int and bit objects. Use **!cast<string>** to compare other types of objects.

!shl(a,b) !srl(a,b) !sra(a,b) !add(a,b) !and(a,b) The usual binary and arithmetic operators.

Note that all of the values have rules specifying how they convert to values for different types. These rules allow you to assign a value like "7" to a "bits<4>" value, for example.

Classes and definitions

As mentioned in the *introduction*, classes and definitions (collectively known as ‘records’) in TableGen are the main high-level unit of information that TableGen collects. Records are defined with a `def` or `class` keyword, the record name, and an optional list of “*template arguments*”. If the record has superclasses, they are specified as a comma separated list that starts with a colon character (“:”). If *value definitions* or *let expressions* are needed for the class, they are enclosed in curly braces (“{ }”); otherwise, the record ends with a semicolon.

Here is a simple TableGen file:

```
class C { bit V = 1; }
def X : C;
def Y : C {
    string Greeting = "hello";
}
```

This example defines two definitions, `X` and `Y`, both of which derive from the `C` class. Because of this, they both get the `V` bit value. The `Y` definition also gets the `Greeting` member as well.

In general, classes are useful for collecting together the commonality between a group of records and isolating it in a single place. Also, classes permit the specification of default values for their subclasses, allowing the subclasses to override them as they wish.

Value definitions

Value definitions define named entries in records. A value must be defined before it can be referred to as the operand for another value definition or before the value is reset with a *let expression*. A value is defined by specifying a *TableGen type* and a name. If an initial value is available, it may be specified after the type with an equal sign. Value definitions require terminating semicolons.

‘let’ expressions

A record-level `let` expression is used to change the value of a value definition in a record. This is primarily useful when a superclass defines a value that a derived class or definition wants to override. `Let expressions` consist of the ‘`let`’ keyword followed by a value name, an equal sign (“=”), and a new value. For example, a new class could be added to the example above, redefining the `V` field for all of its subclasses:

```
class D : C { let V = 0; }
def Z : D;
```

In this case, the `Z` definition will have a zero value for its `V` value, despite the fact that it derives (indirectly) from the `C` class, because the `D` class overrode its value.

Class template arguments

TableGen permits the definition of parameterized classes as well as normal concrete classes. Parameterized TableGen classes specify a list of variable bindings (which may optionally have defaults) that are bound when used. Here is a simple example:

```
class FPFormat<bits<3> val> {
    bits<3> Value = val;
}
def NotFP : FPFormat<0>;
```

```
def ZeroArgFP : FPFormat<1>;
def OneArgFP  : FPFormat<2>;
def OneArgFPRW : FPFormat<3>;
def TwoArgFP   : FPFormat<4>;
def CompareFP  : FPFormat<5>;
def CondMovFP  : FPFormat<6>;
def SpecialFP  : FPFormat<7>;
```

In this case, template arguments are used as a space efficient way to specify a list of “enumeration values”, each with a “Value” field set to the specified integer.

The more esoteric forms of *TableGen expressions* are useful in conjunction with template arguments. As an example:

```
class ModRefVal<bits<2> val> {
  bits<2> Value = val;
}

def None      : ModRefVal<0>;
def Mod       : ModRefVal<1>;
def Ref       : ModRefVal<2>;
def ModRef    : ModRefVal<3>;

class Value<ModRefVal MR> {
  // Decode some information into a more convenient format, while providing
  // a nice interface to the user of the "Value" class.
  bit isMod = MR.Value{0};
  bit isRef = MR.Value{1};

  // other stuff...
}

// Example uses
def bork : Value<Mod>;
def zork : Value<Ref>;
def hork : Value<ModRef>;
```

This is obviously a contrived example, but it shows how template arguments can be used to decouple the interface provided to the user of the class from the actual internal data representation expected by the class. In this case, running `llvm-tblgen` on the example prints the following definitions:

```
def bork {          // Value
  bit isMod = 1;
  bit isRef = 0;
}
def hork {          // Value
  bit isMod = 1;
  bit isRef = 1;
}
def zork {          // Value
  bit isMod = 0;
  bit isRef = 1;
}
```

This shows that TableGen was able to dig into the argument and extract a piece of information that was requested by the designer of the “Value” class. For more realistic examples, please see existing users of TableGen, such as the X86 backend.

Multiclass definitions and instances

While classes with template arguments are a good way to factor commonality between two instances of a definition, multiclasss allow a convenient notation for defining multiple definitions at once (instances of implicitly constructed classes). For example, consider an 3-address instruction set whose instructions come in two forms: “reg = reg op reg” and “reg = reg op imm” (e.g. SPARC). In this case, you’d like to specify in one place that this commonality exists, then in a separate place indicate what all the ops are.

Here is an example TableGen fragment that shows this idea:

```
def ops;
def GPR;
def Imm;
class inst<int opc, string asmstr, dag operandlist>;

multiclass ri_inst<int opc, string asmstr> {
  def _rr : inst<opc, !strconcat(asmstr, " $dst, $src1, $src2"),
    (ops GPR:$dst, GPR:$src1, GPR:$src2)>;
  def _ri : inst<opc, !strconcat(asmstr, " $dst, $src1, $src2"),
    (ops GPR:$dst, GPR:$src1, Imm:$src2)>;
}

// Instantiations of the ri_inst multiclass.
defm ADD : ri_inst<0b111, "add">;
defm SUB : ri_inst<0b101, "sub">;
defm MUL : ri_inst<0b100, "mul">;
...
```

The name of the resultant definitions has the multidef fragment names appended to them, so this defines ADD_rr, ADD_ri, SUB_rr, etc. A defm may inherit from multiple multiclasss, instantiating definitions from each multiclass. Using a multiclass this way is exactly equivalent to instantiating the classes multiple times yourself, e.g. by writing:

```
def ops;
def GPR;
def Imm;
class inst<int opc, string asmstr, dag operandlist>;

class rrinst<int opc, string asmstr>
  : inst<opc, !strconcat(asmstr, " $dst, $src1, $src2"),
    (ops GPR:$dst, GPR:$src1, GPR:$src2)>;

class riinst<int opc, string asmstr>
  : inst<opc, !strconcat(asmstr, " $dst, $src1, $src2"),
    (ops GPR:$dst, GPR:$src1, Imm:$src2)>;

// Instantiations of the ri_inst multiclass.
def ADD_rr : rrinst<0b111, "add">;
def ADD_ri : riinst<0b111, "add">;
def SUB_rr : rrinst<0b101, "sub">;
def SUB_ri : riinst<0b101, "sub">;
def MUL_rr : rrinst<0b100, "mul">;
def MUL_ri : riinst<0b100, "mul">;
...
```

A defm can also be used inside a multiclass providing several levels of multiclass instantiations.

```
class Instruction<bits<4> opc, string Name> {
  bits<4> opcode = opc;
```

```
    string name = Name;
}

multiclass basic_r<bits<4> opc> {
    def rr : Instruction<opc, "rr">;
    def rm : Instruction<opc, "rm">;
}

multiclass basic_s<bits<4> opc> {
    defm SS : basic_r<opc>;
    defm SD : basic_r<opc>;
    def X : Instruction<opc, "x">;
}

multiclass basic_p<bits<4> opc> {
    defm PS : basic_r<opc>;
    defm PD : basic_r<opc>;
    def Y : Instruction<opc, "y">;
}

defm ADD : basic_s<0xf>, basic_p<0xf>;
...

// Results
def ADDPDrm { ...
def ADDPDrr { ...
def ADDPSrm { ...
def ADDPSrr { ...
def ADDSDrm { ...
def ADDSDrr { ...
def ADDY { ...
def ADDX { ...
```

`defm` declarations can inherit from classes too, the rule to follow is that the class list must start after the last multiclass, and there must be at least one multiclass before them.

```
class XD { bits<4> Prefix = 11; }
class XS { bits<4> Prefix = 12; }

class I<bits<4> op> {
    bits<4> opcode = op;
}

multiclass R {
    def rr : I<4>;
    def rm : I<2>;
}

multiclass Y {
    defm SS : R, XD;
    defm SD : R, XS;
}

defm Instr : Y;

// Results
def InstrSDrm {
    bits<4> opcode = { 0, 0, 1, 0 };
```

```

    bits<4> Prefix = { 1, 1, 0, 0 };
}
...
def InstrSSrr {
    bits<4> opcode = { 0, 1, 0, 0 };
    bits<4> Prefix = { 1, 0, 1, 1 };
}

```

File scope entities

File inclusion

TableGen supports the ‘include’ token, which textually substitutes the specified file in place of the include directive. The filename should be specified as a double quoted string immediately after the ‘include’ keyword. Example:

```
include "foo.td"
```

‘let’ expressions

“Let” expressions at file scope are similar to *“let” expressions within a record*, except they can specify a value binding for multiple records at a time, and may be useful in certain other cases. File-scope let expressions are really just another way that TableGen allows the end-user to factor out commonality from the records.

File-scope “let” expressions take a comma-separated list of bindings to apply, and one or more records to bind the values in. Here are some examples:

```

let isTerminator = 1, isReturn = 1, isBarrier = 1, hasCtrlDep = 1 in
  def RET : I<0xC3, RawFrm, (outs), (ins), "ret", [(X86retflag 0)]>;

let isCall = 1 in
  // All calls clobber the non-callee saved registers...
  let Defs = [EAX, ECX, EDX, FP0, FP1, FP2, FP3, FP4, FP5, FP6, ST0,
             MM0, MM1, MM2, MM3, MM4, MM5, MM6, MM7,
             XMM0, XMM1, XMM2, XMM3, XMM4, XMM5, XMM6, XMM7, EFLAGS] in {
    def CALLpcrel32 : Ii32<0xE8, RawFrm, (outs), (ins i32imm:$dst,variable_ops),
                     "call\t${dst:call}", []>;

    def CALL32r      : I<0xFF, MRM2r, (outs), (ins GR32:$dst, variable_ops),
                     "call\t{*}$dst", [(X86call GR32:$dst)]>;

    def CALL32m      : I<0xFF, MRM2m, (outs), (ins i32mem:$dst, variable_ops),
                     "call\t{*}$dst", []>;
  }

```

File-scope “let” expressions are often useful when a couple of definitions need to be added to several records, and the records do not otherwise need to be opened, as in the case with the CALL* instructions above.

It’s also possible to use “let” expressions inside multiclasses, providing more ways to factor out commonality from the records, specially if using several levels of multiclass instantiations. This also avoids the need of using “let” expressions within subsequent records inside a multiclass.

```

multiclass basic_r<bits<4> opc> {
  let Predicates = [HasSSE2] in {
    def rr : Instruction<opc, "rr">;
    def rm : Instruction<opc, "rm">;
  }
}

```

```
let Predicates = [HasSSE3] in
  def rx : Instruction<opc, "rx">;
}

multiclass basic_ss<bits<4> opc> {
  let IsDouble = 0 in
    defm SS : basic_r<opc>;

  let IsDouble = 1 in
    defm SD : basic_r<opc>;
}

defm ADD : basic_ss<0xf>;
```

Looping

TableGen supports the ‘foreach’ block, which textually replicates the loop body, substituting iterator values for iterator references in the body. Example:

```
foreach i = [0, 1, 2, 3] in {
  def R#i : Register<...>;
  def F#i : Register<...>;
}
```

This will create objects R0, R1, R2 and R3. `foreach` blocks may be nested. If there is only one item in the body the braces may be elided:

```
foreach i = [0, 1, 2, 3] in
  def R#i : Register<...>;
```

Code Generator backend info

Expressions used by code generator to describe instructions and isel patterns:

(**implicit a**) an implicitly defined physical register. This tells the dag instruction selection emitter the input pattern’s extra definitions matches implicit physical register definitions.

4.12.4 TableGen Deficiencies

- *Introduction*
- *Known Problems*

Introduction

Despite being very generic, TableGen has some deficiencies that have been pointed out numerous times. The common theme is that, while TableGen allows you to build Domain-Specific-Languages, the final languages that you create lack the power of other DSLs, which in turn increase considerably the size and complexity of TableGen files.

At the same time, TableGen allows you to create virtually any meaning of the basic concepts via custom-made back-ends, which can pervert the original design and make it very hard for newcomers to understand it.

There are some in favour of extending the semantics even more, but making sure back-ends adhere to strict rules. Others suggesting we should move to more powerful DSLs designed with specific purposes, or even re-using existing DSLs.

Known Problems

TODO: Add here frequently asked questions about why TableGen doesn't do what you want, how it might, and how we could extend/restrict it to be more use friendly.

4.12.5 Introduction

TableGen's purpose is to help a human develop and maintain records of domain-specific information. Because there may be a large number of these records, it is specifically designed to allow writing flexible descriptions and for common features of these records to be factored out. This reduces the amount of duplication in the description, reduces the chance of error, and makes it easier to structure domain specific information.

The core part of TableGen parses a file, instantiates the declarations, and hands the result off to a domain-specific *backend* for processing.

The current major users of TableGen are *The LLVM Target-Independent Code Generator* and the *Clang diagnostics and attributes*.

Note that if you work on TableGen much, and use emacs or vim, that you can find an emacs "TableGen mode" and a vim language file in the `llvm/utils/emacs` and `llvm/utils/vim` directories of your LLVM distribution, respectively.

4.12.6 The TableGen program

TableGen files are interpreted by the TableGen program: *llvm-tblgen* available on your build directory under *bin*. It is not installed in the system (or where your sysroot is set to), since it has no use beyond LLVM's build process.

Running TableGen

TableGen runs just like any other LLVM tool. The first (optional) argument specifies the file to read. If a filename is not specified, `llvm-tblgen` reads from standard input.

To be useful, one of the *backends* must be used. These backends are selectable on the command line (type '`llvm-tblgen -help`' for a list). For example, to get a list of all of the definitions that subclass a particular type (which can be useful for building up an enum list of these records), use the `-print-enums` option:

```
$ llvm-tblgen X86.td -print-enums -class=Register
AH, AL, AX, BH, BL, BP, BPL, BX, CH, CL, CX, DH, DI, DIL, DL, DX, EAX, EBP, EBX,
ECX, EDI, EDX, EFLAGS, EIP, ESI, ESP, FP0, FP1, FP2, FP3, FP4, FP5, FP6, IP,
MM0, MM1, MM2, MM3, MM4, MM5, MM6, MM7, R10, R10B, R10D, R10W, R11, R11B, R11D,
R11W, R12, R12B, R12D, R12W, R13, R13B, R13D, R13W, R14, R14B, R14D, R14W, R15,
R15B, R15D, R15W, R8, R8B, R8D, R8W, R9, R9B, R9D, R9W, RAX, RBP, RBX, RCX, RDI,
RDX, RIP, RSI, RSP, SI, SIL, SP, SPL, ST0, ST1, ST2, ST3, ST4, ST5, ST6, ST7,
XMM0, XMM1, XMM10, XMM11, XMM12, XMM13, XMM14, XMM15, XMM2, XMM3, XMM4, XMM5,
XMM6, XMM7, XMM8, XMM9,

$ llvm-tblgen X86.td -print-enums -class=Instruction
```

```
ABS_F, ABS_Fp32, ABS_Fp64, ABS_Fp80, ADC32mi, ADC32mi8, ADC32mr, ADC32ri,
ADC32ri8, ADC32rm, ADC32rr, ADC64mi32, ADC64mi8, ADC64mr, ADC64ri32, ADC64ri8,
ADC64rm, ADC64rr, ADD16mi, ADD16mi8, ADD16mr, ADD16ri, ADD16ri8, ADD16rm,
ADD16rr, ADD32mi, ADD32mi8, ADD32mr, ADD32ri, ADD32ri8, ADD32rm, ADD32rr,
ADD64mi32, ADD64mi8, ADD64mr, ADD64ri32, ...
```

The default backend prints out all of the records.

If you plan to use TableGen, you will most likely have to write a *backend* that extracts the information specific to what you need and formats it in the appropriate way.

Example

With no other arguments, *llvm-tblgen* parses the specified file and prints out all of the classes, then all of the definitions. This is a good way to see what the various definitions expand to fully. Running this on the `X86.td` file prints this (at the time of this writing):

```
...
def ADD32rr {    // Instruction X86Inst I
  string Namespace = "X86";
  dag OutOperandList = (outs GR32:$dst);
  dag InOperandList = (ins GR32:$src1, GR32:$src2);
  string AsmString = "add{l}\t{$src2, $dst|$dst, $src2}";
  list<dag> Pattern = [(set GR32:$dst, (add GR32:$src1, GR32:$src2))];
  list<Register> Uses = [];
  list<Register> Defs = [EFLAGS];
  list<Predicate> Predicates = [];
  int CodeSize = 3;
  int AddedComplexity = 0;
  bit isReturn = 0;
  bit isBranch = 0;
  bit isIndirectBranch = 0;
  bit isBarrier = 0;
  bit isCall = 0;
  bit canFoldAsLoad = 0;
  bit mayLoad = 0;
  bit mayStore = 0;
  bit isImplicitDef = 0;
  bit isConvertibleToThreeAddress = 1;
  bit isCommutable = 1;
  bit isTerminator = 0;
  bit isReMaterializable = 0;
  bit isPredicable = 0;
  bit hasDelaySlot = 0;
  bit usesCustomInserter = 0;
  bit hasCtrlDep = 0;
  bit isNotDuplicable = 0;
  bit hasSideEffects = 0;
  InstrItinClass Itinerary = NoItinerary;
  string Constraints = "";
  string DisableEncoding = "";
  bits<8> Opcode = { 0, 0, 0, 0, 0, 0, 0, 1 };
  Format Form = MRMDestReg;
  bits<6> FormBits = { 0, 0, 0, 0, 1, 1 };
  ImmType ImmT = NoImm;
  bits<3> ImmTypeBits = { 0, 0, 0 };
  bit hasOpSizePrefix = 0;
```

```

bit hasAdSizePrefix = 0;
bits<4> Prefix = { 0, 0, 0, 0 };
bit hasREX_WPrefix = 0;
FPFormat FPForm = ?;
bits<3> FPFormBits = { 0, 0, 0 };
}
...

```

This definition corresponds to the 32-bit register-register add instruction of the x86 architecture. `def ADD32rr` defines a record named `ADD32rr`, and the comment at the end of the line indicates the superclasses of the definition. The body of the record contains all of the data that TableGen assembled for the record, indicating that the instruction is part of the “X86” namespace, the pattern indicating how the instruction is selected by the code generator, that it is a two-address instruction, has a particular encoding, etc. The contents and semantics of the information in the record are specific to the needs of the X86 backend, and are only shown as an example.

As you can see, a lot of information is needed for every instruction supported by the code generator, and specifying it all manually would be unmaintainable, prone to bugs, and tiring to do in the first place. Because we are using TableGen, all of the information was derived from the following definition:

```

let Defs = [EFLAGS],
    isCommutable = 1,                // X = ADD Y,Z --> X = ADD Z,Y
    isConvertibleToThreeAddress = 1 in // Can transform into LEA.
def ADD32rr : I<0x01, MRMDestReg, (outs GR32:$dst),
    (ins GR32:$src1, GR32:$src2),
    "add{l}\t{$src2, $dst|$dst, $src2}",
    [(set GR32:$dst, (add GR32:$src1, GR32:$src2))]>;

```

This definition makes use of the custom class `I` (extended from the custom class `X86Inst`), which is defined in the X86-specific TableGen file, to factor out the common features that instructions of its class share. A key feature of TableGen is that it allows the end-user to define the abstractions they prefer to use when describing their information.

Each `def` record has a special entry called “NAME”. This is the name of the record (“`ADD32rr`” above). In the general case `def` names can be formed from various kinds of string processing expressions and `NAME` resolves to the final value obtained after resolving all of those expressions. The user may refer to `NAME` anywhere she desires to use the ultimate name of the `def`. `NAME` should not be defined anywhere else in user code to avoid conflicts.

4.12.7 Syntax

TableGen has a syntax that is loosely based on C++ templates, with built-in types and specification. In addition, TableGen’s syntax introduces some automation concepts like multiclass, foreach, let, etc.

Basic concepts

TableGen files consist of two key parts: ‘classes’ and ‘definitions’, both of which are considered ‘records’.

TableGen records have a unique name, a list of values, and a list of superclasses. The list of values is the main data that TableGen builds for each record; it is this that holds the domain specific information for the application. The interpretation of this data is left to a specific *backend*, but the structure and format rules are taken care of and are fixed by TableGen.

TableGen definitions are the concrete form of ‘records’. These generally do not have any undefined values, and are marked with the ‘def’ keyword.

```

def FeatureFPARMv8 : SubtargetFeature<"fp-armv8", "HasFPARMv8", "true",
    "Enable ARMv8 FP">;

```

In this example, `FeatureFPARMv8` is `SubtargetFeature` record initialised with some values. The names of the classes are defined via the keyword *class* either on the same file or some other included. Most target TableGen files include the generic ones in `include/llvm/Target`.

TableGen classes are abstract records that are used to build and describe other records. These classes allow the end-user to build abstractions for either the domain they are targeting (such as “Register”, “RegisterClass”, and “Instruction” in the LLVM code generator) or for the implementor to help factor out common properties of records (such as “FPInst”, which is used to represent floating point instructions in the X86 backend). TableGen keeps track of all of the classes that are used to build up a definition, so the backend can find all definitions of a particular class, such as “Instruction”.

```
class ProcNoItin<string Name, list<SubtargetFeature> Features>
    : Processor<Name, NoItineraries, Features>;
```

Here, the class `ProcNoItin`, receiving parameters *Name* of type *string* and a list of target features is specializing the class `Processor` by passing the arguments down as well as hard-coding `NoItineraries`.

TableGen multiclasss are groups of abstract records that are instantiated all at once. Each instantiation can result in multiple TableGen definitions. If a multiclass inherits from another multiclass, the definitions in the sub-multiclass become part of the current multiclass, as if they were declared in the current multiclass.

```
multiclass ro_signed_pats<string T, string Rm, dag Base, dag Offset, dag Extend,
                        dag address, ValueType sty> {
def : Pat<(i32 (!cast<SDNode>("sextload" # sty) address)),
        (!cast<Instruction>("LDRS" # T # "w_" # Rm # "_RegOffset")
         Base, Offset, Extend)>;

def : Pat<(i64 (!cast<SDNode>("sextload" # sty) address)),
        (!cast<Instruction>("LDRS" # T # "x_" # Rm # "_RegOffset")
         Base, Offset, Extend)>;
}

defm : ro_signed_pats<"B", Rm, Base, Offset, Extend,
                    !foreach(decls.pattern, address,
                            !subst(SHIFT, imm_eq0, decls.pattern)),
                    i8>;
```

See the [TableGen Language Introduction](#) for more generic information on the usage of the language, and the [TableGen Language Reference](#) for more in-depth description of the formal language specification.

4.12.8 TableGen backends

TableGen files have no real meaning without a back-end. The default operation of running `llvm-tblgen` is to print the information in a textual format, but that’s only useful for debugging of the TableGen files themselves. The power in TableGen is, however, to interpret the source files into an internal representation that can be generated into anything you want.

Current usage of TableGen is to create huge include files with tables that you can either include directly (if the output is in the language you’re coding), or be used in pre-processing via macros surrounding the include of the file.

Direct output can be used if the back-end already prints a table in C format or if the output is just a list of strings (for error and warning messages). Pre-processed output should be used if the same information needs to be used in different contexts (like Instruction names), so your back-end should print a meta-information list that can be shaped into different compile-time formats.

See the [TableGen BackEnds](#) for more information.

4.12.9 TableGen Deficiencies

Despite being very generic, TableGen has some deficiencies that have been pointed out numerous times. The common theme is that, while TableGen allows you to build Domain-Specific-Languages, the final languages that you create lack the power of other DSLs, which in turn increase considerably the size and complexity of TableGen files.

At the same time, TableGen allows you to create virtually any meaning of the basic concepts via custom-made back-ends, which can pervert the original design and make it very hard for newcomers to understand the evil TableGen file.

There are some in favour of extending the semantics even more, but making sure back-ends adhere to strict rules. Others are suggesting we should move to less, more powerful DSLs designed with specific purposes, or even re-using existing DSLs.

Either way, this is a discussion that will likely span across several years, if not decades. You can read more in the [TableGen Deficiencies](#) document.

4.13 Debugging JIT-ed Code With GDB

4.13.1 Background

Without special runtime support, debugging dynamically generated code with GDB (as well as most debuggers) can be quite painful. Debuggers generally read debug information from the object file of the code, but for JITed code, there is no such file to look for.

In order to communicate the necessary debug info to GDB, an interface for registering JITed code with debuggers has been designed and implemented for GDB and LLVM MCJIT. At a high level, whenever MCJIT generates new machine code, it does so in an in-memory object file that contains the debug information in DWARF format. MCJIT then adds this in-memory object file to a global list of dynamically generated object files and calls a special function (`__jit_debug_register_code`) marked `noinline` that GDB knows about. When GDB attaches to a process, it puts a breakpoint in this function and loads all of the object files in the global list. When MCJIT calls the registration function, GDB catches the breakpoint signal, loads the new object file from the inferior's memory, and resumes the execution. In this way, GDB can get the necessary debug information.

4.13.2 GDB Version

In order to debug code JIT-ed by LLVM, you need GDB 7.0 or newer, which is available on most modern distributions of Linux. The version of GDB that Apple ships with Xcode has been frozen at 6.3 for a while. LLDB may be a better option for debugging JIT-ed code on Mac OS X.

4.13.3 Debugging MCJIT-ed code

The emerging MCJIT component of LLVM allows full debugging of JIT-ed code with GDB. This is due to MCJIT's ability to use the MC emitter to provide full DWARF debugging information to GDB.

Note that `lli` has to be passed the `-use-mcjit` flag to JIT the code with MCJIT instead of the old JIT.

Example

Consider the following C code (with line numbers added to make the example easier to follow):

```
1  int compute_factorial(int n)
2  {
3      if (n <= 1)
4          return 1;
5
6      int f = n;
7      while (--n > 1)
8          f *= n;
9      return f;
10 }
11
12
13 int main(int argc, char** argv)
14 {
15     if (argc < 2)
16         return -1;
17     char firstletter = argv[1][0];
18     int result = compute_factorial(firstletter - '0');
19
20     // Returned result is clipped at 255...
21     return result;
22 }
```

Here is a sample command line session that shows how to build and run this code via `lli` inside GDB:

```
$ $BINPATH/clang -cc1 -O0 -g -emit-llvm showdebug.c
$ gdb --quiet --args $BINPATH/lli -use-mcjit showdebug.ll 5
Reading symbols from $BINPATH/lli...done.
(gdb) b showdebug.c:6
No source file named showdebug.c.
Make breakpoint pending on future shared library load? (y or [n]) y
Breakpoint 1 (showdebug.c:6) pending.
(gdb) r
Starting program: $BINPATH/lli -use-mcjit showdebug.ll 5
[Thread debugging using libthread_db enabled]

Breakpoint 1, compute_factorial (n=5) at showdebug.c:6
6      int f = n;
(gdb) p n
$1 = 5
(gdb) p f
$2 = 0
(gdb) n
7      while (--n > 1)
(gdb) p f
$3 = 5
(gdb) b showdebug.c:9
Breakpoint 2 at 0x7ffff7ed404c: file showdebug.c, line 9.
(gdb) c
Continuing.

Breakpoint 2, compute_factorial (n=1) at showdebug.c:9
9      return f;
(gdb) p f
$4 = 120
(gdb) bt
#0  compute_factorial (n=1) at showdebug.c:9
#1  0x00007ffff7ed40a9 in main (argc=2, argv=0x16677e0) at showdebug.c:18
```

```

#2 0x3500000001652748 in ?? ()
#3 0x00000000016677e0 in ?? ()
#4 0x0000000000000002 in ?? ()
#5 0x0000000000d953b3 in llvm::MCJIT::runFunction (this=0x16151f0, F=0x1603020,
↳ArgValues=...) at /home/ebenders_test/llvm_svn_rw/lib/ExecutionEngine/MCJIT/MCJIT.
↳cpp:161
#6 0x0000000000dc8872 in llvm::ExecutionEngine::runFunctionAsMain (this=0x16151f0,
↳Fn=0x1603020, argv=..., envp=0x7fffffffe040)
    at /home/ebenders_test/llvm_svn_rw/lib/ExecutionEngine/ExecutionEngine.cpp:397
#7 0x000000000059c583 in main (argc=4, argv=0x7fffffffe018, envp=0x7fffffffe040) at /
↳home/ebenders_test/llvm_svn_rw/tools/lli/lli.cpp:324
(gdb) finish
Run till exit from #0 compute_factorial (n=1) at showdebug.c:9
0x00007ffff7ed40a9 in main (argc=2, argv=0x16677e0) at showdebug.c:18
18     int result = compute_factorial(firstletter - '0');
Value returned is $5 = 120
(gdb) p result
$6 = 23406408
(gdb) n
21     return result;
(gdb) p result
$7 = 120
(gdb) c
Continuing.

Program exited with code 0170.
(gdb)

```

4.14 The LLVM gold plugin

4.14.1 Introduction

Building with link time optimization requires cooperation from the system linker. LTO support on Linux systems requires that you use the [gold linker](#) or `ld.bfd` from `binutils` `>= 2.21.51.0.2`, as they support LTO via plugins. This is the same mechanism used by the [GCC LTO](#) project.

The LLVM gold plugin implements the gold plugin interface on top of *libLTO*. The same plugin can also be used by other tools such as `ar` and `nm`.

4.14.2 How to build it

Check for plugin support by running `/usr/bin/ld -plugin`. If it complains “missing argument” then you have plugin support. If not, such as an “unknown option” error then you will either need to build gold or install a recent version of `ld.bfd` with plugin support and then build gold plugin.

- Download, configure and build `ld.bfd` with plugin support:

```

$ git clone --depth 1 git://sourceware.org/git/binutils-gdb.git binutils
$ mkdir build
$ cd build
$ ../binutils/configure --disable-werror # ld.bfd includes plugin support by
↳default
$ make all-ld

```

That should leave you with `build/ld/ld-new` which supports the `-plugin` option. Running `make` will additionally build `build/binutils/ar` and `nm-new` binaries supporting plugins.

- Build the LLVMgold plugin. Run CMake with `-DLLVM_BINUTILS_INCDIR=/path/to/binutils/include`. The correct include path will contain the file `plugin-api.h`.

4.14.3 Usage

The linker takes a `-plugin` option that points to the path of the plugin `.so` file. To find out what link command `gcc` would run in a given situation, run `gcc -v [...]` and look for the line where it runs `collect2`. Replace that with `ld-new -plugin /path/to/LLVMgold.so` to test it out. Once you're ready to switch to using gold, backup your existing `/usr/bin/ld` then replace it with `ld-new`.

You should produce bitcode files from `clang` with the option `-fllvm-gold-plugin`. This flag will also cause `clang` to look for the gold plugin in the `lib` directory under its prefix and pass the `-plugin` option to `ld`. It will not look for an alternate linker, which is why you need gold to be the installed system linker in your path.

`ar` and `nm` also accept the `-plugin` option and it's possible to install `LLVMgold.so` to `/usr/lib/bfd-plugins` for a seamless setup. If you built your own gold, be sure to install the `ar` and `nm-new` you built to `/usr/bin`.

Example of link time optimization

The following example shows a worked example of the gold plugin mixing LLVM bitcode and native code.

```
--- a.c ---
#include <stdio.h>

extern void foo1(void);
extern void foo4(void);

void foo2(void) {
    printf("Foo2\n");
}

void foo3(void) {
    foo4();
}

int main(void) {
    foo1();
}

--- b.c ---
#include <stdio.h>

extern void foo2(void);

void foo1(void) {
    foo2();
}

void foo4(void) {
    printf("Foo4");
}
```

```

--- command lines ---
$ clang -flto a.c -c -o a.o      # <-- a.o is LLVM bitcode file
$ ar q a.a a.o                  # <-- a.a is an archive with LLVM bitcode
$ clang b.c -c -o b.o           # <-- b.o is native object file
$ clang -flto a.a b.o -o main   # <-- link with LLVMgold plugin

```

Gold informs the plugin that `foo3` is never referenced outside the IR, leading LLVM to delete that function. However, unlike in the *libLTO example* gold does not currently eliminate `foo4`.

4.14.4 Quickstart for using LTO with autotooled projects

Once your system `ld`, `ar`, and `nm` all support LLVM bitcode, everything is in place for an easy to use LTO build of autotooled projects:

- Follow the instructions *on how to build LLVMgold.so*.
- Install the newly built binutils to `$PREFIX`
- Copy `Release/lib/LLVMgold.so` to `$PREFIX/lib/bfd-plugins/`
- Set environment variables (`$PREFIX` is where you installed clang and binutils):

```

export CC="$PREFIX/bin/clang -flto"
export CXX="$PREFIX/bin/clang++ -flto"
export AR="$PREFIX/bin/ar"
export NM="$PREFIX/bin/nm"
export RANLIB=/bin/true #ranlib is not needed, and doesn't support .bc files in .a

```

- Or you can just set your path:

```

export PATH="$PREFIX/bin:$PATH"
export CC="clang -flto"
export CXX="clang++ -flto"
export RANLIB=/bin/true

```

- Configure and build the project as usual:

```
% ./configure && make && make check
```

The environment variable settings may work for non-autotooled projects too, but you may need to set the `LD` environment variable as well.

4.14.5 Licensing

Gold is licensed under the GPLv3. LLVMgold uses the interface file `plugin-api.h` from gold which means that the resulting `LLVMgold.so` binary is also GPLv3. This can still be used to link non-GPLv3 programs just as much as gold could without the plugin.

4.15 LLVM's Optional Rich Disassembly Output

- *Introduction*

- *Instruction Annotations*
 - *Contextual markups*
 - *C API Details*

4.15.1 Introduction

LLVM’s default disassembly output is raw text. To allow consumers more ability to introspect the instructions’ textual representation or to reformat for a more user friendly display there is an optional rich disassembly output.

This optional output is sufficient to reference into individual portions of the instruction text. This is intended for clients like disassemblers, list file generators, and pretty-printers, which need more than the raw instructions and the ability to print them.

To provide this functionality the assembly text is marked up with annotations. The markup is simple enough in syntax to be robust even in the case of version mismatches between consumers and producers. That is, the syntax generally does not carry semantics beyond “this text has an annotation,” so consumers can simply ignore annotations they do not understand or do not care about.

After calling `LLVMCreateDisasm()` to create a disassembler context the optional output is enable with this call:

```
LLVMSetDisasmOptions(DC, LLVMDisassembler_Option_UseMarkup);
```

Then subsequent calls to `LLVMDisasmInstruction()` will return output strings with the marked up annotations.

4.15.2 Instruction Annotations

Contextual markups

Annotated assembly display will supply contextual markup to help clients more efficiently implement things like pretty printers. Most markup will be target independent, so clients can effectively provide good display without any target specific knowledge.

Annotated assembly goes through the normal instruction printer, but optionally includes contextual tags on portions of the instruction string. An annotation is any ‘<’ ‘>’ delimited section of text(1).

```
annotation: '<' tag-name tag-modifier-list ':' annotated-text '>'
tag-name: identifier
tag-modifier-list: comma delimited identifier list
```

The tag-name is an identifier which gives the type of the annotation. For the first pass, this will be very simple, with memory references, registers, and immediates having the tag names “mem”, “reg”, and “imm”, respectively.

The tag-modifier-list is typically additional target-specific context, such as register class.

Clients should accept and ignore any tag-names or tag-modifiers they do not understand, allowing the annotations to grow in richness without breaking older clients.

For example, a possible annotation of an ARM load of a stack-relative location might be annotated as:

```
ldr <reg gpr:r0>, <mem regoffset:[<reg gpr:sp>, <imm:#4>]>
```

1: For assembly dialects in which ‘<’ and/or ‘>’ are legal tokens, a literal token is escaped by following immediately with a repeat of the character. For example, a literal ‘<’ character is output as ‘<<’ in an annotated assembly string.

C API Details

The intended consumers of this information use the C API, therefore the new C API function for the disassembler will be added to provide an option to produce disassembled instructions with annotations, `LLVMSetDisasmOptions()` and the `LLVMDisassembler_Option_UseMarkup` option (see above).

4.16 System Library

4.16.1 Abstract

This document provides some details on LLVM's System Library, located in the source at `lib/System` and `include/llvm/System`. The library's purpose is to shield LLVM from the differences between operating systems for the few services LLVM needs from the operating system. Much of LLVM is written using portability features of standard C++. However, in a few areas, system dependent facilities are needed and the System Library is the wrapper around those system calls.

By centralizing LLVM's use of operating system interfaces, we make it possible for the LLVM tool chain and runtime libraries to be more easily ported to new platforms since (theoretically) only `lib/System` needs to be ported. This library also unclutters the rest of LLVM from `#ifdef` use and special cases for specific operating systems. Such uses are replaced with simple calls to the interfaces provided in `include/llvm/System`.

Note that the System Library is not intended to be a complete operating system wrapper (such as the Adaptive Communications Environment (ACE) or Apache Portable Runtime (APR)), but only provides the functionality necessary to support LLVM.

The System Library was written by Reid Spencer who formulated the design based on similar work originating from the eXtensible Programming System (XPS). Several people helped with the effort; especially, Jeff Cohen and Henrik Bach on the Win32 port.

4.16.2 Keeping LLVM Portable

In order to keep LLVM portable, LLVM developers should adhere to a set of portability rules associated with the System Library. Adherence to these rules should help the System Library achieve its goal of shielding LLVM from the variations in operating system interfaces and doing so efficiently. The following sections define the rules needed to fulfill this objective.

Don't Include System Headers

Except in `lib/System`, no LLVM source code should directly `#include` a system header. Care has been taken to remove all such `#includes` from LLVM while `lib/System` was being developed. Specifically this means that header files like `"unistd.h"`, `"windows.h"`, `"stdio.h"`, and `"string.h"` are forbidden to be included by LLVM source code outside the implementation of `lib/System`.

To obtain system-dependent functionality, existing interfaces to the system found in `include/llvm/System` should be used. If an appropriate interface is not available, it should be added to `include/llvm/System` and implemented in `lib/System` for all supported platforms.

Don't Expose System Headers

The System Library must shield LLVM from **all** system headers. To obtain system level functionality, LLVM source must `#include "llvm/System/Thing.h"` and nothing else. This means that `Thing.h` cannot expose any

system header files. This protects LLVM from accidentally using system specific functionality and only allows it via the `lib/System` interface.

Use Standard C Headers

The **standard** C headers (the ones beginning with “c”) are allowed to be exposed through the `lib/System` interface. These headers and the things they declare are considered to be platform agnostic. LLVM source files may include them directly or obtain their inclusion through `lib/System` interfaces.

Use Standard C++ Headers

The **standard** C++ headers from the standard C++ library and standard template library may be exposed through the `lib/System` interface. These headers and the things they declare are considered to be platform agnostic. LLVM source files may include them or obtain their inclusion through `lib/System` interfaces.

High Level Interface

The entry points specified in the interface of `lib/System` must be aimed at completing some reasonably high level task needed by LLVM. We do not want to simply wrap each operating system call. It would be preferable to wrap several operating system calls that are always used in conjunction with one another by LLVM.

For example, consider what is needed to execute a program, wait for it to complete, and return its result code. On Unix, this involves the following operating system calls: `getenv`, `fork`, `execve`, and `wait`. The correct thing for `lib/System` to provide is a function, say `ExecuteProgramAndWait`, that implements the functionality completely. what we don't want is wrappers for the operating system calls involved.

There must **not** be a one-to-one relationship between operating system calls and the System library's interface. Any such interface function will be suspicious.

No Unused Functionality

There must be no functionality specified in the interface of `lib/System` that isn't actually used by LLVM. We're not writing a general purpose operating system wrapper here, just enough to satisfy LLVM's needs. And, LLVM doesn't need much. This design goal aims to keep the `lib/System` interface small and understandable which should foster its actual use and adoption.

No Duplicate Implementations

The implementation of a function for a given platform must be written exactly once. This implies that it must be possible to apply a function's implementation to multiple operating systems if those operating systems can share the same implementation. This rule applies to the set of operating systems supported for a given class of operating system (e.g. Unix, Win32).

No Virtual Methods

The System Library interfaces can be called quite frequently by LLVM. In order to make those calls as efficient as possible, we discourage the use of virtual methods. There is no need to use inheritance for implementation differences, it just adds complexity. The `#include` mechanism works just fine.

No Exposed Functions

Any functions defined by system libraries (i.e. not defined by `lib/System`) must not be exposed through the `lib/System` interface, even if the header file for that function is not exposed. This prevents inadvertent use of system specific functionality.

For example, the `stat` system call is notorious for having variations in the data it provides. `lib/System` must not declare `stat` nor allow it to be declared. Instead it should provide its own interface to discovering information about files and directories. Those interfaces may be implemented in terms of `stat` but that is strictly an implementation detail. The interface provided by the System Library must be implemented on all platforms (even those without `stat`).

No Exposed Data

Any data defined by system libraries (i.e. not defined by `lib/System`) must not be exposed through the `lib/System` interface, even if the header file for that function is not exposed. As with functions, this prevents inadvertent use of data that might not exist on all platforms.

Minimize Soft Errors

Operating system interfaces will generally provide error results for every little thing that could go wrong. In almost all cases, you can divide these error results into two groups: normal/good/soft and abnormal/bad/hard. That is, some of the errors are simply information like “file not found”, “insufficient privileges”, etc. while other errors are much harder like “out of space”, “bad disk sector”, or “system call interrupted”. We’ll call the first group “*soft*” errors and the second group “*hard*” errors.

`lib/System` must always attempt to minimize soft errors. This is a design requirement because the minimization of soft errors can affect the granularity and the nature of the interface. In general, if you find that you’re wanting to throw soft errors, you must review the granularity of the interface because it is likely you’re trying to implement something that is too low level. The rule of thumb is to provide interface functions that **can’t** fail, except when faced with hard errors.

For a trivial example, suppose we wanted to add an “`OpenFileForWriting`” function. For many operating systems, if the file doesn’t exist, attempting to open the file will produce an error. However, `lib/System` should not simply throw that error if it occurs because its a soft error. The problem is that the interface function, `OpenFileForWriting` is too low level. It should be `OpenOrCreateFileForWriting`. In the case of the soft “doesn’t exist” error, this function would just create it and then open it for writing.

This design principle needs to be maintained in `lib/System` because it avoids the propagation of soft error handling throughout the rest of LLVM. Hard errors will generally just cause a termination for an LLVM tool so don’t be bashful about throwing them.

Rules of thumb:

1. Don’t throw soft errors, only hard errors.
2. If you’re tempted to throw a soft error, re-think the interface.
3. Handle internally the most common normal/good/soft error conditions so the rest of LLVM doesn’t have to.

No throw Specifications

None of the `lib/System` interface functions may be declared with C++ `throw()` specifications on them. This requirement makes sure that the compiler does not insert additional exception handling code into the interface functions. This is a performance consideration: `lib/System` functions are at the bottom of many call chains and as such can be frequently called. We need them to be as efficient as possible. However, no routines in the system library should actually throw exceptions.

Code Organization

Implementations of the System Library interface are separated by their general class of operating system. Currently only Unix and Win32 classes are defined but more could be added for other operating system classifications. To distinguish which implementation to compile, the code in `lib/System` uses the `LLVM_ON_UNIX` and `LLVM_ON_WIN32` `#defines` provided via configure through the `llvm/Config/config.h` file. Each source file in `lib/System`, after implementing the generic (operating system independent) functionality needs to include the correct implementation using a set of `#if defined(LLVM_ON_XYZ)` directives. For example, if we had `lib/System/File.cpp`, we'd expect to see in that file:

```
#if defined(LLVM_ON_UNIX)
#include "Unix/File.cpp"
#endif
#if defined(LLVM_ON_WIN32)
#include "Win32/File.cpp"
#endif
```

The implementation in `lib/System/Unix/File.cpp` should handle all Unix variants. The implementation in `lib/System/Win32/File.cpp` should handle all Win32 variants. What this does is quickly differentiate the basic class of operating system that will provide the implementation. The specific details for a given platform must still be determined through the use of `#ifdef`.

Consistent Semantics

The implementation of a `lib/System` interface can vary drastically between platforms. That's okay as long as the end result of the interface function is the same. For example, a function to create a directory is pretty straight forward on all operating system. System V IPC on the other hand isn't even supported on all platforms. Instead of "supporting" System V IPC, `lib/System` should provide an interface to the basic concept of inter-process communications. The implementations might use System V IPC if that was available or named pipes, or whatever gets the job done effectively for a given operating system. In all cases, the interface and the implementation must be semantically consistent.

4.17 Source Level Debugging with LLVM

- *Introduction*
 - *Philosophy behind LLVM debugging information*
 - *Debug information consumers*
 - *Debugging optimized code*
- *Debugging information format*
 - *Debugger intrinsic functions*
 - * `llvm.dbg.declare`
 - * `llvm.dbg.value`
- *Object lifetimes and scoping*
- *C/C++ front-end specific debug information*
 - *C/C++ source file information*

- C/C++ global variable information
- C/C++ function information
- *Debugging information format*
 - *Debugging Information Extension for Objective C Properties*
 - * *Introduction*
 - * *Proposal*
 - * *New DWARF Tags*
 - * *New DWARF Attributes*
 - * *New DWARF Constants*
 - *Name Accelerator Tables*
 - * *Introduction*
 - * *Hash Tables*
 - *Standard Hash Tables*
 - *Name Hash Tables*
 - * *Details*
 - *Header Layout*
 - *Fixed Lookup*
 - * *Contents*
 - * *Language Extensions and File Format Changes*
 - *Objective-C Extensions*
 - *Mach-O Changes*
- *CodeView Debug Info Format*
 - *Format Background*
 - *Working with CodeView*

4.17.1 Introduction

This document is the central repository for all information pertaining to debug information in LLVM. It describes the *actual format that the LLVM debug information takes*, which is useful for those interested in creating front-ends or dealing directly with the information. Further, this document provides specific examples of what debug information for C/C++ looks like.

Philosophy behind LLVM debugging information

The idea of the LLVM debugging information is to capture how the important pieces of the source-language's Abstract Syntax Tree map onto LLVM code. Several design aspects have shaped the solution that appears here. The important ones are:

- Debugging information should have very little impact on the rest of the compiler. No transformations, analyses, or code generators should need to be modified because of debugging information.

- LLVM optimizations should interact in *well-defined and easily described ways* with the debugging information.
- Because LLVM is designed to support arbitrary programming languages, LLVM-to-LLVM tools should not need to know anything about the semantics of the source-level-language.
- Source-level languages are often **widely** different from one another. LLVM should not put any restrictions of the flavor of the source-language, and the debugging information should work with any language.
- With code generator support, it should be possible to use an LLVM compiler to compile a program to native machine code and standard debugging formats. This allows compatibility with traditional machine-code level debuggers, like GDB or DBX.

The approach used by the LLVM implementation is to use a small set of *intrinsic functions* to define a mapping between LLVM program objects and the source-level objects. The description of the source-level program is maintained in LLVM metadata in an *implementation-defined format* (the C/C++ front-end currently uses working draft 7 of the [DWARF 3 standard](#)).

When a program is being debugged, a debugger interacts with the user and turns the stored debug information into source-language specific information. As such, a debugger must be aware of the source-language, and is thus tied to a specific language or family of languages.

Debug information consumers

The role of debug information is to provide meta information normally stripped away during the compilation process. This meta information provides an LLVM user a relationship between generated code and the original program source code.

Currently, there are two backend consumers of debug info: DwarfDebug and CodeViewDebug. DwarfDebug produces DWARF suitable for use with GDB, LLDB, and other DWARF-based debuggers. *CodeViewDebug* produces CodeView, the Microsoft debug info format, which is usable with Microsoft debuggers such as Visual Studio and WinDBG. LLVM's debug information format is mostly derived from and inspired by DWARF, but it is feasible to translate into other target debug info formats such as STABS.

It would also be reasonable to use debug information to feed profiling tools for analysis of generated code, or, tools for reconstructing the original source from generated code.

Debugging optimized code

An extremely high priority of LLVM debugging information is to make it interact well with optimizations and analysis. In particular, the LLVM debug information provides the following guarantees:

- LLVM debug information **always provides information to accurately read the source-level state of the program**, regardless of which LLVM optimizations have been run, and without any modification to the optimizations themselves. However, some optimizations may impact the ability to modify the current state of the program with a debugger, such as setting program variables, or calling functions that have been deleted.
- As desired, LLVM optimizations can be upgraded to be aware of debugging information, allowing them to update the debugging information as they perform aggressive optimizations. This means that, with effort, the LLVM optimizers could optimize debug code just as well as non-debug code.
- LLVM debug information does not prevent optimizations from happening (for example inlining, basic block reordering/merging/cleanup, tail duplication, etc).
- LLVM debug information is automatically optimized along with the rest of the program, using existing facilities. For example, duplicate information is automatically merged by the linker, and unused information is automatically removed.

Basically, the debug information allows you to compile a program with “`-O0 -g`” and get full debug information, allowing you to arbitrarily modify the program as it executes from a debugger. Compiling a program with “`-O3 -g`” gives you full debug information that is always available and accurate for reading (e.g., you get accurate stack traces despite tail call elimination and inlining), but you might lose the ability to modify the program and call functions which were optimized out of the program, or inlined away completely.

The *LLVM test suite* provides a framework to test optimizer’s handling of debugging information. It can be run like this:

```
% cd llvm/projects/test-suite/MultiSource/Benchmarks # or some other level
% make TEST=dbgopt
```

This will test impact of debugging information on optimization passes. If debugging information influences optimization passes then it will be reported as a failure. See *LLVM Testing Infrastructure Guide* for more information on LLVM test infrastructure and how to run various tests.

4.17.2 Debugging information format

LLVM debugging information has been carefully designed to make it possible for the optimizer to optimize the program and debugging information without necessarily having to know anything about debugging information. In particular, the use of metadata avoids duplicated debugging information from the beginning, and the global dead code elimination pass automatically deletes debugging information for a function if it decides to delete the function.

To do this, most of the debugging information (descriptors for types, variables, functions, source files, etc) is inserted by the language front-end in the form of LLVM metadata.

Debug information is designed to be agnostic about the target debugger and debugging information representation (e.g. DWARF/Stabs/etc). It uses a generic pass to decode the information that represents variables, types, functions, namespaces, etc: this allows for arbitrary source-language semantics and type-systems to be used, as long as there is a module written for the target debugger to interpret the information.

To provide basic functionality, the LLVM debugger does have to make some assumptions about the source-level language being debugged, though it keeps these to a minimum. The only common features that the LLVM debugger assumes exist are *source files*, and *program objects*. These abstract objects are used by a debugger to form stack traces, show information about local variables, etc.

This section of the documentation first describes the representation aspects common to any source-language. *C/C++ front-end specific debug information* describes the data layout conventions used by the C and C++ front-ends.

Debug information descriptors are *specialized metadata nodes*, first-class subclasses of `Metadata`.

Debugger intrinsic functions

LLVM uses several intrinsic functions (name prefixed with “`llvm.dbg`”) to provide debug information at various points in generated code.

`llvm.dbg.declare`

```
void @llvm.dbg.declare(metadata, metadata, metadata)
```

This intrinsic provides information about a local element (e.g., variable). The first argument is metadata holding the alloca for the variable. The second argument is a *local variable* containing a description of the variable. The third argument is a *complex expression*. An *llvm.dbg.declare* intrinsic describes the *location* of a source variable.

```
%i.addr = alloca i32, align 4
call void @llvm.dbg.declare(metadata i32* %i.addr, metadata !1, metadata !2), !dbg !3
!1 = !DILocalVariable(name: "i", ...) ; int i
!2 = !DIExpression()
!3 = !DILocation(...)
...
%buffer = alloca [256 x i8], align 8
; The address of i is buffer+64.
call void @llvm.dbg.declare(metadata [256 x i8]* %buffer, metadata !1, metadata !2)
!1 = !DILocalVariable(name: "i", ...) ; int i
!2 = !DIExpression(DW_OP_plus, 64)
```

llvm.dbg.value

```
void @llvm.dbg.value(metadata, i64, metadata, metadata)
```

This intrinsic provides information when a user source variable is set to a new value. The first argument is the new value (wrapped as metadata). The second argument is the offset in the user source variable where the new value is written. The third argument is a [local variable](#) containing a description of the variable. The fourth argument is a [complex expression](#).

4.17.3 Object lifetimes and scoping

In many languages, the local variables in functions can have their lifetimes or scopes limited to a subset of a function. In the C family of languages, for example, variables are only live (readable and writable) within the source block that they are defined in. In functional languages, values are only readable after they have been defined. Though this is a very obvious concept, it is non-trivial to model in LLVM, because it has no notion of scoping in this sense, and does not want to be tied to a language's scoping rules.

In order to handle this, the LLVM debug format uses the metadata attached to llvm instructions to encode line number and scoping information. Consider the following C fragment, for example:

```
1. void foo() {
2.     int X = 21;
3.     int Y = 22;
4.     {
5.         int Z = 23;
6.         Z = X;
7.     }
8.     X = Y;
9. }
```

Compiled to LLVM, this function would be represented like this:

```
; Function Attrs: nounwind ssp uwtable
define void @foo() #0 !dbg !4 {
entry:
    %X = alloca i32, align 4
    %Y = alloca i32, align 4
    %Z = alloca i32, align 4
    call void @llvm.dbg.declare(metadata i32* %X, metadata !11, metadata !13), !dbg !14
    store i32 21, i32* %X, align 4, !dbg !14
    call void @llvm.dbg.declare(metadata i32* %Y, metadata !15, metadata !13), !dbg !16
    store i32 22, i32* %Y, align 4, !dbg !16
```

```

    call void @llvm.dbg.declare(metadata i32* %Z, metadata !17, metadata !13), !dbg !19
    store i32 23, i32* %Z, align 4, !dbg !19
    %0 = load i32, i32* %X, align 4, !dbg !20
    store i32 %0, i32* %Z, align 4, !dbg !21
    %1 = load i32, i32* %Y, align 4, !dbg !22
    store i32 %1, i32* %X, align 4, !dbg !23
    ret void, !dbg !24
}

; Function Attrs: nounwind readnone
declare void @llvm.dbg.declare(metadata, metadata, metadata) #1

attributes #0 = { nounwind ssp uwtable "less-precise-fpmad"="false" "no-frame-pointer-
→elim"="true" "no-frame-pointer-elim-non-leaf" "no-infs-fp-math"="false" "no-nans-fp-
→math"="false" "stack-protector-buffer-size"="8" "unsafe-fp-math"="false" "use-soft-
→float"="false" }
attributes #1 = { nounwind readnone }

!llvm.dbg.cu = !{!0}
!llvm.module.flags = !{!7, !8, !9}
!llvm.ident = !{!10}

!0 = !DICompileUnit(language: DW_LANG_C99, file: !1, producer: "clang version 3.7.0_
→(trunk 231150) (llvm/trunk 231154)", isOptimized: false, runtimeVersion: 0, _
→emissionKind: FullDebug, enums: !2, retainedTypes: !2, subprograms: !3, globals: !2,
→ imports: !2)
!1 = !DIFile(filename: "/dev/stdin", directory: "/Users/dexonsmith/data/llvm/debug-
→info")
!2 = !{}
!3 = !{!4}
!4 = distinct !DISubprogram(name: "foo", scope: !1, file: !1, line: 1, type: !5, _
→isLocal: false, isDefinition: true, scopeLine: 1, isOptimized: false, variables: !2)
!5 = !DISubroutineType(types: !6)
!6 = !{null}
!7 = !{i32 2, !"Dwarf Version", i32 2}
!8 = !{i32 2, !"Debug Info Version", i32 3}
!9 = !{i32 1, !"PIC Level", i32 2}
!10 = !{"clang version 3.7.0 (trunk 231150) (llvm/trunk 231154)"}
!11 = !DILocalVariable(name: "X", scope: !4, file: !1, line: 2, type: !12)
!12 = !DIBasicType(name: "int", size: 32, align: 32, encoding: DW_ATE_signed)
!13 = !DIExpression()
!14 = !DILocation(line: 2, column: 9, scope: !4)
!15 = !DILocalVariable(name: "Y", scope: !4, file: !1, line: 3, type: !12)
!16 = !DILocation(line: 3, column: 9, scope: !4)
!17 = !DILocalVariable(name: "Z", scope: !18, file: !1, line: 5, type: !12)
!18 = distinct !DILexicalBlock(scope: !4, file: !1, line: 4, column: 5)
!19 = !DILocation(line: 5, column: 11, scope: !18)
!20 = !DILocation(line: 6, column: 11, scope: !18)
!21 = !DILocation(line: 6, column: 9, scope: !18)
!22 = !DILocation(line: 8, column: 9, scope: !4)
!23 = !DILocation(line: 8, column: 7, scope: !4)
!24 = !DILocation(line: 9, column: 3, scope: !4)

```

This example illustrates a few important details about LLVM debugging information. In particular, it shows how the `llvm.dbg.declare` intrinsic and location information, which are attached to an instruction, are applied together to allow a debugger to analyze the relationship between statements, variable definitions, and the code used to implement the function.

```
call void @llvm.dbg.declare(metadata i32* %X, metadata !11, metadata !13), !dbg !14
; [debug line = 2:7] [debug variable = X]
```

The first intrinsic `@llvm.dbg.declare` encodes debugging information for the variable `X`. The metadata `!dbg !14` attached to the intrinsic provides scope information for the variable `X`.

```
!14 = !DILocation(line: 2, column: 9, scope: !4)
!4 = distinct !DISubprogram(name: "foo", scope: !1, file: !1, line: 1, type: !5,
                           isLocal: false, isDefinition: true, scopeLine: 1,
                           isOptimized: false, variables: !2)
```

Here `!14` is metadata providing [location information](#). In this example, scope is encoded by `!4`, a [subprogram descriptor](#). This way the location information attached to the intrinsics indicates that the variable `X` is declared at line number 2 at a function level scope in function `foo`.

Now let's take another example.

```
call void @llvm.dbg.declare(metadata i32* %Z, metadata !17, metadata !13), !dbg !19
; [debug line = 5:9] [debug variable = Z]
```

The third intrinsic `@llvm.dbg.declare` encodes debugging information for variable `Z`. The metadata `!dbg !19` attached to the intrinsic provides scope information for the variable `Z`.

```
!18 = distinct !DILexicalBlock(scope: !4, file: !1, line: 4, column: 5)
!19 = !DILocation(line: 5, column: 11, scope: !18)
```

Here `!19` indicates that `Z` is declared at line number 5 and column number 0 inside of lexical scope `!18`. The lexical scope itself resides inside of subprogram `!4` described above.

The scope information attached with each instruction provides a straightforward way to find instructions covered by a scope.

4.17.4 C/C++ front-end specific debug information

The C and C++ front-ends represent information about the program in a format that is effectively identical to [DWARF 3.0](#) in terms of information content. This allows code generators to trivially support native debuggers by generating standard dwarf information, and contains enough information for non-dwarf targets to translate it as needed.

This section describes the forms used to represent C and C++ programs. Other languages could pattern themselves after this (which itself is tuned to representing programs in the same way that DWARF 3 does), or they could choose to provide completely different forms if they don't fit into the DWARF model. As support for debugging information gets added to the various LLVM source-language front-ends, the information used should be documented here.

The following sections provide examples of a few C/C++ constructs and the debug information that would best describe those constructs. The canonical references are the `DIDescriptor` classes defined in `include/llvm/IR/DebugInfo.h` and the implementations of the helper functions in `lib/IR/DIBuilder.cpp`.

C/C++ source file information

`llvm::Instruction` provides easy access to metadata attached with an instruction. One can extract line number information encoded in LLVM IR using `Instruction::getDebugLoc()` and `DILocation::getLine()`.

```
if (DILocation *Loc = I->getDebugLoc()) { // Here I is an LLVM instruction
    unsigned Line = Loc->getLine();
   StringRef File = Loc->getFilename();
```



```
StringRef Dir = Loc->getDirectory();
}
```

C/C++ global variable information

Given an integer global variable declared as follows:

```
_Alignas(8) int MyGlobal = 100;
```

a C/C++ front-end would generate the following descriptors:

```
;;
;; Define the global itself.
;;
@MyGlobal = global i32 100, align 8, !dbg !0

;;
;; List of debug info of globals
;;
!llvm.dbg.cu = !{!1}

;; Some unrelated metadata.
!llvm.module.flags = !{!6, !7}
!llvm.ident = !{!8}

;; Define the global variable itself
!0 = distinct !DIGlobalVariable(name: "MyGlobal", scope: !1, file: !2, line: 1, type:
↳!5, isLocal: false, isDefinition: true, align: 64)

;; Define the compile unit.
!1 = distinct !DICompileUnit(language: DW_LANG_C99, file: !2,
    producer: "clang version 4.0.0 (http://llvm.org/git/
↳clang.git ae4deadbea242e8ea517eef662c30443f75bd086) (http://llvm.org/git/llvm.git
↳818b4c1539df3e51dc7e62c89ead4abfd348827d)",
    isOptimized: false, runtimeVersion: 0, emissionKind:
↳FullDebug,
    enums: !3, globals: !4)

;;
;; Define the file
;;
!2 = !DIFile(filename: "/dev/stdin",
    directory: "/Users/dexonsmith/data/llvm/debug-info")

;; An empty array.
!3 = !{}

;; The Array of Global Variables
!4 = !{!0}

;;
;; Define the type
;;
!5 = !DIBasicType(name: "int", size: 32, encoding: DW_ATE_signed)

;; Dwarf version to output.
```

```
!6 = !{i32 2, !"Dwarf Version", i32 4}

;; Debug info schema version.
!7 = !{i32 2, !"Debug Info Version", i32 3}

;; Compiler identification
!8 = !{"clang version 4.0.0 (http://llvm.org/git/clang.git_
↪ae4deadbea242e8ea517eef662c30443f75bd086) (http://llvm.org/git/_
↪818b4c1539df3e51dc7e62c89ead4abfd348827d)"} }
```

The align value in DIGlobalVariable description specifies variable alignment in case it was forced by C11 `_Alignas()`, C++11 `alignas()` keywords or compiler attribute `__attribute__((aligned ()))`. In other case (when this field is missing) alignment is considered default. This is used when producing DWARF output for DW_AT_alignment value.

C/C++ function information

Given a function declared as follows:

```
int main(int argc, char *argv[]) {
    return 0;
}
```

a C/C++ front-end would generate the following descriptors:

```
;;
;; Define the anchor for subprograms.
;;
!4 = !DISubprogram(name: "main", scope: !1, file: !1, line: 1, type: !5,
                  isLocal: false, isDefinition: true, scopeLine: 1,
                  flags: DIFlagPrototyped, isOptimized: false,
                  variables: !2)

;;
;; Define the subprogram itself.
;;
define i32 @main(i32 %argc, i8** %argv) !dbg !4 {
...
}
```

4.17.5 Debugging information format

Debugging Information Extension for Objective C Properties

Introduction

Objective C provides a simpler way to declare and define accessor methods using declared properties. The language provides features to declare a property and to let compiler synthesize accessor methods.

The debugger lets developer inspect Objective C interfaces and their instance variables and class variables. However, the debugger does not know anything about the properties defined in Objective C interfaces. The debugger consumes information generated by compiler in DWARF format. The format does not support encoding of Objective C properties. This proposal describes DWARF extensions to encode Objective C properties, which the debugger can use to let developers inspect Objective C properties.

Proposal

Objective C properties exist separately from class members. A property can be defined only by “setter” and “getter” selectors, and be calculated anew on each access. Or a property can just be a direct access to some declared ivar. Finally it can have an ivar “automatically synthesized” for it by the compiler, in which case the property can be referred to in user code directly using the standard C dereference syntax as well as through the property “dot” syntax, but there is no entry in the `@interface` declaration corresponding to this ivar.

To facilitate debugging, these properties we will add a new DWARF TAG into the `DW_TAG_structure_type` definition for the class to hold the description of a given property, and a set of DWARF attributes that provide said description. The property tag will also contain the name and declared type of the property.

If there is a related ivar, there will also be a DWARF property attribute placed in the `DW_TAG_member` DIE for that ivar referring back to the property TAG for that property. And in the case where the compiler synthesizes the ivar directly, the compiler is expected to generate a `DW_TAG_member` for that ivar (with the `DW_AT_artificial` set to 1), whose name will be the name used to access this ivar directly in code, and with the property attribute pointing back to the property it is backing.

The following examples will serve as illustration for our discussion:

```
@interface I1 {
    int n2;
}

@property int p1;
@property int p2;
@end

@implementation I1
@synthesize p1;
@synthesize p2 = n2;
@end
```

This produces the following DWARF (this is a “pseudo dwarfdump” output):

```
0x00000100: TAG_structure_type [7] *
            AT_APPLE_runtime_class( 0x10 )
            AT_name( "I1" )
            AT_decl_file( "Objc_Property.m" )
            AT_decl_line( 3 )

0x00000110: TAG_APPLE_property
            AT_name ( "p1" )
            AT_type ( {0x00000150} ( int ) )

0x00000120: TAG_APPLE_property
            AT_name ( "p2" )
            AT_type ( {0x00000150} ( int ) )

0x00000130: TAG_member [8]
            AT_name( "_p1" )
            AT_APPLE_property ( {0x00000110} "p1" )
            AT_type( {0x00000150} ( int ) )
            AT_artificial ( 0x1 )

0x00000140: TAG_member [8]
            AT_name( "n2" )
            AT_APPLE_property ( {0x00000120} "p2" )
```

```

        AT_type( {0x00000150} ( int ) )

0x00000150:  AT_type( ( int ) )

```

Note, the current convention is that the name of the ivar for an auto-synthesized property is the name of the property from which it derives with an underscore prepended, as is shown in the example. But we actually don't need to know this convention, since we are given the name of the ivar directly.

Also, it is common practice in ObjC to have different property declarations in the @interface and @implementation - e.g. to provide a read-only property in the interface, and a read-write interface in the implementation. In that case, the compiler should emit whichever property declaration will be in force in the current translation unit.

Developers can decorate a property with attributes which are encoded using DW_AT_APPLE_property_attribute.

```
@property (readonly, nonatomic) int pr;
```

```

TAG_APPLE_property [8]
  AT_name( "pr" )
  AT_type ( {0x00000147} (int) )
  AT_APPLE_property_attribute (DW_APPLE_PROPERTY_readonly, DW_APPLE_PROPERTY_
↪nonatomic)

```

The setter and getter method names are attached to the property using DW_AT_APPLE_property_setter and DW_AT_APPLE_property_getter attributes.

```

@interface I1
@property (setter=myOwnP3Setter:) int p3;
-(void)myOwnP3Setter:(int)a;
@end

@implementation I1
@synthesize p3;
-(void)myOwnP3Setter:(int)a{ }
@end

```

The DWARF for this would be:

```

0x000003bd: TAG_structure_type [7] *
        AT_APPLE_runtime_class( 0x10 )
        AT_name( "I1" )
        AT_decl_file( "Objc_Property.m" )
        AT_decl_line( 3 )

0x000003cd      TAG_APPLE_property
        AT_name ( "p3" )
        AT_APPLE_property_setter ( "myOwnP3Setter:" )
        AT_type( {0x00000147} ( int ) )

0x000003f3:      TAG_member [8]
        AT_name( "_p3" )
        AT_type ( {0x00000147} ( int ) )
        AT_APPLE_property ( {0x000003cd} )
        AT_artificial ( 0x1 )

```

New DWARF Tags

TAG	Value
DW_TAG_APPLE_property	0x4200

New DWARF Attributes

Attribute	Value	Classes
DW_AT_APPLE_property	0x3fed	Reference
DW_AT_APPLE_property_getter	0x3fe9	String
DW_AT_APPLE_property_setter	0x3fea	String
DW_AT_APPLE_property_attribute	0x3feb	Constant

New DWARF Constants

Name	Value
DW_APPLE_PROPERTY_readonly	0x01
DW_APPLE_PROPERTY_getter	0x02
DW_APPLE_PROPERTY_assign	0x04
DW_APPLE_PROPERTY_readwrite	0x08
DW_APPLE_PROPERTY_retain	0x10
DW_APPLE_PROPERTY_copy	0x20
DW_APPLE_PROPERTY_nonatomic	0x40
DW_APPLE_PROPERTY_setter	0x80
DW_APPLE_PROPERTY_atomic	0x100
DW_APPLE_PROPERTY_weak	0x200
DW_APPLE_PROPERTY_strong	0x400
DW_APPLE_PROPERTY_unsafe_unretained	0x800
DW_APPLE_PROPERTY_nullability	0x1000
DW_APPLE_PROPERTY_null_resettable	0x2000
DW_APPLE_PROPERTY_class	0x4000

Name Accelerator Tables

Introduction

The “.debug_pubnames” and “.debug_pubtypes” formats are not what a debugger needs. The “pub” in the section name indicates that the entries in the table are publicly visible names only. This means no static or hidden functions show up in the “.debug_pubnames”. No static variables or private class variables are in the “.debug_pubtypes”. Many compilers add different things to these tables, so we can’t rely upon the contents between gcc, icc, or clang.

The typical query given by users tends not to match up with the contents of these tables. For example, the DWARF spec states that “In the case of the name of a function member or static data member of a C++ structure, class or union, the name presented in the “.debug_pubnames” section is not the simple name given by the DW_AT_name attribute of the referenced debugging information entry, but rather the fully qualified name of the data or function member.” So the only names in these tables for complex C++ entries is a fully qualified name. Debugger users

tend not to enter their search strings as `"a::b::c(int, const Foo&) const"`, but rather as `"c"`, `"b::c"`, or `"a::b::c"`. So the name entered in the name table must be demangled in order to chop it up appropriately and additional names must be manually entered into the table to make it effective as a name lookup table for debuggers to use.

All debuggers currently ignore the `".debug_pubnames"` table as a result of its inconsistent and useless public-only name content making it a waste of space in the object file. These tables, when they are written to disk, are not sorted in any way, leaving every debugger to do its own parsing and sorting. These tables also include an inlined copy of the string values in the table itself making the tables much larger than they need to be on disk, especially for large C++ programs.

Can't we just fix the sections by adding all of the names we need to this table? No, because that is not what the tables are defined to contain and we won't know the difference between the old bad tables and the new good tables. At best we could make our own renamed sections that contain all of the data we need.

These tables are also insufficient for what a debugger like LLDB needs. LLDB uses clang for its expression parsing where LLDB acts as a PCH. LLDB is then often asked to look for type `"foo"` or namespace `"bar"`, or list items in namespace `"baz"`. Namespaces are not included in the pubnames or pubtypes tables. Since clang asks a lot of questions when it is parsing an expression, we need to be very fast when looking up names, as it happens a lot. Having new accelerator tables that are optimized for very quick lookups will benefit this type of debugging experience greatly.

We would like to generate name lookup tables that can be mapped into memory from disk, and used as is, with little or no up-front parsing. We would also be able to control the exact content of these different tables so they contain exactly what we need. The Name Accelerator Tables were designed to fix these issues. In order to solve these issues we need to:

- Have a format that can be mapped into memory from disk and used as is
- Lookups should be very fast
- Extensible table format so these tables can be made by many producers
- Contain all of the names needed for typical lookups out of the box
- Strict rules for the contents of tables

Table size is important and the accelerator table format should allow the reuse of strings from common string tables so the strings for the names are not duplicated. We also want to make sure the table is ready to be used as-is by simply mapping the table into memory with minimal header parsing.

The name lookups need to be fast and optimized for the kinds of lookups that debuggers tend to do. Optimally we would like to touch as few parts of the mapped table as possible when doing a name lookup and be able to quickly find the name entry we are looking for, or discover there are no matches. In the case of debuggers we optimized for lookups that fail most of the time.

Each table that is defined should have strict rules on exactly what is in the accelerator tables and documented so clients can rely on the content.

Hash Tables

Standard Hash Tables

Typical hash tables have a header, buckets, and each bucket points to the bucket contents:



```
| DATA |
|-----|
```

The BUCKETS are an array of offsets to DATA for each hash:

```
.-----
| 0x00001000 | BUCKETS[0]
| 0x00002000 | BUCKETS[1]
| 0x00002200 | BUCKETS[2]
| 0x000034f0 | BUCKETS[3]
|             | ...
| 0xFFFFFFFF | BUCKETS[n_buckets]
|-----|
```

So for `bucket[3]` in the example above, we have an offset into the table `0x000034f0` which points to a chain of entries for the bucket. Each bucket must contain a next pointer, full 32 bit hash value, the string itself, and the data for the current string value.

```
0x000034f0: .-----
| 0x00003500 | next pointer
| 0x12345678 | 32 bit hash
| "erase"    | string value
| data[n]    | HashData for this bucket
|-----|
0x00003500: | 0x00003550 | next pointer
| 0x29273623 | 32 bit hash
| "dump"     | string value
| data[n]    | HashData for this bucket
|-----|
0x00003550: | 0x00000000 | next pointer
| 0x82638293 | 32 bit hash
| "main"     | string value
| data[n]    | HashData for this bucket
|-----|
```

The problem with this layout for debuggers is that we need to optimize for the negative lookup case where the symbol we’re searching for is not present. So if we were to lookup “`printf`” in the table above, we would make a 32-bit hash for “`printf`”, it might match `bucket[3]`. We would need to go to the offset `0x000034f0` and start looking to see if our 32 bit hash matches. To do so, we need to read the next pointer, then read the hash, compare it, and skip to the next bucket. Each time we are skipping many bytes in memory and touching new pages just to do the compare on the full 32 bit hash. All of these accesses then tell us that we didn’t have a match.

Name Hash Tables

To solve the issues mentioned above we have structured the hash tables a bit differently: a header, buckets, an array of all unique 32 bit hash values, followed by an array of hash value data offsets, one for each hash value, then the data for all hash values:

```
.-----
| HEADER     |
|-----|
| BUCKETS    |
|-----|
| HASHES     |
|-----|
| OFFSETS    |
|-----|
```

```

|-----|
| DATA |
|-----|

```

The BUCKETS in the name tables are an index into the HASHES array. By making all of the full 32 bit hash values contiguous in memory, we allow ourselves to efficiently check for a match while touching as little memory as possible. Most often checking the 32 bit hash values is as far as the lookup goes. If it does match, it usually is a match with no collisions. So for a table with “n_buckets” buckets, and “n_hashes” unique 32 bit hash values, we can clarify the contents of the BUCKETS, HASHES and OFFSETS as:

```

|-----|
| HEADER.magic      | uint32_t
| HEADER.version    | uint16_t
| HEADER.hash_function | uint16_t
| HEADER.bucket_count | uint32_t
| HEADER.hashes_count | uint32_t
| HEADER.header_data_len | uint32_t
| HEADER_DATA       | HeaderData
|-----|
| BUCKETS           | uint32_t[n_buckets] // 32 bit hash indexes
|-----|
| HASHES            | uint32_t[n_hashes] // 32 bit hash values
|-----|
| OFFSETS           | uint32_t[n_hashes] // 32 bit offsets to hash value data
|-----|
| ALL HASH DATA    |
|-----|

```

So taking the exact same data from the standard hash example above we end up with:

```

|-----|
| HEADER |
|-----|
|         0 | BUCKETS[0]
|         2 | BUCKETS[1]
|         5 | BUCKETS[2]
|         6 | BUCKETS[3]
|         ...
|         ... | BUCKETS[n_buckets]
|-----|
| 0x..... | HASHES[0]
| 0x..... | HASHES[1]
| 0x..... | HASHES[2]
| 0x..... | HASHES[3]
| 0x..... | HASHES[4]
| 0x..... | HASHES[5]
| 0x12345678 | HASHES[6]    hash for BUCKETS[3]
| 0x29273623 | HASHES[7]    hash for BUCKETS[3]
| 0x82638293 | HASHES[8]    hash for BUCKETS[3]
| 0x..... | HASHES[9]
| 0x..... | HASHES[10]
| 0x..... | HASHES[11]
| 0x..... | HASHES[12]
| 0x..... | HASHES[13]
| 0x..... | HASHES[n_hashes]
|-----|
| 0x..... | OFFSETS[0]
| 0x..... | OFFSETS[1]

```



```

| 0x..... | OFFSETS[2]
| 0x..... | OFFSETS[3]
| 0x..... | OFFSETS[4]
| 0x..... | OFFSETS[5]
| 0x000034f0 | OFFSETS[6]   offset for BUCKETS[3]
| 0x00003500 | OFFSETS[7]   offset for BUCKETS[3]
| 0x00003550 | OFFSETS[8]   offset for BUCKETS[3]
| 0x..... | OFFSETS[9]
| 0x..... | OFFSETS[10]
| 0x..... | OFFSETS[11]
| 0x..... | OFFSETS[12]
| 0x..... | OFFSETS[13]
| 0x..... | OFFSETS[n_hashes]
|-----|
|
|
|
|
|-----|
0x000034f0: | 0x00001203 | .debug_str ("erase")
| 0x00000004 | A 32 bit array count - number of HashData with name "erase"
| 0x..... | HashData[0]
| 0x..... | HashData[1]
| 0x..... | HashData[2]
| 0x..... | HashData[3]
| 0x00000000 | String offset into .debug_str (terminate data for hash)
|-----|
0x00003500: | 0x00001203 | String offset into .debug_str ("collision")
| 0x00000002 | A 32 bit array count - number of HashData with name
↪ "collision"
| 0x..... | HashData[0]
| 0x..... | HashData[1]
| 0x00001203 | String offset into .debug_str ("dump")
| 0x00000003 | A 32 bit array count - number of HashData with name "dump"
| 0x..... | HashData[0]
| 0x..... | HashData[1]
| 0x..... | HashData[2]
| 0x00000000 | String offset into .debug_str (terminate data for hash)
|-----|
0x00003550: | 0x00001203 | String offset into .debug_str ("main")
| 0x00000009 | A 32 bit array count - number of HashData with name "main"
| 0x..... | HashData[0]
| 0x..... | HashData[1]
| 0x..... | HashData[2]
| 0x..... | HashData[3]
| 0x..... | HashData[4]
| 0x..... | HashData[5]
| 0x..... | HashData[6]
| 0x..... | HashData[7]
| 0x..... | HashData[8]
| 0x00000000 | String offset into .debug_str (terminate data for hash)
|-----|

```

So we still have all of the same data, we just organize it more efficiently for debugger lookup. If we repeat the same “printf” lookup from above, we would hash “printf” and find it matches BUCKETS[3] by taking the 32 bit hash value and modulo it by n_buckets. BUCKETS[3] contains “6” which is the index into the HASHES table. We would then compare any consecutive 32 bit hashes values in the HASHES array as long as the hashes would be in

BUCKETS[3]. We do this by verifying that each subsequent hash value modulo `n_buckets` is still 3. In the case of a failed lookup we would access the memory for BUCKETS[3], and then compare a few consecutive 32 bit hashes before we know that we have no match. We don't end up marching through multiple words of memory and we really keep the number of processor data cache lines being accessed as small as possible.

The string hash that is used for these lookup tables is the Daniel J. Bernstein hash which is also used in the ELF GNU_HASH sections. It is a very good hash for all kinds of names in programs with very few hash collisions.

Empty buckets are designated by using an invalid hash index of `UINT32_MAX`.

Details

These name hash tables are designed to be generic where specializations of the table get to define additional data that goes into the header ("HeaderData"), how the string value is stored ("KeyType") and the content of the data for each hash value.

Header Layout

The header has a fixed part, and the specialized part. The exact format of the header is:

```
struct Header
{
    uint32_t    magic;           // 'HASH' magic value to allow endian detection
    uint16_t    version;         // Version number
    uint16_t    hash_function;    // The hash function enumeration that was used
    uint32_t    bucket_count;    // The number of buckets in this hash table
    uint32_t    hashes_count;    // The total number of unique hash values and hash data_
    ↪offsets in this table
    uint32_t    header_data_len; // The bytes to skip to get to the hash indexes_
    ↪(buckets) for correct alignment
                                // Specifically the length of the following HeaderData_
    ↪field - this does not
                                // include the size of the preceding fields
    HeaderData header_data;      // Implementation specific header data
};
```

The header starts with a 32 bit “magic” value which must be 'HASH' encoded as an ASCII integer. This allows the detection of the start of the hash table and also allows the table's byte order to be determined so the table can be correctly extracted. The “magic” value is followed by a 16 bit version number which allows the table to be revised and modified in the future. The current version number is 1. `hash_function` is a `uint16_t` enumeration that specifies which hash function was used to produce this table. The current values for the hash function enumerations include:

```
enum HashFunctionType
{
    eHashFunctionDJB = 0u, // Daniel J Bernstein hash function
};
```

`bucket_count` is a 32 bit unsigned integer that represents how many buckets are in the BUCKETS array. `hashes_count` is the number of unique 32 bit hash values that are in the HASHES array, and is the same number of offsets are contained in the OFFSETS array. `header_data_len` specifies the size in bytes of the HeaderData that is filled in by specialized versions of this table.

Fixed Lookup

The header is followed by the buckets, hashes, offsets, and hash value data.

```
struct FixedTable
{
    uint32_t buckets[Header.bucket_count]; // An array of hash indexes into the
    ↪ "hashes[]" array below
    uint32_t hashes [Header.hashes_count]; // Every unique 32 bit hash for the entire
    ↪ table is in this table
    uint32_t offsets[Header.hashes_count]; // An offset that corresponds to each item
    ↪ in the "hashes[]" array above
};
```

buckets is an array of 32 bit indexes into the hashes array. The hashes array contains all of the 32 bit hash values for all names in the hash table. Each hash in the hashes table has an offset in the offsets array that points to the data for the hash value.

This table setup makes it very easy to repurpose these tables to contain different data, while keeping the lookup mechanism the same for all tables. This layout also makes it possible to save the table to disk and map it in later and do very efficient name lookups with little or no parsing.

DWARF lookup tables can be implemented in a variety of ways and can store a lot of information for each name. We want to make the DWARF tables extensible and able to store the data efficiently so we have used some of the DWARF features that enable efficient data storage to define exactly what kind of data we store for each name.

The HeaderData contains a definition of the contents of each HashData chunk. We might want to store an offset to all of the debug information entries (DIEs) for each name. To keep things extensible, we create a list of items, or Atoms, that are contained in the data for each name. First comes the type of the data in each atom:

```
enum AtomType
{
    eAtomTypeNULL      = 0u,
    eAtomTypeDIEOffset  = 1u, // DIE offset, check form for encoding
    eAtomTypeCUOffset   = 2u, // DIE offset of the compiler unit header that contains
    ↪ the item in question
    eAtomTypeTag        = 3u, // DW_TAG_XXX value, should be encoded as DW_FORM_data1
    ↪ (if no tags exceed 255) or DW_FORM_data2
    eAtomTypeNameFlags  = 4u, // Flags from enum NameFlags
    eAtomTypeTypeFlags  = 5u, // Flags from enum TypeFlags
};
```

The enumeration values and their meanings are:

```
eAtomTypeNULL      - a termination atom that specifies the end of the atom list
eAtomTypeDIEOffset  - an offset into the .debug_info section for the DWARF DIE for
    ↪ this name
eAtomTypeCUOffset   - an offset into the .debug_info section for the CU that contains
    ↪ the DIE
eAtomTypeDIETag     - The DW_TAG_XXX enumeration value so you don't have to parse the
    ↪ DWARF to see what it is
eAtomTypeNameFlags  - Flags for functions and global variables (isFunction, isInlined,
    ↪ isExternal...)
eAtomTypeTypeFlags  - Flags for types (isCXXClass, isObjCClass, ...)
```

Then we allow each atom type to define the atom type and how the data for each atom type data is encoded:

```

struct Atom
{
    uint16_t type;    // AtomType enum value
    uint16_t form;    // DWARF DW_FORM_XXX defines
};

```

The `form` type above is from the DWARF specification and defines the exact encoding of the data for the `Atom` type. See the DWARF specification for the `DW_FORM_` definitions.

```

struct HeaderData
{
    uint32_t die_offset_base;
    uint32_t atom_count;
    Atoms    atoms[atom_count0];
};

```

`HeaderData` defines the base DIE offset that should be added to any atoms that are encoded using the `DW_FORM_ref1`, `DW_FORM_ref2`, `DW_FORM_ref4`, `DW_FORM_ref8` or `DW_FORM_ref_udata`. It also defines what is contained in each `HashData` object – `Atom.form` tells us how large each field will be in the `HashData` and the `Atom.type` tells us how this data should be interpreted.

For the current implementations of the “`.apple_names`” (all functions + globals), the “`.apple_types`” (names of all types that are defined), and the “`.apple_namespaces`” (all namespaces), we currently set the `Atom` array to be:

```

HeaderData.atom_count = 1;
HeaderData.atoms[0].type = eAtomTypeDIEOffset;
HeaderData.atoms[0].form = DW_FORM_data4;

```

This defines the contents to be the DIE offset (`eAtomTypeDIEOffset`) that is encoded as a 32 bit value (`DW_FORM_data4`). This allows a single name to have multiple matching DIEs in a single file, which could come up with an inlined function for instance. Future tables could include more information about the DIE such as flags indicating if the DIE is a function, method, block, or inlined.

The `KeyType` for the DWARF table is a 32 bit string table offset into the “`.debug_str`” table. The “`.debug_str`” is the string table for the DWARF which may already contain copies of all of the strings. This helps make sure, with help from the compiler, that we reuse the strings between all of the DWARF sections and keeps the hash table size down. Another benefit to having the compiler generate all strings as `DW_FORM_strp` in the debug info, is that DWARF parsing can be made much faster.

After a lookup is made, we get an offset into the hash data. The hash data needs to be able to deal with 32 bit hash collisions, so the chunk of data at the offset in the hash data consists of a triple:

```

uint32_t str_offset
uint32_t hash_data_count
HashData[hash_data_count]

```

If “`str_offset`” is zero, then the bucket contents are done. 99.9% of the hash data chunks contain a single item (no 32 bit hash collision):

```

.------.
| 0x00001023 | uint32_t KeyType (.debug_str[0x0001023] => "main")
| 0x00000004 | uint32_t HashData count
| 0x.....   | uint32_t HashData[0] DIE offset
| 0x.....   | uint32_t HashData[1] DIE offset
| 0x.....   | uint32_t HashData[2] DIE offset
| 0x.....   | uint32_t HashData[3] DIE offset

```

```
| 0x00000000 | uint32_t KeyType (end of hash chain)
|-----|
```

If there are collisions, you will have multiple valid string offsets:

```
|-----|
| 0x00001023 | uint32_t KeyType (.debug_str[0x0001023] => "main")
| 0x00000004 | uint32_t HashData count
| 0x..... | uint32_t HashData[0] DIE offset
| 0x..... | uint32_t HashData[1] DIE offset
| 0x..... | uint32_t HashData[2] DIE offset
| 0x..... | uint32_t HashData[3] DIE offset
| 0x00002023 | uint32_t KeyType (.debug_str[0x0002023] => "print")
| 0x00000002 | uint32_t HashData count
| 0x..... | uint32_t HashData[0] DIE offset
| 0x..... | uint32_t HashData[1] DIE offset
| 0x00000000 | uint32_t KeyType (end of hash chain)
|-----|
```

Current testing with real world C++ binaries has shown that there is around 1 32 bit hash collision per 100,000 name entries.

Contents

As we said, we want to strictly define exactly what is included in the different tables. For DWARF, we have 3 tables: “.apple_names”, “.apple_types”, and “.apple_namespaces”.

“.apple_names” sections should contain an entry for each DWARF DIE whose DW_TAG is a DW_TAG_label, DW_TAG_inlined_subroutine, or DW_TAG_subprogram that has address attributes: DW_AT_low_pc, DW_AT_high_pc, DW_AT_ranges or DW_AT_entry_pc. It also contains DW_TAG_variable DIEs that have a DW_OP_addr in the location (global and static variables). All global and static variables should be included, including those scoped within functions and classes. For example using the following code:

```
static int var = 0;

void f ()
{
    static int var = 0;
}
```

Both of the static var variables would be included in the table. All functions should emit both their full names and their basenames. For C or C++, the full name is the mangled name (if available) which is usually in the DW_AT_MIPS_linkage_name attribute, and the DW_AT_name contains the function basename. If global or static variables have a mangled name in a DW_AT_MIPS_linkage_name attribute, this should be emitted along with the simple name found in the DW_AT_name attribute.

“.apple_types” sections should contain an entry for each DWARF DIE whose tag is one of:

- DW_TAG_array_type
- DW_TAG_class_type
- DW_TAG_enumeration_type
- DW_TAG_pointer_type
- DW_TAG_reference_type
- DW_TAG_string_type

- DW_TAG_structure_type
- DW_TAG_subroutine_type
- DW_TAG_typedef
- DW_TAG_union_type
- DW_TAG_ptr_to_member_type
- DW_TAG_set_type
- DW_TAG_subrange_type
- DW_TAG_base_type
- DW_TAG_const_type
- DW_TAG_file_type
- DW_TAG_namelist
- DW_TAG_packed_type
- DW_TAG_volatile_type
- DW_TAG_restrict_type
- DW_TAG_atomic_type
- DW_TAG_interface_type
- DW_TAG_unspecified_type
- DW_TAG_shared_type

Only entries with a DW_AT_name attribute are included, and the entry must not be a forward declaration (DW_AT_declaration attribute with a non-zero value). For example, using the following code:

```
int main ()
{
    int *b = 0;
    return *b;
}
```

We get a few type DIEs:

```
0x00000067:      TAG_base_type [5]
                AT_encoding( DW_ATE_signed )
                AT_name( "int" )
                AT_byte_size( 0x04 )

0x0000006e:      TAG_pointer_type [6]
                AT_type( {0x00000067} ( int ) )
                AT_byte_size( 0x08 )
```

The DW_TAG_pointer_type is not included because it does not have a DW_AT_name.

“.apple_namespaces” section should contain all DW_TAG_namespace DIEs. If we run into a namespace that has no name this is an anonymous namespace, and the name should be output as “(anonymous namespace)” (without the quotes). Why? This matches the output of the `abi::cxa_demangle()` that is in the standard C++ library that demangles mangled names.

Language Extensions and File Format Changes

Objective-C Extensions

“.apple_objc” section should contain all DW_TAG_subprogram DIEs for an Objective-C class. The name used in the hash table is the name of the Objective-C class itself. If the Objective-C class has a category, then an entry is made for both the class name without the category, and for the class name with the category. So if we have a DIE at offset 0x1234 with a name of method “-[NSString(my_additions) stringWithSpecialString:]”, we would add an entry for “NSString” that points to DIE 0x1234, and an entry for “NSString(my_additions)” that points to 0x1234. This allows us to quickly track down all Objective-C methods for an Objective-C class when doing expressions. It is needed because of the dynamic nature of Objective-C where anyone can add methods to a class. The DWARF for Objective-C methods is also emitted differently from C++ classes where the methods are not usually contained in the class definition, they are scattered about across one or more compile units. Categories can also be defined in different shared libraries. So we need to be able to quickly find all of the methods and class functions given the Objective-C class name, or quickly find all methods and class functions for a class + category name. This table does not contain any selector names, it just maps Objective-C class names (or class names + category) to all of the methods and class functions. The selectors are added as function basenames in the “.debug_names” section.

In the “.apple_names” section for Objective-C functions, the full name is the entire function name with the brackets (“-[NSString stringWithCString:]”) and the basename is the selector only (“stringWithCString:”).

Mach-O Changes

The sections names for the apple hash tables are for non-mach-o files. For mach-o files, the sections should be contained in the __DWARF segment with names as follows:

- “.apple_names” -> “__apple_names”
- “.apple_types” -> “__apple_types”
- “.apple_namespaces” -> “__apple_namespac” (16 character limit)
- “.apple_objc” -> “__apple_objc”

4.17.6 CodeView Debug Info Format

LLVM supports emitting CodeView, the Microsoft debug info format, and this section describes the design and implementation of that support.

Format Background

CodeView as a format is clearly oriented around C++ debugging, and in C++, the majority of debug information tends to be type information. Therefore, the overriding design constraint of CodeView is the separation of type information from other “symbol” information so that type information can be efficiently merged across translation units. Both type information and symbol information is generally stored as a sequence of records, where each record begins with a 16-bit record size and a 16-bit record kind.

Type information is usually stored in the .debug\$T section of the object file. All other debug info, such as line info, string table, symbol info, and inlinee info, is stored in one or more .debug\$S sections. There may only be one .debug\$T section per object file, since all other debug info refers to it. If a PDB (enabled by the /Zi MSVC option) was used during compilation, the .debug\$T section will contain only an LF_TYPESERVER2 record pointing to the PDB. When using PDBs, symbol information appears to remain in the object file .debug\$S sections.

Type records are referred to by their index, which is the number of records in the stream before a given record plus `0x1000`. Many common basic types, such as the basic integral types and unqualified pointers to them, are represented using type indices less than `0x1000`. Such basic types are built in to CodeView consumers and do not require type records.

Each type record may only contain type indices that are less than its own type index. This ensures that the graph of type stream references is acyclic. While the source-level type graph may contain cycles through pointer types (consider a linked list struct), these cycles are removed from the type stream by always referring to the forward declaration record of user-defined record types. Only “symbol” records in the `.debug$S` streams may refer to complete, non-forward-declaration type records.

Working with CodeView

These are instructions for some common tasks for developers working to improve LLVM’s CodeView support. Most of them revolve around using the CodeView dumper embedded in `llvm-readobj`.

- Testing MSVC’s output:

```
$ cl -c -Z7 foo.cpp # Use /Z7 to keep types in the object file
$ llvm-readobj -codeview foo.obj
```

- Getting LLVM IR debug info out of Clang:

```
$ clang -g -gcodeview --target=x86_64-windows-msvc foo.cpp -S -emit-llvm
```

Use this to generate LLVM IR for LLVM test cases.

- Generate and dump CodeView from LLVM IR metadata:

```
$ llc foo.ll -filetype=obj -o foo.obj
$ llvm-readobj -codeview foo.obj > foo.txt
```

Use this pattern in lit test cases and FileCheck the output of `llvm-readobj`

Improving LLVM’s CodeView support is a process of finding interesting type records, constructing a C++ test case that makes MSVC emit those records, dumping the records, understanding them, and then generating equivalent records in LLVM’s backend.

4.18 Auto-Vectorization in LLVM

- *The Loop Vectorizer*
 - *Usage*
 - * *Command line flags*
 - * *Pragma loop hint directives*
 - *Diagnostics*
 - *Features*
 - * *Loops with unknown trip count*
 - * *Runtime Checks of Pointers*

- * *Reductions*
- * *Inductions*
- * *If Conversion*
- * *Pointer Induction Variables*
- * *Reverse Iterators*
- * *Scatter / Gather*
- * *Vectorization of Mixed Types*
- * *Global Structures Alias Analysis*
- * *Vectorization of function calls*
- * *Partial unrolling during vectorization*
- *Performance*
- *Ongoing Development Directions*
- *The SLP Vectorizer*
 - *Details*
 - *Usage*

LLVM has two vectorizers: The *Loop Vectorizer*, which operates on Loops, and the *SLP Vectorizer*. These vectorizers focus on different optimization opportunities and use different techniques. The SLP vectorizer merges multiple scalars that are found in the code into vectors while the Loop Vectorizer widens instructions in loops to operate on multiple consecutive iterations.

Both the Loop Vectorizer and the SLP Vectorizer are enabled by default.

4.18.1 The Loop Vectorizer

Usage

The Loop Vectorizer is enabled by default, but it can be disabled through clang using the command line flag:

```
$ clang ... -fno-vectorize file.c
```

Command line flags

The loop vectorizer uses a cost model to decide on the optimal vectorization factor and unroll factor. However, users of the vectorizer can force the vectorizer to use specific values. Both ‘clang’ and ‘opt’ support the flags below.

Users can control the vectorization SIMD width using the command line flag “-force-vector-width”.

```
$ clang -mllvm -force-vector-width=8 ...
$ opt -loop-vectorize -force-vector-width=8 ...
```

Users can control the unroll factor using the command line flag “-force-vector-interleave”

```
$ clang -mllvm -force-vector-interleave=2 ...
$ opt -loop-vectorize -force-vector-interleave=2 ...
```

Pragma loop hint directives

The `#pragma clang loop` directive allows loop vectorization hints to be specified for the subsequent `for`, `while`, `do-while`, or `c++11` range-based `for` loop. The directive allows vectorization and interleaving to be enabled or disabled. Vector width as well as interleave count can also be manually specified. The following example explicitly enables vectorization and interleaving:

```
#pragma clang loop vectorize(enable) interleave(enable)
while(...) {
    ...
}
```

The following example implicitly enables vectorization and interleaving by specifying a vector width and interleaving count:

```
#pragma clang loop vectorize_width(2) interleave_count(2)
for(...) {
    ...
}
```

See the Clang [language extensions](#) for details.

Diagnostics

Many loops cannot be vectorized including loops with complicated control flow, unvectorizable types, and unvectorizable calls. The loop vectorizer generates optimization remarks which can be queried using command line options to identify and diagnose loops that are skipped by the loop-vectorizer.

Optimization remarks are enabled using:

`-Rpass=loop-vectorize` identifies loops that were successfully vectorized.

`-Rpass-missed=loop-vectorize` identifies loops that failed vectorization and indicates if vectorization was specified.

`-Rpass-analysis=loop-vectorize` identifies the statements that caused vectorization to fail. If in addition `-fsave-optimization-record` is provided, multiple causes of vectorization failure may be listed (this behavior might change in the future).

Consider the following loop:

```
#pragma clang loop vectorize(enable)
for (int i = 0; i < Length; i++) {
    switch(A[i]) {
        case 0: A[i] = i*2; break;
        case 1: A[i] = i;   break;
        default: A[i] = 0;
    }
}
```

The command line `-Rpass-missed=loop-vectorized` prints the remark:

```
no_switch.cpp:4:5: remark: loop not vectorized: vectorization is explicitly enabled [-
↪Rpass-missed=loop-vectorize]
```

And the command line `-Rpass-analysis=loop-vectorize` indicates that the switch statement cannot be vectorized.

```
no_switch.cpp:4:5: remark: loop not vectorized: loop contains a switch statement [-
↳Rpass-analysis=loop-vectorize]
    switch(A[i]) {
      ^
```

To ensure line and column numbers are produced include the command line options `-gline-tables-only` and `-gcolumn-info`. See the Clang [user manual](#) for details

Features

The LLVM Loop Vectorizer has a number of features that allow it to vectorize complex loops.

Loops with unknown trip count

The Loop Vectorizer supports loops with an unknown trip count. In the loop below, the iteration start and finish points are unknown, and the Loop Vectorizer has a mechanism to vectorize loops that do not start at zero. In this example, ‘n’ may not be a multiple of the vector width, and the vectorizer has to execute the last few iterations as scalar code. Keeping a scalar copy of the loop increases the code size.

```
void bar(float *A, float* B, float K, int start, int end) {
    for (int i = start; i < end; ++i)
        A[i] *= B[i] + K;
}
```

Runtime Checks of Pointers

In the example below, if the pointers A and B point to consecutive addresses, then it is illegal to vectorize the code because some elements of A will be written before they are read from array B.

Some programmers use the ‘restrict’ keyword to notify the compiler that the pointers are disjointed, but in our example, the Loop Vectorizer has no way of knowing that the pointers A and B are unique. The Loop Vectorizer handles this loop by placing code that checks, at runtime, if the arrays A and B point to disjointed memory locations. If arrays A and B overlap, then the scalar version of the loop is executed.

```
void bar(float *A, float* B, float K, int n) {
    for (int i = 0; i < n; ++i)
        A[i] *= B[i] + K;
}
```

Reductions

In this example the sum variable is used by consecutive iterations of the loop. Normally, this would prevent vectorization, but the vectorizer can detect that ‘sum’ is a reduction variable. The variable ‘sum’ becomes a vector of integers, and at the end of the loop the elements of the array are added together to create the correct result. We support a number of different reduction operations, such as addition, multiplication, XOR, AND and OR.

```
int foo(int *A, int *B, int n) {
    unsigned sum = 0;
    for (int i = 0; i < n; ++i)
        sum += A[i] + 5;
}
```

```
    return sum;
}
```

We support floating point reduction operations when *-ffast-math* is used.

Inductions

In this example the value of the induction variable `i` is saved into an array. The Loop Vectorizer knows to vectorize induction variables.

```
void bar(float *A, float* B, float K, int n) {
    for (int i = 0; i < n; ++i)
        A[i] = i;
}
```

If Conversion

The Loop Vectorizer is able to “flatten” the IF statement in the code and generate a single stream of instructions. The Loop Vectorizer supports any control flow in the innermost loop. The innermost loop may contain complex nesting of IFs, ELSEs and even GOTOs.

```
int foo(int *A, int *B, int n) {
    unsigned sum = 0;
    for (int i = 0; i < n; ++i)
        if (A[i] > B[i])
            sum += A[i] + 5;
    return sum;
}
```

Pointer Induction Variables

This example uses the “accumulate” function of the standard c++ library. This loop uses C++ iterators, which are pointers, and not integer indices. The Loop Vectorizer detects pointer induction variables and can vectorize this loop. This feature is important because many C++ programs use iterators.

```
int baz(int *A, int n) {
    return std::accumulate(A, A + n, 0);
}
```

Reverse Iterators

The Loop Vectorizer can vectorize loops that count backwards.

```
int foo(int *A, int *B, int n) {
    for (int i = n; i > 0; --i)
        A[i] += 1;
}
```

Scatter / Gather

The Loop Vectorizer can vectorize code that becomes a sequence of scalar instructions that scatter/gathers memory.

```
int foo(int * A, int * B, int n) {
    for (intptr_t i = 0; i < n; ++i)
        A[i] += B[i * 4];
}
```

In many situations the cost model will inform LLVM that this is not beneficial and LLVM will only vectorize such code if forced with “-mllvm -force-vector-width=#”.

Vectorization of Mixed Types

The Loop Vectorizer can vectorize programs with mixed types. The Vectorizer cost model can estimate the cost of the type conversion and decide if vectorization is profitable.

```
int foo(int *A, char *B, int n, int k) {
    for (int i = 0; i < n; ++i)
        A[i] += 4 * B[i];
}
```

Global Structures Alias Analysis

Access to global structures can also be vectorized, with alias analysis being used to make sure accesses don’t alias. Run-time checks can also be added on pointer access to structure members.

Many variations are supported, but some that rely on undefined behaviour being ignored (as other compilers do) are still being left un-vectorized.

```
struct { int A[100], K, B[100]; } Foo;

int foo() {
    for (int i = 0; i < 100; ++i)
        Foo.A[i] = Foo.B[i] + 100;
}
```

Vectorization of function calls

The Loop Vectorizer can vectorize intrinsic math functions. See the table below for a list of these functions.

pow	exp	exp2
sin	cos	sqrt
log	log2	log10
fabs	floor	ceil
fma	trunc	nearbyint
		fmuladd

The loop vectorizer knows about special instructions on the target and will vectorize a loop containing a function call that maps to the instructions. For example, the loop below will be vectorized on Intel x86 if the SSE4.1 roundps instruction is available.

```
void foo(float *f) {
    for (int i = 0; i != 1024; ++i)
        f[i] = floorf(f[i]);
}
```

Partial unrolling during vectorization

Modern processors feature multiple execution units, and only programs that contain a high degree of parallelism can fully utilize the entire width of the machine. The Loop Vectorizer increases the instruction level parallelism (ILP) by performing partial-unrolling of loops.

In the example below the entire array is accumulated into the variable ‘sum’. This is inefficient because only a single execution port can be used by the processor. By unrolling the code the Loop Vectorizer allows two or more execution ports to be used simultaneously.

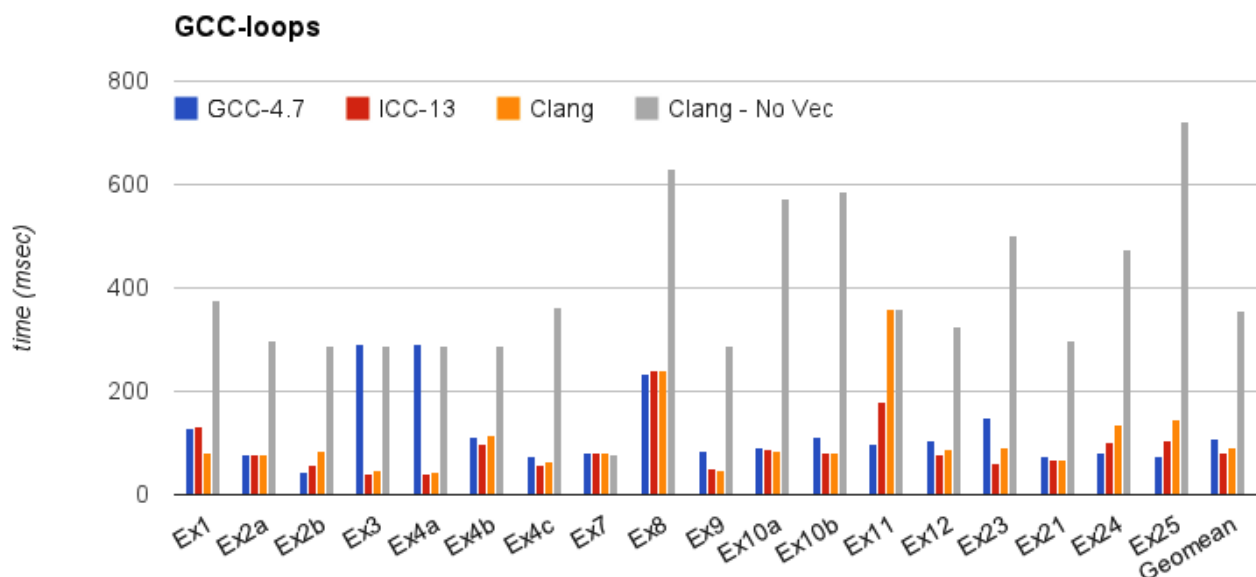
```
int foo(int *A, int *B, int n) {
    unsigned sum = 0;
    for (int i = 0; i < n; ++i)
        sum += A[i];
    return sum;
}
```

The Loop Vectorizer uses a cost model to decide when it is profitable to unroll loops. The decision to unroll the loop depends on the register pressure and the generated code size.

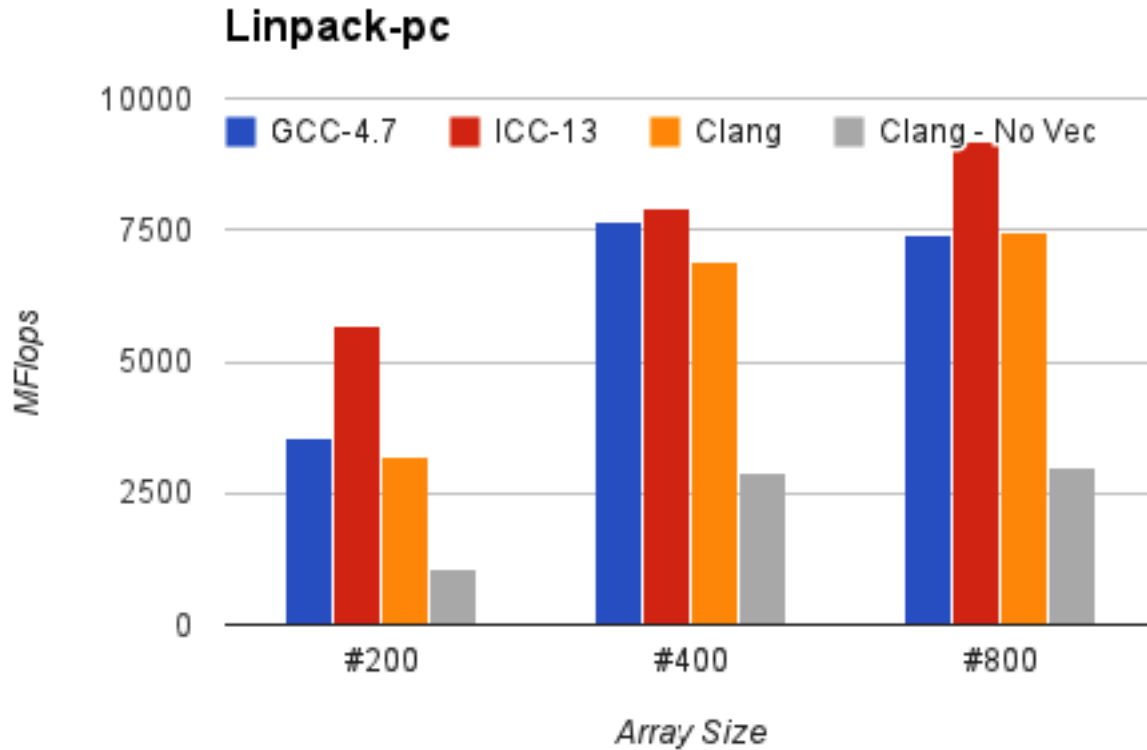
Performance

This section shows the execution time of Clang on a simple benchmark: [gcc-loops](#). This benchmark is a collection of loops from the GCC autovectorization [page](#) by Dorit Nuzman.

The chart below compares GCC-4.7, ICC-13, and Clang-SVN with and without loop vectorization at -O3, tuned for “corei7-avx”, running on a Sandybridge iMac. The Y-axis shows the time in msec. Lower is better. The last column shows the geomean of all the kernels.



And Linpack-pc with the same configuration. Result is Mflops, higher is better.



Ongoing Development Directions

Vectorization Plan

- *Abstract*
- *High-level Design*
 - *Vectorization Workflow*
 - *Design Guidelines*
- *Low-level Design*
 - *Related LLVM components*
 - *References*

Abstract

The vectorization transformation can be rather complicated, involving several potential alternatives, especially for outer-loops¹ but also possibly for innermost loops. These alternatives may have significant performance impact, both positive and negative. A cost model is therefore employed to identify the best alternative, including the alternative of avoiding any transformation altogether.

¹ “Outer-loop vectorization: revisited for short SIMD architectures”, Dorit Nuzman and Ayal Zaks, PACT 2008.

The Vectorization Plan is an explicit model for describing vectorization candidates. It serves for both optimizing candidates including estimating their cost reliably, and for performing their final translation into IR. This facilitates dealing with multiple vectorization candidates.

High-level Design

Vectorization Workflow

VPlan-based vectorization involves three major steps, taking a “scenario-based approach” to vectorization planning:

1. Legal Step: check if a loop can be legally vectorized; encode constraints and artifacts if so.
2. Plan Step:
 - (a) Build initial VPlans following the constraints and decisions taken by Legal Step 1, and compute their cost.
 - (b) Apply optimizations to the VPlans, possibly forking additional VPlans. Prune sub-optimal VPlans having relatively high cost.
3. Execute Step: materialize the best VPlan. Note that this is the only step that modifies the IR.

Design Guidelines

In what follows, the term “input IR” refers to code that is fed into the vectorizer whereas the term “output IR” refers to code that is generated by the vectorizer. The output IR contains code that has been vectorized or “widened” according to a loop Vectorization Factor (VF), and/or loop unroll-and-jammed according to an Unroll Factor (UF). The design of VPlan follows several high-level guidelines:

1. Analysis-like: building and manipulating VPlans must not modify the input IR. In particular, if the best option is not to vectorize at all, the vectorization process terminates before reaching Step 3, and compilation should proceed as if VPlans had not been built.
2. Align Cost & Execute: each VPlan must support both estimating the cost and generating the output IR code, such that the cost estimation evaluates the to-be-generated code reliably.
3. Support vectorizing additional constructs:
 - (a) Outer-loop vectorization. In particular, VPlan must be able to model the control-flow of the output IR which may include multiple basic-blocks and nested loops.
 - (b) SLP vectorization.
 - (c) Combinations of the above, including nested vectorization: vectorizing both an inner loop and an outer-loop at the same time (each with its own VF and UF), mixed vectorization: vectorizing a loop with SLP patterns inside⁴, (re)vectorizing input IR containing vector code.
 - (d) Function vectorization².
4. Support multiple candidates efficiently. In particular, similar candidates related to a range of possible VF’s and UF’s must be represented efficiently. Potential versioning needs to be supported efficiently.
5. Support vectorizing idioms, such as interleaved groups of strided loads or stores. This is achieved by modeling a sequence of output instructions using a “Recipe”, which is responsible for computing its cost and generating its code.

⁴ “Exploiting mixed SIMD parallelism by reducing data reorganization overhead”, Hao Zhou and Jingling Xue, CGO 2016.

² “Proposal for function vectorization and loop vectorization with function calls”, Xinmin Tian, [cfe-dev]., March 2, 2016. See also [review](#).

6. Encapsulate Single-Entry Single-Exit regions (SESE). During vectorization such regions may need to be, for example, predicated and linearized, or replicated $VF \cdot UF$ times to handle scalarized and predicated instructions. Innerloops are also modelled as SESE regions.

Low-level Design

The low-level design of VPlan comprises of the following classes.

LoopVectorizationPlanner A LoopVectorizationPlanner is designed to handle the vectorization of a loop or a loop nest. It can construct, optimize and discard one or more VPlans, each VPlan modelling a distinct way to vectorize the loop or the loop nest. Once the best VPlan is determined, including the best VF and UF, this VPlan drives the generation of output IR.

VPlan A model of a vectorized candidate for a given input IR loop or loop nest. This candidate is represented using a Hierarchical CFG. VPlan supports estimating the cost and driving the generation of the output IR code it represents.

Hierarchical CFG A control-flow graph whose nodes are basic-blocks or Hierarchical CFG's. The Hierarchical CFG data structure is similar to the Tile Tree⁵, where cross-Tile edges are lifted to connect Tiles instead of the original basic-blocks as in Sharir⁶, promoting the Tile encapsulation. The terms Region and Block are used rather than Tile⁵ to avoid confusion with loop tiling.

VPBlockBase The building block of the Hierarchical CFG. A pure-virtual base-class of VPBasicBlock and VPRegionBlock, see below. VPBlockBase models the hierarchical control-flow relations with other VPBlocks. Note that in contrast to the IR BasicBlock, a VPBlockBase models its control-flow successors and predecessors directly, rather than through a Terminator branch or through predecessor branches that “use” the VPBlockBase.

VPBasicBlock VPBasicBlock is a subclass of VPBlockBase, and serves as the leaves of the Hierarchical CFG. It represents a sequence of output IR instructions that will appear consecutively in an output IR basic-block. The instructions of this basic-block originate from one or more VPBasicBlocks. VPBasicBlock holds a sequence of zero or more VPRecipes that model the cost and generation of the output IR instructions.

VPRegionBlock VPRegionBlock is a subclass of VPBlockBase. It models a collection of VPBasicBlocks and VPRegionBlocks which form a SESE subgraph of the output IR CFG. A VPRegionBlock may indicate that its contents are to be replicated a constant number of times when output IR is generated, effectively representing a loop with constant trip-count that will be completely unrolled. This is used to support scalarized and predicated instructions with a single model for multiple candidate VF's and UF's.

VPRecipeBase A pure-virtual base class modeling a sequence of one or more output IR instructions, possibly based on one or more input IR instructions. These input IR instructions are referred to as “Ingredients” of the Recipe. A Recipe may specify how its ingredients are to be transformed to produce the output IR instructions; e.g., cloned once, replicated multiple times or widened according to selected VF.

VPTransformState Stores information used for generating output IR, passed from LoopVectorizationPlanner to its selected VPlan for execution, and used to pass additional information down to VPBlocks and VPRecipes.

⁵ “Register Allocation via Hierarchical Graph Coloring”, David Callahan and Brian Koblenz, PLDI 1991

⁶ “Structural analysis: A new approach to flow analysis in optimizing compilers”, M. Sharir, Journal of Computer Languages, Jan. 1980

Related LLVM components

1. SLP Vectorizer: one can compare the VPlan model with LLVM's existing SLP tree, where TSLP³ adds Plan Step 2.b.
2. RegionInfo: one can compare VPlan's H-CFG with the Region Analysis as used by Polly⁷.

References

Vectorization Plan Modeling the process and upgrading the infrastructure of LLVM's Loop Vectorizer.

4.18.2 The SLP Vectorizer

Details

The goal of SLP vectorization (a.k.a. superword-level parallelism) is to combine similar independent instructions into vector instructions. Memory accesses, arithmetic operations, comparison operations, PHI-nodes, can all be vectorized using this technique.

For example, the following function performs very similar operations on its inputs (a1, b1) and (a2, b2). The basic-block vectorizer may combine these into vector operations.

```
void foo(int a1, int a2, int b1, int b2, int *A) {  
    A[0] = a1*(a1 + b1)/b1 + 50*b1/a1;  
    A[1] = a2*(a2 + b2)/b2 + 50*b2/a2;  
}
```

The SLP-vectorizer processes the code bottom-up, across basic blocks, in search of scalars to combine.

Usage

The SLP Vectorizer is enabled by default, but it can be disabled through clang using the command line flag:

```
$ clang -fno-slp-vectorize file.c
```

LLVM has a second basic block vectorization phase which is more compile-time intensive (The BB vectorizer). This optimization can be enabled through clang using the command line flag:

```
$ clang -fslp-vectorize-aggressive file.c
```

4.19 Writing an LLVM Backend

4.19.1 How To Use Instruction Mappings

- *Introduction*

³ “Throttling Automatic Vectorization: When Less is More”, Vasileios Porpodas and Tim Jones, PACT 2015 and LLVM Developers’ Meeting 2015.

⁷ “Enabling Polyhedral Optimizations in LLVM”, Tobias Grosser, Diploma thesis, 2011.

- *InstrMapping Class Overview*
 - *Sample Example*

Introduction

This document contains information about adding instruction mapping support for a target. The motivation behind this feature comes from the need to switch between different instruction formats during various optimizations. One approach could be to use switch cases which list all the instructions along with formats they can transition to. However, it has large maintenance overhead because of the hardcoded instruction names. Also, whenever a new instruction is added in the .td files, all the relevant switch cases should be modified accordingly. Instead, the same functionality could be achieved with TableGen and some support from the .td files for a fraction of maintenance cost.

InstrMapping Class Overview

TableGen uses relationship models to map instructions with each other. These models are described using InstrMapping class as a base. Each model sets various fields of the InstrMapping class such that they can uniquely describe all the instructions using that model. TableGen parses all the relation models and uses the information to construct relation tables which relate instructions with each other. These tables are emitted in the XXXInstrInfo.inc file along with the functions to query them. Following is the definition of InstrMapping class defined in Target.td file:

```
class InstrMapping {
  // Used to reduce search space only to the instructions using this
  // relation model.
  string FilterClass;

  // List of fields/attributes that should be same for all the instructions in
  // a row of the relation table. Think of this as a set of properties shared
  // by all the instructions related by this relationship.
  list<string> RowFields = [];

  // List of fields/attributes that are same for all the instructions
  // in a column of the relation table.
  list<string> ColFields = [];

  // Values for the fields/attributes listed in 'ColFields' corresponding to
  // the key instruction. This is the instruction that will be transformed
  // using this relation model.
  list<string> KeyCol = [];

  // List of values for the fields/attributes listed in 'ColFields', one for
  // each column in the relation table. These are the instructions a key
  // instruction will be transformed into.
  list<list<string> > ValueCols = [];
}
```

Sample Example

Let's say that we want to have a function `int getPredOpcode(uint16_t Opcode, enum PredSense inPredSense)` which takes a non-predicated instruction and returns its predicated true or false form depending on some input flag, `inPredSense`. The first step in the process is to define a relationship model that relates predicated instructions to their non-predicated form by assigning appropriate values to the InstrMapping fields. For this

relationship, non-predicated instructions are treated as key instruction since they are the one used to query the interface function.

```
def getPredOpcode : InstrMapping {
  // Choose a FilterClass that is used as a base class for all the
  // instructions modeling this relationship. This is done to reduce the
  // search space only to these set of instructions.
  let FilterClass = "PredRel";

  // Instructions with same values for all the fields in RowFields form a
  // row in the resulting relation table.
  // For example, if we want to relate 'ADD' (non-predicated) with 'Add_pt'
  // (predicated true) and 'Add_pf' (predicated false), then all 3
  // instructions need to have same value for BaseOpcode field. It can be any
  // unique value (Ex: XYZ) and should not be shared with any other
  // instruction not related to 'add'.
  let RowFields = ["BaseOpcode"];

  // List of attributes that can be used to define key and column instructions
  // for a relation. Key instruction is passed as an argument
  // to the function used for querying relation tables. Column instructions
  // are the instructions they (key) can transform into.
  //
  // Here, we choose 'PredSense' as ColFields since this is the unique
  // attribute of the key (non-predicated) and column (true/false)
  // instructions involved in this relationship model.
  let ColFields = ["PredSense"];

  // The key column contains non-predicated instructions.
  let KeyCol = ["none"];

  // Two value columns - first column contains instructions with
  // PredSense=true while second column has instructions with PredSense=false.
  let ValueCols = ["true", "false"];
}
```

TableGen uses the above relationship model to emit relation table that maps non-predicated instructions with their predicated forms. It also outputs the interface function `int getPredOpcode(uint16_t Opcode, enum PredSense inPredSense)` to query the table. Here, Function `getPredOpcode` takes two arguments, opcode of the current instruction and `PredSense` of the desired instruction, and returns predicated form of the instruction, if found in the relation table. In order for an instruction to be added into the relation table, it needs to include relevant information in its definition. For example, consider following to be the current definitions of `ADD`, `ADD_pt` (true) and `ADD_pf` (false) instructions:

```
def ADD : ALU32_rr<(outs IntRegs:$dst), (ins IntRegs:$a, IntRegs:$b),
  "$dst = add($a, $b)",
  [(set (i32 IntRegs:$dst), (add (i32 IntRegs:$a),
                                (i32 IntRegs:$b)))]>;

def ADD_Pt : ALU32_rr<(outs IntRegs:$dst),
  (ins PredRegs:$p, IntRegs:$a, IntRegs:$b),
  "if ($p) $dst = add($a, $b)",
  []>;

def ADD_Pf : ALU32_rr<(outs IntRegs:$dst),
  (ins PredRegs:$p, IntRegs:$a, IntRegs:$b),
  "if (!$p) $dst = add($a, $b)",
  []>;
```

In this step, we modify these instructions to include the information required by the relationship model, `<tt>getPredOpcode</tt>`, so that they can be related.

```
def ADD : PredRel, ALU32_rr<(outs IntRegs:$dst), (ins IntRegs:$a, IntRegs:$b),
    "$dst = add($a, $b)",
    [(set (i32 IntRegs:$dst), (add (i32 IntRegs:$a),
                                   (i32 IntRegs:$b)))]> {
    let BaseOpcode = "ADD";
    let PredSense = "none";
}

def ADD_Pt : PredRel, ALU32_rr<(outs IntRegs:$dst),
    (ins PredRegs:$p, IntRegs:$a, IntRegs:$b),
    "if ($p) $dst = add($a, $b)",
    []> {
    let BaseOpcode = "ADD";
    let PredSense = "true";
}

def ADD_Pf : PredRel, ALU32_rr<(outs IntRegs:$dst),
    (ins PredRegs:$p, IntRegs:$a, IntRegs:$b),
    "if (!$p) $dst = add($a, $b)",
    []> {
    let BaseOpcode = "ADD";
    let PredSense = "false";
}
```

Please note that all the above instructions use `PredRel` as a base class. This is extremely important since `TableGen` uses it as a filter for selecting instructions for `getPredOpcode` model. Any instruction not derived from `PredRel` is excluded from the analysis. `BaseOpcode` is another important field. Since it's selected as a `RowFields` of the model, it is required to have the same value for all 3 instructions in order to be related. Next, `PredSense` is used to determine their column positions by comparing its value with `KeyCol` and `ValueCols`. If an instruction sets its `PredSense` value to something not used in the relation model, it will not be assigned a column in the relation table.

- *Introduction*
 - *Audience*
 - *Prerequisite Reading*
 - *Basic Steps*
 - *Preliminaries*
- *Target Machine*
- *Target Registration*
- *Register Set and Register Classes*
 - *Defining a Register*
 - *Defining a Register Class*
 - *Implement a subclass of `TargetRegisterInfo`*
- *Instruction Set*
 - *Instruction Operand Mapping*
 - * *Instruction Operand Name Mapping*

- * *Instruction Operand Types*
 - *Instruction Scheduling*
 - *Instruction Relation Mapping*
 - *Implement a subclass of `TargetInstrInfo`*
 - *Branch Folding and If Conversion*
- *Instruction Selector*
 - *The SelectionDAG Legalize Phase*
 - * *Promote*
 - * *Expand*
 - * *Custom*
 - * *Legal*
 - *Calling Conventions*
- *Assembly Printer*
- *Subtarget Support*
- *JIT Support*
 - *Machine Code Emitter*
 - *Target JIT Info*

4.19.2 Introduction

This document describes techniques for writing compiler backends that convert the LLVM Intermediate Representation (IR) to code for a specified machine or other languages. Code intended for a specific machine can take the form of either assembly code or binary code (usable for a JIT compiler).

The backend of LLVM features a target-independent code generator that may create output for several types of target CPUs — including X86, PowerPC, ARM, and SPARC. The backend may also be used to generate code targeted at SPU of the Cell processor or GPUs to support the execution of compute kernels.

The document focuses on existing examples found in subdirectories of `llvm/lib/Target` in a downloaded LLVM release. In particular, this document focuses on the example of creating a static compiler (one that emits text assembly) for a SPARC target, because SPARC has fairly standard characteristics, such as a RISC instruction set and straightforward calling conventions.

Audience

The audience for this document is anyone who needs to write an LLVM backend to generate code for a specific hardware or software target.

Prerequisite Reading

These essential documents must be read before reading this document:

- [LLVM Language Reference Manual](#) — a reference manual for the LLVM assembly language.

- *The LLVM Target-Independent Code Generator* — a guide to the components (classes and code generation algorithms) for translating the LLVM internal representation into machine code for a specified target. Pay particular attention to the descriptions of code generation stages: Instruction Selection, Scheduling and Formation, SSA-based Optimization, Register Allocation, Prolog/Epilog Code Insertion, Late Machine Code Optimizations, and Code Emission.
- *TableGen* — a document that describes the TableGen (tblgen) application that manages domain-specific information to support LLVM code generation. TableGen processes input from a target description file (.td suffix) and generates C++ code that can be used for code generation.
- *Writing an LLVM Pass* — The assembly printer is a FunctionPass, as are several SelectionDAG processing steps.

To follow the SPARC examples in this document, have a copy of [The SPARC Architecture Manual, Version 8](#) for reference. For details about the ARM instruction set, refer to the [ARM Architecture Reference Manual](#). For more about the GNU Assembler format (GAS), see [Using As](#), especially for the assembly printer. “Using As” contains a list of target machine dependent features.

Basic Steps

To write a compiler backend for LLVM that converts the LLVM IR to code for a specified target (machine or other language), follow these steps:

- Create a subclass of the `TargetMachine` class that describes characteristics of your target machine. Copy existing examples of specific `TargetMachine` class and header files; for example, start with `SparcTargetMachine.cpp` and `SparcTargetMachine.h`, but change the file names for your target. Similarly, change code that references “Sparc” to reference your target.
- Describe the register set of the target. Use TableGen to generate code for register definition, register aliases, and register classes from a target-specific `RegisterInfo.td` input file. You should also write additional code for a subclass of the `TargetRegisterInfo` class that represents the class register file data used for register allocation and also describes the interactions between registers.
- Describe the instruction set of the target. Use TableGen to generate code for target-specific instructions from target-specific versions of `TargetInstrFormats.td` and `TargetInstrInfo.td`. You should write additional code for a subclass of the `TargetInstrInfo` class to represent machine instructions supported by the target machine.
- Describe the selection and conversion of the LLVM IR from a Directed Acyclic Graph (DAG) representation of instructions to native target-specific instructions. Use TableGen to generate code that matches patterns and selects instructions based on additional information in a target-specific version of `TargetInstrInfo.td`. Write code for `XXXISelDAGToDAG.cpp`, where XXX identifies the specific target, to perform pattern matching and DAG-to-DAG instruction selection. Also write code in `XXXISelLowering.cpp` to replace or remove operations and data types that are not supported natively in a `SelectionDAG`.
- Write code for an assembly printer that converts LLVM IR to a GAS format for your target machine. You should add assembly strings to the instructions defined in your target-specific version of `TargetInstrInfo.td`. You should also write code for a subclass of `AsmPrinter` that performs the LLVM-to-assembly conversion and a trivial subclass of `TargetAsmInfo`.
- Optionally, add support for subtargets (i.e., variants with different capabilities). You should also write code for a subclass of the `TargetSubtarget` class, which allows you to use the `-mcpu=` and `-mattr=` command-line options.
- Optionally, add JIT support and create a machine code emitter (subclass of `TargetJITInfo`) that is used to emit binary code directly into memory.

In the .cpp and .h. files, initially stub up these methods and then implement them later. Initially, you may not know which private members that the class will need and which components will need to be subclassed.

Preliminaries

To actually create your compiler backend, you need to create and modify a few files. The absolute minimum is discussed here. But to actually use the LLVM target-independent code generator, you must perform the steps described in the *LLVM Target-Independent Code Generator* document.

First, you should create a subdirectory under `lib/Target` to hold all the files related to your target. If your target is called “Dummy”, create the directory `lib/Target/Dummy`.

In this new directory, create a `CMakeLists.txt`. It is easiest to copy a `CMakeLists.txt` of another target and modify it. It should at least contain the `LLVM_TARGET_DEFINITIONS` variable. The library can be named `LLVMDummy` (for example, see the MIPS target). Alternatively, you can split the library into `LLVMDummyCodeGen` and `LLVMDummyAsmPrinter`, the latter of which should be implemented in a subdirectory below `lib/Target/Dummy` (for example, see the PowerPC target).

Note that these two naming schemes are hardcoded into `llvm-config`. Using any other naming scheme will confuse `llvm-config` and produce a lot of (seemingly unrelated) linker errors when linking `llc`.

To make your target actually do something, you need to implement a subclass of `TargetMachine`. This implementation should typically be in the file `lib/Target/DummyTargetMachine.cpp`, but any file in the `lib/Target` directory will be built and should work. To use LLVM’s target independent code generator, you should do what all current machine backends do: create a subclass of `LLVMTargetMachine`. (To create a target from scratch, create a subclass of `TargetMachine`.)

To get LLVM to actually build and link your target, you need to run `cmake` with `-DLLVM_EXPERIMENTAL_TARGETS_TO_BUILD=Dummy`. This will build your target without needing to add it to the list of all the targets.

Once your target is stable, you can add it to the `LLVM_ALL_TARGETS` variable located in the main `CMakeLists.txt`.

4.19.3 Target Machine

`LLVMTargetMachine` is designed as a base class for targets implemented with the LLVM target-independent code generator. The `LLVMTargetMachine` class should be specialized by a concrete target class that implements the various virtual methods. `LLVMTargetMachine` is defined as a subclass of `TargetMachine` in `include/llvm/Target/TargetMachine.h`. The `TargetMachine` class implementation (`TargetMachine.cpp`) also processes numerous command-line options.

To create a concrete target-specific subclass of `LLVMTargetMachine`, start by copying an existing `TargetMachine` class and header. You should name the files that you create to reflect your specific target. For instance, for the SPARC target, name the files `SparcTargetMachine.h` and `SparcTargetMachine.cpp`.

For a target machine XXX, the implementation of `XXXTargetMachine` must have access methods to obtain objects that represent target components. These methods are named `get*Info`, and are intended to obtain the instruction set (`getInstrInfo`), register set (`getRegisterInfo`), stack frame layout (`getFrameInfo`), and similar information. `XXXTargetMachine` must also implement the `getDataLayout` method to access an object with target-specific data characteristics, such as data type size and alignment requirements.

For instance, for the SPARC target, the header file `SparcTargetMachine.h` declares prototypes for several `get*Info` and `getDataLayout` methods that simply return a class member.

```
namespace llvm {  
  
class Module;  
  
class SparcTargetMachine : public LLVMTargetMachine {  
    const DataLayout DataLayout;           // Calculates type size & alignment  
};
```



```

SparcSubtarget Subtarget;
SparcInstrInfo InstrInfo;
TargetFrameInfo FrameInfo;

protected:
    virtual const TargetAsmInfo *createTargetAsmInfo() const;

public:
    SparcTargetMachine(const Module &M, const std::string &FS);

    virtual const SparcInstrInfo *getInstrInfo() const {return &InstrInfo; }
    virtual const TargetFrameInfo *getFrameInfo() const {return &FrameInfo; }
    virtual const TargetSubtarget *getSubtargetImpl() const{return &Subtarget; }
    virtual const TargetRegisterInfo *getRegisterInfo() const {
        return &InstrInfo.getRegisterInfo();
    }
    virtual const DataLayout *getDataLayout() const { return &DataLayout; }
    static unsigned getModuleMatchQuality(const Module &M);

    // Pass Pipeline Configuration
    virtual bool addInstSelector(PassManagerBase &PM, bool Fast);
    virtual bool addPreEmitPass(PassManagerBase &PM, bool Fast);
};

} // end namespace llvm

```

- `getInstrInfo()`
- `getRegisterInfo()`
- `getFrameInfo()`
- `getDataLayout()`
- `getSubtargetImpl()`

For some targets, you also need to support the following methods:

- `getTargetLowering()`
- `getJITInfo()`

Some architectures, such as GPUs, do not support jumping to an arbitrary program location and implement branching using masked execution and loop using special instructions around the loop body. In order to avoid CFG modifications that introduce irreducible control flow not handled by such hardware, a target must call *setRequiresStructuredCFG(true)* when being initialized.

In addition, the `XXXTargetMachine` constructor should specify a `TargetDescription` string that determines the data layout for the target machine, including characteristics such as pointer size, alignment, and endianness. For example, the constructor for `SparcTargetMachine` contains the following:

```

SparcTargetMachine::SparcTargetMachine(const Module &M, const std::string &FS)
: DataLayout("E-p:32:32-f128:128:128"),
  Subtarget(M, FS), InstrInfo(Subtarget),
  FrameInfo(TargetFrameInfo::StackGrowsDown, 8, 0) {
}

```

Hyphens separate portions of the `TargetDescription` string.

- An upper-case “E” in the string indicates a big-endian target data model. A lower-case “e” indicates little-endian.

- “p:” is followed by pointer information: size, ABI alignment, and preferred alignment. If only two figures follow “p:”, then the first value is pointer size, and the second value is both ABI and preferred alignment.
- Then a letter for numeric type alignment: “i”, “f”, “v”, or “a” (corresponding to integer, floating point, vector, or aggregate). “i”, “v”, or “a” are followed by ABI alignment and preferred alignment. “f” is followed by three values: the first indicates the size of a long double, then ABI alignment, and then ABI preferred alignment.

4.19.4 Target Registration

You must also register your target with the `TargetRegistry`, which is what other LLVM tools use to be able to lookup and use your target at runtime. The `TargetRegistry` can be used directly, but for most targets there are helper templates which should take care of the work for you.

All targets should declare a global `Target` object which is used to represent the target during registration. Then, in the target’s `TargetInfo` library, the target should define that object and use the `RegisterTarget` template to register the target. For example, the Sparc registration code looks like this:

```
Target llvm::getTheSparcTarget();

extern "C" void LLVMInitializeSparcTargetInfo() {
    RegisterTarget<Triple::sparc, /*HasJIT=*/false>
        X(getTheSparcTarget(), "sparc", "Sparc");
}
```

This allows the `TargetRegistry` to look up the target by name or by target triple. In addition, most targets will also register additional features which are available in separate libraries. These registration steps are separate, because some clients may wish to only link in some parts of the target — the JIT code generator does not require the use of the assembler printer, for example. Here is an example of registering the Sparc assembly printer:

```
extern "C" void LLVMInitializeSparcAsmPrinter() {
    RegisterAsmPrinter<SparcAsmPrinter> X(getTheSparcTarget());
}
```

For more information, see “[llvm/Target/TargetRegistry.h](#)”.

4.19.5 Register Set and Register Classes

You should describe a concrete target-specific class that represents the register file of a target machine. This class is called `XXXRegisterInfo` (where XXX identifies the target) and represents the class register file data that is used for register allocation. It also describes the interactions between registers.

You also need to define register classes to categorize related registers. A register class should be added for groups of registers that are all treated the same way for some instruction. Typical examples are register classes for integer, floating-point, or vector registers. A register allocator allows an instruction to use any register in a specified register class to perform the instruction in a similar manner. Register classes allocate virtual registers to instructions from these sets, and register classes let the target-independent register allocator automatically choose the actual registers.

Much of the code for registers, including register definition, register aliases, and register classes, is generated by TableGen from `XXXRegisterInfo.td` input files and placed in `XXXGenRegisterInfo.h.inc` and `XXXGenRegisterInfo.inc` output files. Some of the code in the implementation of `XXXRegisterInfo` requires hand-coding.

Defining a Register

The `XXXRegisterInfo.td` file typically starts with register definitions for a target machine. The `Register` class (specified in `Target.td`) is used to define an object for each register. The specified string `n` becomes the `Name` of the register. The basic `Register` object does not have any subregisters and does not specify any aliases.

```
class Register<string n> {
  string Namespace = "";
  string AsmName = n;
  string Name = n;
  int SpillSize = 0;
  int SpillAlignment = 0;
  list<Register> Aliases = [];
  list<Register> SubRegs = [];
  list<int> DwarfNumbers = [];
}
```

For example, in the `X86RegisterInfo.td` file, there are register definitions that utilize the `Register` class, such as:

```
def AL : Register<"AL">, DwarfRegNum<[0, 0, 0]>;
```

This defines the register `AL` and assigns it values (with `DwarfRegNum`) that are used by `gcc`, `gdb`, or a debug information writer to identify a register. For register `AL`, `DwarfRegNum` takes an array of 3 values representing 3 different modes: the first element is for X86-64, the second for exception handling (EH) on X86-32, and the third is generic. `-1` is a special Dwarf number that indicates the gcc number is undefined, and `-2` indicates the register number is invalid for this mode.

From the previously described line in the `X86RegisterInfo.td` file, TableGen generates this code in the `X86GenRegisterInfo.inc` file:

```
static const unsigned GR8[] = { X86::AL, ... };

const unsigned AL_AliasSet[] = { X86::AX, X86::EAX, X86::RAX, 0 };

const TargetRegisterDesc RegisterDescriptors[] = {
  ...
  { "AL", "AL", AL_AliasSet, Empty_SubRegsSet, Empty_SubRegsSet, AL_SuperRegsSet }, ...
}
```

From the register info file, TableGen generates a `TargetRegisterDesc` object for each register. `TargetRegisterDesc` is defined in `include/llvm/Target/TargetRegisterInfo.h` with the following fields:

```
struct TargetRegisterDesc {
  const char *AsmName;           // Assembly language name for the register
  const char *Name;              // Printable name for the reg (for debugging)
  const unsigned *AliasSet;       // Register Alias Set
  const unsigned *SubRegs;        // Sub-register set
  const unsigned *ImmSubRegs;     // Immediate sub-register set
  const unsigned *SuperRegs;     // Super-register set
};
```

TableGen uses the entire target description file (`.td`) to determine text names for the register (in the `AsmName` and `Name` fields of `TargetRegisterDesc`) and the relationships of other registers to the defined register (in the other `TargetRegisterDesc` fields). In this example, other definitions establish the registers “AX”, “EAX”, and “RAX” as aliases for one another, so TableGen generates a null-terminated array (`AL_AliasSet`) for this register alias set.

The `Register` class is commonly used as a base class for more complex classes. In `Target.td`, the `Register`

class is the base for the `RegisterWithSubRegs` class that is used to define registers that need to specify subregisters in the `SubRegs` list, as shown here:

```
class RegisterWithSubRegs<string n, list<Register> subregs> : Register<n> {
    let SubRegs = subregs;
}
```

In `SparcRegisterInfo.td`, additional register classes are defined for SPARC: a `Register` subclass, `SparcReg`, and further subclasses: `Ri`, `Rf`, and `Rd`. SPARC registers are identified by 5-bit ID numbers, which is a feature common to these subclasses. Note the use of “let” expressions to override values that are initially defined in a superclass (such as `SubRegs` field in the `Rd` class).

```
class SparcReg<string n> : Register<n> {
    field bits<5> Num;
    let Namespace = "SP";
}
// Ri - 32-bit integer registers
class Ri<bits<5> num, string n> :
SparcReg<n> {
    let Num = num;
}
// Rf - 32-bit floating-point registers
class Rf<bits<5> num, string n> :
SparcReg<n> {
    let Num = num;
}
// Rd - Slots in the FP register file for 64-bit floating-point values.
class Rd<bits<5> num, string n, list<Register> subregs> : SparcReg<n> {
    let Num = num;
    let SubRegs = subregs;
}
```

In the `SparcRegisterInfo.td` file, there are register definitions that utilize these subclasses of `Register`, such as:

```
def G0 : Ri< 0, "G0">, DwarfRegNum<[0]>;
def G1 : Ri< 1, "G1">, DwarfRegNum<[1]>;
...
def F0 : Rf< 0, "F0">, DwarfRegNum<[32]>;
def F1 : Rf< 1, "F1">, DwarfRegNum<[33]>;
...
def D0 : Rd< 0, "F0", [F0, F1]>, DwarfRegNum<[32]>;
def D1 : Rd< 2, "F2", [F2, F3]>, DwarfRegNum<[34]>;
```

The last two registers shown above (`D0` and `D1`) are double-precision floating-point registers that are aliases for pairs of single-precision floating-point sub-registers. In addition to aliases, the sub-register and super-register relationships of the defined register are in fields of a register’s `TargetRegisterDesc`.

Defining a Register Class

The `RegisterClass` class (specified in `Target.td`) is used to define an object that represents a group of related registers and also defines the default allocation order of the registers. A target description file `XXXRegisterInfo.td` that uses `Target.td` can construct register classes using the following class:

```
class RegisterClass<string namespace,
list<ValueType> regTypes, int alignment, dag regList> {
    string Namespace = namespace;
```

```

list<ValueType> RegTypes = regTypes;
int Size = 0; // spill size, in bits; zero lets tblgen pick the size
int Alignment = alignment;

// CopyCost is the cost of copying a value between two registers
// default value 1 means a single instruction
// A negative value means copying is extremely expensive or impossible
int CopyCost = 1;
dag MemberList = regList;

// for register classes that are subregisters of this class
list<RegisterClass> SubRegClassList = [];

code MethodProtos = [{}]; // to insert arbitrary code
code MethodBodies = [{}];
}

```

To define a `RegisterClass`, use the following 4 arguments:

- The first argument of the definition is the name of the namespace.
- The second argument is a list of `ValueType` register type values that are defined in `include/llvm/CodeGen/ValueTypes.td`. Defined values include integer types (such as `i16`, `i32`, and `i1` for Boolean), floating-point types (`f32`, `f64`), and vector types (for example, `v8i16` for an $8 \times i16$ vector). All registers in a `RegisterClass` must have the same `ValueType`, but some registers may store vector data in different configurations. For example a register that can process a 128-bit vector may be able to handle 16 8-bit integer elements, 8 16-bit integers, 4 32-bit integers, and so on.
- The third argument of the `RegisterClass` definition specifies the alignment required of the registers when they are stored or loaded to memory.
- The final argument, `regList`, specifies which registers are in this class. If an alternative allocation order method is not specified, then `regList` also defines the order of allocation used by the register allocator. Besides simply listing registers with `(add R0, R1, ...)`, more advanced set operators are available. See `include/llvm/Target/Target.td` for more information.

In `SparcRegisterInfo.td`, three `RegisterClass` objects are defined: `FPPRegs`, `DFFPPRegs`, and `IntRegs`. For all three register classes, the first argument defines the namespace with the string “SP”. `FPPRegs` defines a group of 32 single-precision floating-point registers (F0 to F31); `DFFPPRegs` defines a group of 16 double-precision registers (D0–D15).

```

// F0, F1, F2, ..., F31
def FPPRegs : RegisterClass<"SP", [f32], 32, (sequence "F%u", 0, 31)>;

def DFFPPRegs : RegisterClass<"SP", [f64], 64,
    (add D0, D1, D2, D3, D4, D5, D6, D7, D8,
     D9, D10, D11, D12, D13, D14, D15)>;

def IntRegs : RegisterClass<"SP", [i32], 32,
    (add L0, L1, L2, L3, L4, L5, L6, L7,
     I0, I1, I2, I3, I4, I5,
     O0, O1, O2, O3, O4, O5, O7,
     G1,
     // Non-allocatable regs:
     G2, G3, G4,
     O6, // stack ptr
     I6, // frame ptr
     I7, // return address
     G0, // constant zero

```

```
G5, G6, G7 // reserved for kernel
);
```

Using `SparcRegisterInfo.td` with `TableGen` generates several output files that are intended for inclusion in other source code that you write. `SparcRegisterInfo.td` generates `SparcGenRegisterInfo.h.inc`, which should be included in the header file for the implementation of the SPARC register implementation that you write (`SparcRegisterInfo.h`). In `SparcGenRegisterInfo.h.inc` a new structure is defined called `SparcGenRegisterInfo` that uses `TargetRegisterInfo` as its base. It also specifies types, based upon the defined register classes: `DFPRegsClass`, `FPRegsClass`, and `IntRegsClass`.

`SparcRegisterInfo.td` also generates `SparcGenRegisterInfo.inc`, which is included at the bottom of `SparcRegisterInfo.cpp`, the SPARC register implementation. The code below shows only the generated integer registers and associated register classes. The order of registers in `IntRegs` reflects the order in the definition of `IntRegs` in the target description file.

```
// IntRegs Register Class...
static const unsigned IntRegs[] = {
    SP::L0, SP::L1, SP::L2, SP::L3, SP::L4, SP::L5,
    SP::L6, SP::L7, SP::I0, SP::I1, SP::I2, SP::I3,
    SP::I4, SP::I5, SP::O0, SP::O1, SP::O2, SP::O3,
    SP::O4, SP::O5, SP::O7, SP::G1, SP::G2, SP::G3,
    SP::G4, SP::O6, SP::I6, SP::I7, SP::G0, SP::G5,
    SP::G6, SP::G7,
};

// IntRegsVTs Register Class Value Types...
static const MVT::ValueType IntRegsVTs[] = {
    MVT::i32, MVT::Other
};

namespace SP { // Register class instances
    DFPRegsClass    DFPRegsRegClass;
    FPRegsClass     FPRegsRegClass;
    IntRegsClass     IntRegsRegClass;
...
    // IntRegs Sub-register Classes...
    static const TargetRegisterClass* const IntRegsSubRegClasses [] = {
        NULL
    };
...
    // IntRegs Super-register Classes..
    static const TargetRegisterClass* const IntRegsSuperRegClasses [] = {
        NULL
    };
...
    // IntRegs Register Class sub-classes...
    static const TargetRegisterClass* const IntRegsSubclasses [] = {
        NULL
    };
...
    // IntRegs Register Class super-classes...
    static const TargetRegisterClass* const IntRegsSuperclasses [] = {
        NULL
    };
};

IntRegsClass::IntRegsClass() : TargetRegisterClass(IntRegsRegClassID,
    IntRegsVTs, IntRegsSubclasses, IntRegsSuperclasses, IntRegsSubRegClasses,
    IntRegsSuperRegClasses, 4, 4, 1, IntRegs, IntRegs + 32) {}
```

```
}

```

The register allocators will avoid using reserved registers, and callee saved registers are not used until all the volatile registers have been used. That is usually good enough, but in some cases it may be necessary to provide custom allocation orders.

Implement a subclass of `TargetRegisterInfo`

The final step is to hand code portions of `XXXRegisterInfo`, which implements the interface described in `TargetRegisterInfo.h` (see *The `TargetRegisterInfo` class*). These functions return 0, `NULL`, or `false`, unless overridden. Here is a list of functions that are overridden for the SPARC implementation in `SparcRegisterInfo.cpp`:

- `getCalleeSavedRegs` — Returns a list of callee-saved registers in the order of the desired callee-save stack frame offset.
- `getReservedRegs` — Returns a bitset indexed by physical register numbers, indicating if a particular register is unavailable.
- `hasFP` — Return a Boolean indicating if a function should have a dedicated frame pointer register.
- `eliminateCallFramePseudoInstr` — If call frame setup or destroy pseudo instructions are used, this can be called to eliminate them.
- `eliminateFrameIndex` — Eliminate abstract frame indices from instructions that may use them.
- `emitPrologue` — Insert prologue code into the function.
- `emitEpilogue` — Insert epilogue code into the function.

4.19.6 Instruction Set

During the early stages of code generation, the LLVM IR code is converted to a `SelectionDAG` with nodes that are instances of the `SDNode` class containing target instructions. An `SDNode` has an opcode, operands, type requirements, and operation properties. For example, is an operation commutative, does an operation load from memory. The various operation node types are described in the `include/llvm/CodeGen/SelectionDAGNodes.h` file (values of the `NodeType` enum in the `ISD` namespace).

TableGen uses the following target description (`.td`) input files to generate much of the code for instruction definition:

- `Target.td` — Where the `Instruction`, `Operand`, `InstrInfo`, and other fundamental classes are defined.
- `TargetSelectionDAG.td` — Used by `SelectionDAG` instruction selection generators, contains `SDTC*` classes (selection DAG type constraint), definitions of `SelectionDAG` nodes (such as `imm`, `cond`, `bb`, `add`, `fadd`, `sub`), and pattern support (`Pattern`, `Pat`, `PatFrag`, `PatLeaf`, `ComplexPattern`).
- `XXXInstrFormats.td` — Patterns for definitions of target-specific instructions.
- `XXXInstrInfo.td` — Target-specific definitions of instruction templates, condition codes, and instructions of an instruction set. For architecture modifications, a different file name may be used. For example, for Pentium with SSE instruction, this file is `X86InstrSSE.td`, and for Pentium with MMX, this file is `X86InstrMMX.td`.

There is also a target-specific `XXX.td` file, where `XXX` is the name of the target. The `XXX.td` file includes the other `.td` input files, but its contents are only directly important for subtargets.

You should describe a concrete target-specific class `XXXInstrInfo` that represents machine instructions supported by a target machine. `XXXInstrInfo` contains an array of `XXXInstrDescriptor` objects, each of which describes one instruction. An instruction descriptor defines:

- Opcode mnemonic
- Number of operands
- List of implicit register definitions and uses
- Target-independent properties (such as memory access, is commutable)
- Target-specific flags

The `Instruction` class (defined in `Target.td`) is mostly used as a base for more complex instruction classes.

```
class Instruction {
  string Namespace = "";
  dag OutOperandList;    // A dag containing the MI def operand list.
  dag InOperandList;     // A dag containing the MI use operand list.
  string AsmString = ""; // The .s format to print the instruction with.
  list<dag> Pattern;     // Set to the DAG pattern for this instruction.
  list<Register> Uses = [];
  list<Register> Defs = [];
  list<Predicate> Predicates = []; // predicates turned into isel match code
  ... remainder not shown for space ...
}
```

A `SelectionDAG` node (`SDNode`) should contain an object representing a target-specific instruction that is defined in `XXXInstrInfo.td`. The instruction objects should represent instructions from the architecture manual of the target machine (such as the SPARC Architecture Manual for the SPARC target).

A single instruction from the architecture manual is often modeled as multiple target instructions, depending upon its operands. For example, a manual might describe an add instruction that takes a register or an immediate operand. An LLVM target could model this with two instructions named `ADDri` and `ADDrr`.

You should define a class for each instruction category and define each opcode as a subclass of the category with appropriate parameters such as the fixed binary encoding of opcodes and extended opcodes. You should map the register bits to the bits of the instruction in which they are encoded (for the JIT). Also you should specify how the instruction should be printed when the automatic assembly printer is used.

As is described in the SPARC Architecture Manual, Version 8, there are three major 32-bit formats for instructions. Format 1 is only for the `CALL` instruction. Format 2 is for branch on condition codes and `SETHI` (set high bits of a register) instructions. Format 3 is for other instructions.

Each of these formats has corresponding classes in `SparcInstrFormat.td`. `InstSP` is a base class for other instruction classes. Additional base classes are specified for more precise formats: for example in `SparcInstrFormat.td`, `F2_1` is for `SETHI`, and `F2_2` is for branches. There are three other base classes: `F3_1` for register/register operations, `F3_2` for register/immediate operations, and `F3_3` for floating-point operations. `SparcInstrInfo.td` also adds the base class `Pseudo` for synthetic SPARC instructions.

`SparcInstrInfo.td` largely consists of operand and instruction definitions for the SPARC target. In `SparcInstrInfo.td`, the following target description file entry, `LDrr`, defines the Load Integer instruction for a Word (the `LD` SPARC opcode) from a memory address to a register. The first parameter, the value 3 (11_2), is the operation value for this category of operation. The second parameter (000000_2) is the specific operation value for `LD/Load Word`. The third parameter is the output destination, which is a register operand and defined in the `Register` target description file (`IntRegs`).

```
def LDrr : F3_1 <3, 0b000000, (outs IntRegs:$dst), (ins MEMrr:$addr),
    "ld [$addr], $dst",
    [(set i32:$dst, (load ADDRrr:$addr))]>;
```


The fourth parameter is the input source, which uses the address operand `MEMrr` that is defined earlier in `SparcInstrInfo.td`:

```
def MEMrr : Operand<i32> {
  let PrintMethod = "printMemOperand";
  let MIOperandInfo = (ops IntRegs, IntRegs);
}
```

The fifth parameter is a string that is used by the assembly printer and can be left as an empty string until the assembly printer interface is implemented. The sixth and final parameter is the pattern used to match the instruction during the SelectionDAG Select Phase described in *The LLVM Target-Independent Code Generator*. This parameter is detailed in the next section, *Instruction Selector*.

Instruction class definitions are not overloaded for different operand types, so separate versions of instructions are needed for register, memory, or immediate value operands. For example, to perform a Load Integer instruction for a Word from an immediate operand to a register, the following instruction class is defined:

```
def LDri : F3_2 <3, 0b000000, (outs IntRegs:$dst), (ins MEMri:$addr),
  "ld [$addr], $dst",
  [(set i32:$dst, (load ADDRri:$addr))]>;
```

Writing these definitions for so many similar instructions can involve a lot of cut and paste. In `.td` files, the `multiclass` directive enables the creation of templates to define several instruction classes at once (using the `defm` directive). For example in `SparcInstrInfo.td`, the `multiclass` pattern `F3_12` is defined to create 2 instruction classes each time `F3_12` is invoked:

```
multiclass F3_12 <string OpcStr, bits<6> Op3Val, SDNode OpNode> {
  def rr : F3_1 <2, Op3Val,
    (outs IntRegs:$dst), (ins IntRegs:$b, IntRegs:$c),
    !strconcat(OpcStr, " $b, $c, $dst"),
    [(set i32:$dst, (OpNode i32:$b, i32:$c))]>;
  def ri : F3_2 <2, Op3Val,
    (outs IntRegs:$dst), (ins IntRegs:$b, i32imm:$c),
    !strconcat(OpcStr, " $b, $c, $dst"),
    [(set i32:$dst, (OpNode i32:$b, simm13:$c))]>;
}
```

So when the `defm` directive is used for the XOR and ADD instructions, as seen below, it creates four instruction objects: `XORrr`, `XORri`, `ADDrr`, and `ADDri`.

```
defm XOR : F3_12<"xor", 0b000011, xor>;
defm ADD : F3_12<"add", 0b000000, add>;
```

`SparcInstrInfo.td` also includes definitions for condition codes that are referenced by branch instructions. The following definitions in `SparcInstrInfo.td` indicate the bit location of the SPARC condition code. For example, the 10th bit represents the “greater than” condition for integers, and the 22nd bit represents the “greater than” condition for floats.

```
def ICC_NE : ICC_VAL< 9>; // Not Equal
def ICC_E : ICC_VAL< 1>; // Equal
def ICC_G : ICC_VAL<10>; // Greater
...
def FCC_U : FCC_VAL<23>; // Unordered
def FCC_G : FCC_VAL<22>; // Greater
def FCC_UG : FCC_VAL<21>; // Unordered or Greater
...
```

(Note that `Sparc.h` also defines enums that correspond to the same SPARC condition codes. Care must be taken to ensure the values in `Sparc.h` correspond to the values in `SparcInstrInfo.td`. I.e., `SPCC::ICC_NE = 9`, `SPCC::FCC_U = 23` and so on.)

Instruction Operand Mapping

The code generator backend maps instruction operands to fields in the instruction. Operands are assigned to unbound fields in the instruction in the order they are defined. Fields are bound when they are assigned a value. For example, the Sparc target defines the `XNORrr` instruction as a `F3_1` format instruction having three operands.

```
def XNORrr : F3_1<2, 0b000111,  
    (outs IntRegs:$dst), (ins IntRegs:$b, IntRegs:$c),  
    "xnor $b, $c, $dst",  
    [(set i32:$dst, (not (xor i32:$b, i32:$c)))]>;
```

The instruction templates in `SparcInstrFormats.td` show the base class for `F3_1` is `InstSP`.

```
class InstSP<dag outs, dag ins, string asmstr, list<dag> pattern> : Instruction {  
  field bits<32> Inst;  
  let Namespace = "SP";  
  bits<2> op;  
  let Inst{31-30} = op;  
  dag OutOperandList = outs;  
  dag InOperandList = ins;  
  let AsmString = asmstr;  
  let Pattern = pattern;  
}
```

`InstSP` leaves the `op` field unbound.

```
class F3<dag outs, dag ins, string asmstr, list<dag> pattern>  
  : InstSP<outs, ins, asmstr, pattern> {  
  bits<5> rd;  
  bits<6> op3;  
  bits<5> rs1;  
  let op{1} = 1; // Op = 2 or 3  
  let Inst{29-25} = rd;  
  let Inst{24-19} = op3;  
  let Inst{18-14} = rs1;  
}
```

`F3` binds the `op` field and defines the `rd`, `op3`, and `rs1` fields. `F3` format instructions will bind the operands `rd`, `op3`, and `rs1` fields.

```
class F3_1<bits<2> opVal, bits<6> op3val, dag outs, dag ins,  
    string asmstr, list<dag> pattern> : F3<outs, ins, asmstr, pattern> {  
  bits<8> asi = 0; // asi not currently used  
  bits<5> rs2;  
  let op = opVal;  
  let op3 = op3val;  
  let Inst{13} = 0; // i field = 0  
  let Inst{12-5} = asi; // address space identifier  
  let Inst{4-0} = rs2;  
}
```

`F3_1` binds the `op3` field and defines the `rs2` fields. `F3_1` format instructions will bind the operands to the `rd`, `rs1`, and `rs2` fields. This results in the `XNORrr` instruction binding `$dst`, `$b`, and `$c` operands to the `rd`, `rs1`, and `rs2`

fields respectively.

Instruction Operand Name Mapping

TableGen will also generate a function called `getNamedOperandIdx()` which can be used to look up an operand's index in a `MachineInstr` based on its TableGen name. Setting the `UseNamedOperandTable` bit in an instruction's TableGen definition will add all of its operands to an enumeration in the `llvm::XXX::OpName` namespace and also add an entry for it into the `OperandMap` table, which can be queried using `getNamedOperandIdx()`

```
int DstIndex = SP::getNamedOperandIdx(SP::XNORrr, SP::OpName::dst); // => 0
int BIndex = SP::getNamedOperandIdx(SP::XNORrr, SP::OpName::b);    // => 1
int CIndex = SP::getNamedOperandIdx(SP::XNORrr, SP::OpName::c);    // => 2
int DIndex = SP::getNamedOperandIdx(SP::XNORrr, SP::OpName::d);    // => -1

...
```

The entries in the `OpName` enum are taken verbatim from the TableGen definitions, so operands with lowercase names will have lower case entries in the enum.

To include the `getNamedOperandIdx()` function in your backend, you will need to define a few preprocessor macros in `XXXInstrInfo.cpp` and `XXXInstrInfo.h`. For example:

`XXXInstrInfo.cpp`:

```
#define GET_INSTRINFO_NAMED_OPS // For getNamedOperandIdx() function
#include "XXXGenInstrInfo.inc"
```

`XXXInstrInfo.h`:

```
#define GET_INSTRINFO_OPERAND_ENUM // For OpName enum
#include "XXXGenInstrInfo.inc"

namespace XXX {
    int16_t getNamedOperandIdx(uint16_t Opcode, uint16_t NamedIndex);
} // End namespace XXX
```

Instruction Operand Types

TableGen will also generate an enumeration consisting of all named Operand types defined in the backend, in the `llvm::XXX::OpTypes` namespace. Some common immediate Operand types (for instance `i8`, `i32`, `i64`, `f32`, `f64`) are defined for all targets in `include/llvm/Target/Target.td`, and are available in each Target's `OpTypes` enum. Also, only named Operand types appear in the enumeration: anonymous types are ignored. For example, the X86 backend defines `brtarget` and `brtarget8`, both instances of the TableGen `Operand` class, which represent branch target operands:

```
def brtarget : Operand<OtherVT>;
def brtarget8 : Operand<OtherVT>;
```

This results in:

```
namespace X86 {
namespace OpTypes {
enum OperandType {
    ...
    brtarget,
```

```
brtarget8,  
...  
i32imm,  
i64imm,  
...  
OPERAND_TYPE_LIST_END  
} // End namespace OpTypes  
} // End namespace X86
```

In typical TableGen fashion, to use the enum, you will need to define a preprocessor macro:

```
#define GET_INSTRINFO_OPERAND_TYPES_ENUM // For OpTypes enum  
#include "XXXGenInstrInfo.inc"
```

Instruction Scheduling

Instruction itineraries can be queried using `MCDesc::getSchedClass()`. The value can be named by an enumeration in `llvm::XXX::Sched` namespace generated by TableGen in `XXXGenInstrInfo.inc`. The name of the schedule classes are the same as provided in `XXXSchedule.td` plus a default `NoItinerary` class.

Instruction Relation Mapping

This TableGen feature is used to relate instructions with each other. It is particularly useful when you have multiple instruction formats and need to switch between them after instruction selection. This entire feature is driven by relation models which can be defined in `XXXInstrInfo.td` files according to the target-specific instruction set. Relation models are defined using `InstrMapping` class as a base. TableGen parses all the models and generates instruction relation maps using the specified information. Relation maps are emitted as tables in the `XXXGenInstrInfo.inc` file along with the functions to query them. For the detailed information on how to use this feature, please refer to *How To Use Instruction Mappings*.

Implement a subclass of `TargetInstrInfo`

The final step is to hand code portions of `XXXInstrInfo`, which implements the interface described in `TargetInstrInfo.h` (see *The TargetInstrInfo class*). These functions return 0 or a Boolean or they assert, unless overridden. Here's a list of functions that are overridden for the SPARC implementation in `SparcInstrInfo.cpp`:

- `isLoadFromStackSlot` — If the specified machine instruction is a direct load from a stack slot, return the register number of the destination and the `FrameIndex` of the stack slot.
- `isStoreToStackSlot` — If the specified machine instruction is a direct store to a stack slot, return the register number of the destination and the `FrameIndex` of the stack slot.
- `copyPhysReg` — Copy values between a pair of physical registers.
- `storeRegToStackSlot` — Store a register value to a stack slot.
- `loadRegFromStackSlot` — Load a register value from a stack slot.
- `storeRegToAddr` — Store a register value to memory.
- `loadRegFromAddr` — Load a register value from memory.
- `foldMemoryOperand` — Attempt to combine instructions of any load or store instruction for the specified operand(s).

Branch Folding and If Conversion

Performance can be improved by combining instructions or by eliminating instructions that are never reached. The `AnalyzeBranch` method in `XXXInstrInfo` may be implemented to examine conditional instructions and remove unnecessary instructions. `AnalyzeBranch` looks at the end of a machine basic block (MBB) for opportunities for improvement, such as branch folding and if conversion. The `BranchFolder` and `IfConverter` machine function passes (see the source files `BranchFolding.cpp` and `IfConversion.cpp` in the `lib/CodeGen` directory) call `AnalyzeBranch` to improve the control flow graph that represents the instructions.

Several implementations of `AnalyzeBranch` (for ARM, Alpha, and X86) can be examined as models for your own `AnalyzeBranch` implementation. Since SPARC does not implement a useful `AnalyzeBranch`, the ARM target implementation is shown below.

`AnalyzeBranch` returns a Boolean value and takes four parameters:

- `MachineBasicBlock &MBB` — The incoming block to be examined.
- `MachineBasicBlock *&TBB` — A destination block that is returned. For a conditional branch that evaluates to true, TBB is the destination.
- `MachineBasicBlock *&FBB` — For a conditional branch that evaluates to false, FBB is returned as the destination.
- `std::vector<MachineOperand> &Cond` — List of operands to evaluate a condition for a conditional branch.

In the simplest case, if a block ends without a branch, then it falls through to the successor block. No destination blocks are specified for either TBB or FBB, so both parameters return `NULL`. The start of the `AnalyzeBranch` (see code below for the ARM target) shows the function parameters and the code for the simplest case.

```
bool ARMInstrInfo::AnalyzeBranch(MachineBasicBlock &MBB,
                                MachineBasicBlock *&TBB,
                                MachineBasicBlock *&FBB,
                                std::vector<MachineOperand> &Cond) const
{
    MachineBasicBlock::iterator I = MBB.end();
    if (I == MBB.begin() || !isUnpredicatedTerminator(--I))
        return false;
```

If a block ends with a single unconditional branch instruction, then `AnalyzeBranch` (shown below) should return the destination of that branch in the TBB parameter.

```
if (LastOpc == ARM::B || LastOpc == ARM::tB) {
    TBB = LastInst->getOperand(0).getMBB();
    return false;
}
```

If a block ends with two unconditional branches, then the second branch is never reached. In that situation, as shown below, remove the last branch instruction and return the penultimate branch in the TBB parameter.

```
if ((SecondLastOpc == ARM::B || SecondLastOpc == ARM::tB) &&
    (LastOpc == ARM::B || LastOpc == ARM::tB)) {
    TBB = SecondLastInst->getOperand(0).getMBB();
    I = LastInst;
    I->eraseFromParent();
    return false;
}
```

A block may end with a single conditional branch instruction that falls through to successor block if the condition evaluates to false. In that case, `AnalyzeBranch` (shown below) should return the destination of that conditional branch in the `TBB` parameter and a list of operands in the `Cond` parameter to evaluate the condition.

```
if (LastOpc == ARM::Bcc || LastOpc == ARM::tBcc) {
    // Block ends with fall-through condbranch.
    TBB = LastInst->getOperand(0).getMBB();
    Cond.push_back(LastInst->getOperand(1));
    Cond.push_back(LastInst->getOperand(2));
    return false;
}
```

If a block ends with both a conditional branch and an ensuing unconditional branch, then `AnalyzeBranch` (shown below) should return the conditional branch destination (assuming it corresponds to a conditional evaluation of “true”) in the `TBB` parameter and the unconditional branch destination in the `FBB` (corresponding to a conditional evaluation of “false”). A list of operands to evaluate the condition should be returned in the `Cond` parameter.

```
unsigned SecondLastOpc = SecondLastInst->getOpcode();

if ((SecondLastOpc == ARM::Bcc && LastOpc == ARM::B) ||
    (SecondLastOpc == ARM::tBcc && LastOpc == ARM::tB)) {
    TBB = SecondLastInst->getOperand(0).getMBB();
    Cond.push_back(SecondLastInst->getOperand(1));
    Cond.push_back(SecondLastInst->getOperand(2));
    FBB = LastInst->getOperand(0).getMBB();
    return false;
}
```

For the last two cases (ending with a single conditional branch or ending with one conditional and one unconditional branch), the operands returned in the `Cond` parameter can be passed to methods of other instructions to create new branches or perform other operations. An implementation of `AnalyzeBranch` requires the helper methods `RemoveBranch` and `InsertBranch` to manage subsequent operations.

`AnalyzeBranch` should return false indicating success in most circumstances. `AnalyzeBranch` should only return true when the method is stumped about what to do, for example, if a block has three terminating branches. `AnalyzeBranch` may return true if it encounters a terminator it cannot handle, such as an indirect branch.

4.19.7 Instruction Selector

LLVM uses a `SelectionDAG` to represent LLVM IR instructions, and nodes of the `SelectionDAG` ideally represent native target instructions. During code generation, instruction selection passes are performed to convert non-native DAG instructions into native target-specific instructions. The pass described in `XXXISelDAGToDAG.cpp` is used to match patterns and perform DAG-to-DAG instruction selection. Optionally, a pass may be defined (in `XXXBranchSelector.cpp`) to perform similar DAG-to-DAG operations for branch instructions. Later, the code in `XXXISelLowering.cpp` replaces or removes operations and data types not supported natively (legalizes) in a `SelectionDAG`.

TableGen generates code for instruction selection using the following target description input files:

- `XXXInstrInfo.td` — Contains definitions of instructions in a target-specific instruction set, generates `XXXGenDAGISel.inc`, which is included in `XXXISelDAGToDAG.cpp`.
- `XXXCallingConv.td` — Contains the calling and return value conventions for the target architecture, and it generates `XXXGenCallingConv.inc`, which is included in `XXXISelLowering.cpp`.

The implementation of an instruction selection pass must include a header that declares the `FunctionPass` class or a subclass of `FunctionPass`. In `XXXTargetMachine.cpp`, a Pass Manager (PM) should add each instruction selection pass into the queue of passes to run.

The LLVM static compiler (`llc`) is an excellent tool for visualizing the contents of DAGs. To display the SelectionDAG before or after specific processing phases, use the command line options for `llc`, described at [SelectionDAG Instruction Selection Process](#).

To describe instruction selector behavior, you should add patterns for lowering LLVM code into a SelectionDAG as the last parameter of the instruction definitions in `XXXInstrInfo.td`. For example, in `SparcInstrInfo.td`, this entry defines a register store operation, and the last parameter describes a pattern with the store DAG operator.

```
def STrr : F3_1< 3, 0b000100, (outs), (ins MEMrr:$addr, IntRegs:$src),
    "st $src, [$addr]", [(store i32:$src, ADDRrr:$addr)]>;
```

`ADDRrr` is a memory mode that is also defined in `SparcInstrInfo.td`:

```
def ADDRrr : ComplexPattern<i32, 2, "SelectADDRrr", [], []>;
```

The definition of `ADDRrr` refers to `SelectADDRrr`, which is a function defined in an implementation of the Instruction Selector (such as `SparcISelDAGToDAG.cpp`).

In `lib/Target/TargetSelectionDAG.td`, the DAG operator for store is defined below:

```
def store : PatFrag<(ops node:$val, node:$ptr),
    (st node:$val, node:$ptr), [{
  if (StoreSDNode *ST = dyn_cast<StoreSDNode>(N))
    return !ST->isTruncatingStore() &&
        ST->getAddressingMode() == ISD::UNINDEXED;
  return false;
}]>;
```

`XXXInstrInfo.td` also generates (in `XXXGenDAGISel.inc`) the `SelectCode` method that is used to call the appropriate processing method for an instruction. In this example, `SelectCode` calls `Select_ISD_STORE` for the `ISD::STORE` opcode.

```
SDNode *SelectCode(SDValue N) {
  ...
  MVT::ValueType NVT = N.getNode()->getValueType(0);
  switch (N.getOpcode()) {
  case ISD::STORE: {
    switch (NVT) {
    default:
      return Select_ISD_STORE(N);
    break;
    }
    break;
  }
  ...
}
```

The pattern for `STrr` is matched, so elsewhere in `XXXGenDAGISel.inc`, code for `STrr` is created for `Select_ISD_STORE`. The `Emit_22` method is also generated in `XXXGenDAGISel.inc` to complete the processing of this instruction.

```
SDNode *Select_ISD_STORE(const SDValue &N) {
  SDValue Chain = N.getOperand(0);
  if (Predicate_store(N.getNode())) {
    SDValue N1 = N.getOperand(1);
    SDValue N2 = N.getOperand(2);
    SDValue CPTmp0;
    SDValue CPTmp1;
```

```
// Pattern: (st:void i32:i32:$src,
//          ADDRrr:i32:$addr)<<P:Predicate_store>>
// Emits: (STrr:void ADDRrr:i32:$addr, IntRegs:i32:$src)
// Pattern complexity = 13  cost = 1  size = 0
if (SelectADDRrr(N, N2, CPTmp0, CPTmp1) &&
    N1.getNode()->getValueType(0) == MVT::i32 &&
    N2.getNode()->getValueType(0) == MVT::i32) {
    return Emit_22(N, SP::STrr, CPTmp0, CPTmp1);
}
...
```

The SelectionDAG Legalize Phase

The Legalize phase converts a DAG to use types and operations that are natively supported by the target. For natively unsupported types and operations, you need to add code to the target-specific `XXXTargetLowering` implementation to convert unsupported types and operations to supported ones.

In the constructor for the `XXXTargetLowering` class, first use the `addRegisterClass` method to specify which types are supported and which register classes are associated with them. The code for the register classes are generated by TableGen from `XXXRegisterInfo.td` and placed in `XXXGenRegisterInfo.h.inc`. For example, the implementation of the constructor for the `SparcISelLowering` class (in `SparcISelLowering.cpp`) starts with the following code:

```
addRegisterClass(MVT::i32, SP::IntRegsRegisterClass);
addRegisterClass(MVT::f32, SP::FPRegsRegisterClass);
addRegisterClass(MVT::f64, SP::DFPRegsRegisterClass);
```

You should examine the node types in the ISD namespace (include/llvm/CodeGen/SelectionDAGNodes.h) and determine which operations the target natively supports. For operations that do **not** have native support, add a callback to the constructor for the `XXXTargetLowering` class, so the instruction selection process knows what to do. The `TargetLowering` class callback methods (declared in `llvm/Target/TargetLowering.h`) are:

- `setOperationAction` — General operation.
- `setLoadExtAction` — Load with extension.
- `setTruncStoreAction` — Truncating store.
- `setIndexedLoadAction` — Indexed load.
- `setIndexedStoreAction` — Indexed store.
- `setConvertAction` — Type conversion.
- `setCondCodeAction` — Support for a given condition code.

Note: on older releases, `setLoadXAction` is used instead of `setLoadExtAction`. Also, on older releases, `setCondCodeAction` may not be supported. Examine your release to see what methods are specifically supported.

These callbacks are used to determine that an operation does or does not work with a specified type (or types). And in all cases, the third parameter is a `LegalAction` type enum value: `Promote`, `Expand`, `Custom`, or `Legal`. `SparcISelLowering.cpp` contains examples of all four `LegalAction` values.

Promote

For an operation without native support for a given type, the specified type may be promoted to a larger type that is supported. For example, SPARC does not support a sign-extending load for Boolean values (`i1` type), so in

`SparcISelLowering.cpp` the third parameter below, `Promote`, changes `i1` type values to a large type before loading.

```
setLoadExtAction(ISD::SEXTLOAD, MVT::i1, Promote);
```

Expand

For a type without native support, a value may need to be broken down further, rather than promoted. For an operation without native support, a combination of other operations may be used to similar effect. In SPARC, the floating-point sine and cosine trig operations are supported by expansion to other operations, as indicated by the third parameter, `Expand`, to `setOperationAction`:

```
setOperationAction(ISD::FSIN, MVT::f32, Expand);
setOperationAction(ISD::FCOS, MVT::f32, Expand);
```

Custom

For some operations, simple type promotion or operation expansion may be insufficient. In some cases, a special intrinsic function must be implemented.

For example, a constant value may require special treatment, or an operation may require spilling and restoring registers in the stack and working with register allocators.

As seen in `SparcISelLowering.cpp` code below, to perform a type conversion from a floating point value to a signed integer, first the `setOperationAction` should be called with `Custom` as the third parameter:

```
setOperationAction(ISD::FP_TO_SINT, MVT::i32, Custom);
```

In the `LowerOperation` method, for each `Custom` operation, a case statement should be added to indicate what function to call. In the following code, an `FP_TO_SINT` opcode will call the `LowerFP_TO_SINT` method:

```
SDValue SparcTargetLowering::LowerOperation(SDValue Op, SelectionDAG &DAG) {
    switch (Op.getOpcode()) {
        case ISD::FP_TO_SINT: return LowerFP_TO_SINT(Op, DAG);
        ...
    }
}
```

Finally, the `LowerFP_TO_SINT` method is implemented, using an FP register to convert the floating-point value to an integer.

```
static SDValue LowerFP_TO_SINT(SDValue Op, SelectionDAG &DAG) {
    assert(Op.getValueType() == MVT::i32);
    Op = DAG.getNode(SPISD::FTOI, MVT::f32, Op.getOperand(0));
    return DAG.getNode(ISD::BITCAST, MVT::i32, Op);
}
```

Legal

The `Legal` `LegalizeAction` enum value simply indicates that an operation is natively supported. `Legal` represents the default condition, so it is rarely used. In `SparcISelLowering.cpp`, the action for `CTPOP` (an operation to count the bits set in an integer) is natively supported only for SPARC v9. The following code enables the `Expand` conversion technique for non-v9 SPARC implementations.

```
setOperationAction(ISD::CTPOP, MVT::i32, Expand);
...
if (TM.getSubtarget<SparcSubtarget>().isV9())
    setOperationAction(ISD::CTPOP, MVT::i32, Legal);
```

Calling Conventions

To support target-specific calling conventions, `XXXGenCallingConv.td` uses interfaces (such as `CCIfType` and `CCAssignToReg`) that are defined in `lib/Target/TargetCallingConv.td`. `TableGen` can take the target descriptor file `XXXGenCallingConv.td` and generate the header file `XXXGenCallingConv.inc`, which is typically included in `XXXISelLowering.cpp`. You can use the interfaces in `TargetCallingConv.td` to specify:

- The order of parameter allocation.
- Where parameters and return values are placed (that is, on the stack or in registers).
- Which registers may be used.
- Whether the caller or callee unwinds the stack.

The following example demonstrates the use of the `CCIfType` and `CCAssignToReg` interfaces. If the `CCIfType` predicate is true (that is, if the current argument is of type `f32` or `f64`), then the action is performed. In this case, the `CCAssignToReg` action assigns the argument value to the first available register: either `R0` or `R1`.

```
CCIfType<[f32, f64], CCAssignToReg<[R0, R1]>>
```

`SparcCallingConv.td` contains definitions for a target-specific return-value calling convention (`RetCC_Sparc32`) and a basic 32-bit C calling convention (`CC_Sparc32`). The definition of `RetCC_Sparc32` (shown below) indicates which registers are used for specified scalar return types. A single-precision float is returned to register `F0`, and a double-precision float goes to register `D0`. A 32-bit integer is returned in register `I0` or `I1`.

```
def RetCC_Sparc32 : CallingConv<[
    CCIfType<[i32], CCAssignToReg<[I0, I1]>>,
    CCIfType<[f32], CCAssignToReg<[F0]>>,
    CCIfType<[f64], CCAssignToReg<[D0]>>
]>;
```

The definition of `CC_Sparc32` in `SparcCallingConv.td` introduces `CCAssignToStack`, which assigns the value to a stack slot with the specified size and alignment. In the example below, the first parameter, 4, indicates the size of the slot, and the second parameter, also 4, indicates the stack alignment along 4-byte units. (Special cases: if size is zero, then the ABI size is used; if alignment is zero, then the ABI alignment is used.)

```
def CC_Sparc32 : CallingConv<[
    // All arguments get passed in integer registers if there is space.
    CCIfType<[i32, f32, f64], CCAssignToReg<[I0, I1, I2, I3, I4, I5]>>,
    CCAssignToStack<4, 4>
]>;
```

`CCDelegateTo` is another commonly used interface, which tries to find a specified sub-calling convention, and, if a match is found, it is invoked. In the following example (in `X86CallingConv.td`), the definition of `RetCC_X86_32_C` ends with `CCDelegateTo`. After the current value is assigned to the register `ST0` or `ST1`, the `RetCC_X86Common` is invoked.

```
def RetCC_X86_32_C : CallingConv<[
    CCIfType<[f32], CCAssignToReg<[ST0, ST1]>>,
    CCIfType<[f64], CCAssignToReg<[ST0, ST1]>>,
```

```
CCDelegateTo<RetCC_X86Common>
]>;
```

CCIfCC is an interface that attempts to match the given name to the current calling convention. If the name identifies the current calling convention, then a specified action is invoked. In the following example (in `X86CallingConv.td`), if the `Fast` calling convention is in use, then `RetCC_X86_32_Fast` is invoked. If the `SSECall` calling convention is in use, then `RetCC_X86_32_SSE` is invoked.

```
def RetCC_X86_32 : CallingConv<[
  CCIfCC<"CallingConv::Fast", CCDelegateTo<RetCC_X86_32_Fast>>,
  CCIfCC<"CallingConv::X86_SSECall", CCDelegateTo<RetCC_X86_32_SSE>>,
  CCDelegateTo<RetCC_X86_32_C>
]>;
```

Other calling convention interfaces include:

- `CCIf <predicate, action>` — If the predicate matches, apply the action.
- `CCIfInReg <action>` — If the argument is marked with the “inreg” attribute, then apply the action.
- `CCIfNest <action>` — If the argument is marked with the “nest” attribute, then apply the action.
- `CCIfNotVarArg <action>` — If the current function does not take a variable number of arguments, apply the action.
- `CCAssignToRegWithShadow <registerList, shadowList>` — similar to `CCAssignToReg`, but with a shadow list of registers.
- `CCPassByVal <size, align>` — Assign value to a stack slot with the minimum specified size and alignment.
- `CCPromoteToType <type>` — Promote the current value to the specified type.
- `CallingConv <[actions]>` — Define each calling convention that is supported.

4.19.8 Assembly Printer

During the code emission stage, the code generator may utilize an LLVM pass to produce assembly output. To do this, you want to implement the code for a printer that converts LLVM IR to a GAS-format assembly language for your target machine, using the following steps:

- Define all the assembly strings for your target, adding them to the instructions defined in the `XXXInstrInfo.td` file. (See *Instruction Set*.) TableGen will produce an output file (`XXXGenAsmWriter.inc`) with an implementation of the `printInstruction` method for the `XXXAsmPrinter` class.
- Write `XXXTargetAsmInfo.h`, which contains the bare-bones declaration of the `XXXTargetAsmInfo` class (a subclass of `TargetAsmInfo`).
- Write `XXXTargetAsmInfo.cpp`, which contains target-specific values for `TargetAsmInfo` properties and sometimes new implementations for methods.
- Write `XXXAsmPrinter.cpp`, which implements the `AsmPrinter` class that performs the LLVM-to-assembly conversion.

The code in `XXXTargetAsmInfo.h` is usually a trivial declaration of the `XXXTargetAsmInfo` class for use in `XXXTargetAsmInfo.cpp`. Similarly, `XXXTargetAsmInfo.cpp` usually has a few declarations of `XXXTargetAsmInfo` replacement values that override the default values in `TargetAsmInfo.cpp`. For example in `SparcTargetAsmInfo.cpp`:

```
SparcTargetAsmInfo::SparcTargetAsmInfo(const SparcTargetMachine &TM) {
    Data16bitsDirective = "\t.half\t";
    Data32bitsDirective = "\t.word\t";
    Data64bitsDirective = 0; // .xword is only supported by V9.
    ZeroDirective = "\t.skip\t";
    CommentString = "!";
    ConstantPoolSection = "\t.section \".rodata\",#alloc\n";
}
```

The X86 assembly printer implementation (X86TargetAsmInfo) is an example where the target specific TargetAsmInfo class uses an overridden methods: ExpandInlineAsm.

A target-specific implementation of AsmPrinter is written in XXXAsmPrinter.cpp, which implements the AsmPrinter class that converts the LLVM to printable assembly. The implementation must include the following headers that have declarations for the AsmPrinter and MachineFunctionPass classes. The MachineFunctionPass is a subclass of FunctionPass.

```
#include "llvm/CodeGen/AsmPrinter.h"
#include "llvm/CodeGen/MachineFunctionPass.h"
```

As a FunctionPass, AsmPrinter first calls doInitialization to set up the AsmPrinter. In SparcAsmPrinter, a Mangler object is instantiated to process variable names.

In XXXAsmPrinter.cpp, the runOnMachineFunction method (declared in MachineFunctionPass) must be implemented for XXXAsmPrinter. In MachineFunctionPass, the runOnFunction method invokes runOnMachineFunction. Target-specific implementations of runOnMachineFunction differ, but generally do the following to process each machine function:

- Call SetupMachineFunction to perform initialization.
- Call EmitConstantPool to print out (to the output stream) constants which have been spilled to memory.
- Call EmitJumpTableInfo to print out jump tables used by the current function.
- Print out the label for the current function.
- Print out the code for the function, including basic block labels and the assembly for the instruction (using printInstruction)

The XXXAsmPrinter implementation must also include the code generated by TableGen that is output in the XXXGenAsmWriter.inc file. The code in XXXGenAsmWriter.inc contains an implementation of the printInstruction method that may call these methods:

- printOperand
- printMemOperand
- printCCOperand (for conditional statements)
- printDataDirective
- printDeclare
- printImplicitDef
- printInlineAsm

The implementations of printDeclare, printImplicitDef, printInlineAsm, and printLabel in AsmPrinter.cpp are generally adequate for printing assembly and do not need to be overridden.

The printOperand method is implemented with a long switch/case statement for the type of operand: register, immediate, basic block, external symbol, global address, constant pool index, or jump table index. For an instruction

with a memory address operand, the `printMemOperand` method should be implemented to generate the proper output. Similarly, `printCCOperand` should be used to print a conditional operand.

`doFinalization` should be overridden in `XXXAsmPrinter`, and it should be called to shut down the assembly printer. During `doFinalization`, global variables and constants are printed to output.

4.19.9 Subtarget Support

Subtarget support is used to inform the code generation process of instruction set variations for a given chip set. For example, the LLVM SPARC implementation provided covers three major versions of the SPARC microprocessor architecture: Version 8 (V8, which is a 32-bit architecture), Version 9 (V9, a 64-bit architecture), and the UltraSPARC architecture. V8 has 16 double-precision floating-point registers that are also usable as either 32 single-precision or 8 quad-precision registers. V8 is also purely big-endian. V9 has 32 double-precision floating-point registers that are also usable as 16 quad-precision registers, but cannot be used as single-precision registers. The UltraSPARC architecture combines V9 with UltraSPARC Visual Instruction Set extensions.

If subtarget support is needed, you should implement a target-specific `XXXSubtarget` class for your architecture. This class should process the command-line options `-mcpu=` and `-mattr=`.

TableGen uses definitions in the `Target.td` and `Sparc.td` files to generate code in `SparcGenSubtarget.inc`. In `Target.td`, shown below, the `SubtargetFeature` interface is defined. The first 4 string parameters of the `SubtargetFeature` interface are a feature name, an attribute set by the feature, the value of the attribute, and a description of the feature. (The fifth parameter is a list of features whose presence is implied, and its default value is an empty array.)

```
class SubtargetFeature<string n, string a, string v, string d,
                      list<SubtargetFeature> i = []> {
    string Name = n;
    string Attribute = a;
    string Value = v;
    string Desc = d;
    list<SubtargetFeature> Implies = i;
}
```

In the `Sparc.td` file, the `SubtargetFeature` is used to define the following features.

```
def FeatureV9 : SubtargetFeature<"v9", "IsV9", "true",
                                "Enable SPARC-V9 instructions">;
def FeatureV8Deprecated : SubtargetFeature<"deprecated-v8",
                                            "V8DeprecatedInsts", "true",
                                            "Enable deprecated V8 instructions in V9 mode">;
def FeatureVIS : SubtargetFeature<"vis", "IsVIS", "true",
                                "Enable UltraSPARC Visual Instruction Set extensions">;
```

Elsewhere in `Sparc.td`, the `Proc` class is defined and then is used to define particular SPARC processor subtypes that may have the previously described features.

```
class Proc<string Name, list<SubtargetFeature> Features>
  : Processor<Name, NoItineraries, Features>;

def : Proc<"generic",          []>;
def : Proc<"v8",               []>;
def : Proc<"supersparc",       []>;
def : Proc<"sparclite",        []>;
def : Proc<"f934",             []>;
def : Proc<"hypersparc",       []>;
def : Proc<"sparclite86x",     []>;
```

```
def : Proc<"sparclet",      []>;
def : Proc<"tsc701",      []>;
def : Proc<"v9",          [FeatureV9]>;
def : Proc<"ultrasparc",   [FeatureV9, FeatureV8Deprecated]>;
def : Proc<"ultrasparc3",  [FeatureV9, FeatureV8Deprecated]>;
def : Proc<"ultrasparc3-vis", [FeatureV9, FeatureV8Deprecated, FeatureVIS]>;
```

From `Target.td` and `Sparc.td` files, the resulting `SparcGenSubtarget.inc` specifies enum values to identify the features, arrays of constants to represent the CPU features and CPU subtypes, and the `ParseSubtargetFeatures` method that parses the features string that sets specified subtarget options. The generated `SparcGenSubtarget.inc` file should be included in the `SparcSubtarget.cpp`. The target-specific implementation of the `XXXSubtarget` method should follow this pseudocode:

```
XXXSubtarget::XXXSubtarget(const Module &M, const std::string &FS) {
    // Set the default features
    // Determine default and user specified characteristics of the CPU
    // Call ParseSubtargetFeatures(FS, CPU) to parse the features string
    // Perform any additional operations
}
```

4.19.10 JIT Support

The implementation of a target machine optionally includes a Just-In-Time (JIT) code generator that emits machine code and auxiliary structures as binary output that can be written directly to memory. To do this, implement JIT code generation by performing the following steps:

- Write an `XXXCodeEmitter.cpp` file that contains a machine function pass that transforms target-machine instructions into relocatable machine code.
- Write an `XXXJITInfo.cpp` file that implements the JIT interfaces for target-specific code-generation activities, such as emitting machine code and stubs.
- Modify `XXXTargetMachine` so that it provides a `TargetJITInfo` object through its `getJITInfo` method.

There are several different approaches to writing the JIT support code. For instance, `TableGen` and target descriptor files may be used for creating a JIT code generator, but are not mandatory. For the Alpha and PowerPC target machines, `TableGen` is used to generate `XXXGenCodeEmitter.inc`, which contains the binary coding of machine instructions and the `getBinaryCodeForInstr` method to access those codes. Other JIT implementations do not.

Both `XXXJITInfo.cpp` and `XXXCodeEmitter.cpp` must include the `llvm/CodeGen/MachineCodeEmitter.h` header file that defines the `MachineCodeEmitter` class containing code for several callback functions that write data (in bytes, words, strings, etc.) to the output stream.

Machine Code Emitter

In `XXXCodeEmitter.cpp`, a target-specific of the `Emitter` class is implemented as a function pass (subclass of `MachineFunctionPass`). The target-specific implementation of `runOnMachineFunction` (invoked by `runOnFunction` in `MachineFunctionPass`) iterates through the `MachineBasicBlock` calls `emitInstruction` to process each instruction and emit binary code. `emitInstruction` is largely implemented with case statements on the instruction types defined in `XXXInstrInfo.h`. For example, in `X86CodeEmitter.cpp`, the `emitInstruction` method is built around the following switch/case statements:

```
switch (Desc->TSFlags & X86::FormMask) {
case X86II::Pseudo: // for not yet implemented instructions
```

```

...                // or pseudo-instructions
break;
case X86II::RawFrm: // for instructions with a fixed opcode value
...
break;
case X86II::AddRegFrm: // for instructions that have one register operand
...                // added to their opcode
break;
case X86II::MRMDestReg: // for instructions that use the Mod/RM byte
...                // to specify a destination (register)
break;
case X86II::MRMDestMem: // for instructions that use the Mod/RM byte
...                // to specify a destination (memory)
break;
case X86II::MRMSrcReg: // for instructions that use the Mod/RM byte
...                // to specify a source (register)
break;
case X86II::MRMSrcMem: // for instructions that use the Mod/RM byte
...                // to specify a source (memory)
break;
case X86II::MRM0r: case X86II::MRM1r: // for instructions that operate on
case X86II::MRM2r: case X86II::MRM3r: // a REGISTER r/m operand and
case X86II::MRM4r: case X86II::MRM5r: // use the Mod/RM byte and a field
case X86II::MRM6r: case X86II::MRM7r: // to hold extended opcode data
...
break;
case X86II::MRM0m: case X86II::MRM1m: // for instructions that operate on
case X86II::MRM2m: case X86II::MRM3m: // a MEMORY r/m operand and
case X86II::MRM4m: case X86II::MRM5m: // use the Mod/RM byte and a field
case X86II::MRM6m: case X86II::MRM7m: // to hold extended opcode data
...
break;
case X86II::MRMInitReg: // for instructions whose source and
...                // destination are the same register
break;
}

```

The implementations of these case statements often first emit the opcode and then get the operand(s). Then depending upon the operand, helper methods may be called to process the operand(s). For example, in `X86CodeEmitter.cpp`, for the `X86II::AddRegFrm` case, the first data emitted (by `emitByte`) is the opcode added to the register operand. Then an object representing the machine operand, `MO1`, is extracted. The helper methods such as `isImmediate`, `isGlobalAddress`, `isExternalSymbol`, `isConstantPoolIndex`, and `isJumpTableIndex` determine the operand type. (`X86CodeEmitter.cpp` also has private methods such as `emitConstant`, `emitGlobalAddress`, `emitExternalSymbolAddress`, `emitConstPoolAddress`, and `emitJumpTableAddress` that emit the data into the output stream.)

```

case X86II::AddRegFrm:
    MCE.emitByte(BaseOpcode + getX86RegNum(MI.getOperand(CurOp++).getReg()));

    if (CurOp != NumOps) {
        const MachineOperand &MO1 = MI.getOperand(CurOp++);
        unsigned Size = X86InstrInfo::sizeofImm(Desc);
        if (MO1.isImmediate())
            emitConstant(MO1.getImm(), Size);
        else {
            unsigned rt = Is64BitMode ? X86::reloc_pcrel_word
                : (IsPIC ? X86::reloc_picrel_word : X86::reloc_absolute_word);

```



```

if (Opcode == X86::MOV64ri)
    rt = X86::reloc_absolute_dword; // FIXME: add X86II flag?
if (MO1.isGlobalAddress()) {
    bool NeedStub = isa<Function>(MO1.getGlobal());
    bool isLazy = gvNeedsLazyPtr(MO1.getGlobal());
    emitGlobalAddress(MO1.getGlobal(), rt, MO1.getOffset(), 0,
                      NeedStub, isLazy);
} else if (MO1.isExternalSymbol())
    emitExternalSymbolAddress(MO1.getSymbolName(), rt);
else if (MO1.isConstantPoolIndex())
    emitConstPoolAddress(MO1.getIndex(), rt);
else if (MO1.isJumpTableIndex())
    emitJumpTableAddress(MO1.getIndex(), rt);
}
break;

```

In the previous example, `XXXCodeEmitter.cpp` uses the variable `rt`, which is a `RelocationType` enum that may be used to relocate addresses (for example, a global address with a PIC base offset). The `RelocationType` enum for that target is defined in the short target-specific `XXXRelocations.h` file. The `RelocationType` is used by the `relocate` method defined in `XXXJITInfo.cpp` to rewrite addresses for referenced global symbols.

For example, `X86Relocations.h` specifies the following relocation types for the X86 addresses. In all four cases, the relocated value is added to the value already in memory. For `reloc_pcrel_word` and `reloc_picrel_word`, there is an additional initial adjustment.

```

enum RelocationType {
    reloc_pcrel_word = 0, // add reloc value after adjusting for the PC loc
    reloc_picrel_word = 1, // add reloc value after adjusting for the PIC base
    reloc_absolute_word = 2, // absolute relocation; no additional adjustment
    reloc_absolute_dword = 3 // absolute relocation; no additional adjustment
};

```

Target JIT Info

`XXXJITInfo.cpp` implements the JIT interfaces for target-specific code-generation activities, such as emitting machine code and stubs. At minimum, a target-specific version of `XXXJITInfo` implements the following:

- `getLazyResolverFunction` — Initializes the JIT, gives the target a function that is used for compilation.
- `emitFunctionStub` — Returns a native function with a specified address for a callback function.
- `relocate` — Changes the addresses of referenced globals, based on relocation types.
- Callback function that are wrappers to a function stub that is used when the real target is not initially known.

`getLazyResolverFunction` is generally trivial to implement. It makes the incoming parameter as the global `JITCompilerFunction` and returns the callback function that will be used as a function wrapper. For the Alpha target (in `AlphaJITInfo.cpp`), the `getLazyResolverFunction` implementation is simply:

```

TargetJITInfo::LazyResolverFn AlphaJITInfo::getLazyResolverFunction(
    JITCompilerFn F) {
    JITCompilerFunction = F;
    return AlphaCompilationCallback;
}

```

For the X86 target, the `getLazyResolverFunction` implementation is a little more complicated, because it returns a different callback function for processors with SSE instructions and XMM registers.

The callback function initially saves and later restores the callee register values, incoming arguments, and frame and return address. The callback function needs low-level access to the registers or stack, so it is typically implemented with assembler.

4.20 Garbage Collection with LLVM

- *Abstract*
- *Quick Start*
- *Introduction*
 - *What is Garbage Collection?*
 - *Goals and non-goals*
- *LLVM IR Features*
 - *Specifying GC code generation: `gc "..."`*
 - *Identifying GC roots on the stack*
 - * *Using `gc.statepoint`*
 - * *Using `llvm.gcwrite`*
 - *Reading and writing references in the heap*
 - * *Write barrier: `llvm.gcwrite`*
 - * *Read barrier: `llvm.gcread`*
- *Built In GC Strategies*
 - *The Shadow Stack GC*
 - *The ‘Erlang’ and ‘Ocaml’ GCs*
 - *The Statepoint Example GC*
 - *The CoreCLR GC*
- *Custom GC Strategies*
 - *Collector Requirements*
 - *Implementing a collector plugin*
 - *Overview of available features*
 - *Computing stack maps*
 - *Initializing roots to null: `InitRoots`*
 - *Custom lowering of intrinsics: `CustomRoots`, `CustomReadBarriers`, and `CustomWriteBarriers`*
 - *Generating safe points: `NeededSafePoints`*
 - *Emitting assembly code: `GCMetadataPrinter`*
- *References*

4.20.1 Abstract

This document covers how to integrate LLVM into a compiler for a language which supports garbage collection. **Note that LLVM itself does not provide a garbage collector.** You must provide your own.

4.20.2 Quick Start

First, you should pick a collector strategy. LLVM includes a number of built in ones, but you can also implement a loadable plugin with a custom definition. Note that the collector strategy is a description of how LLVM should generate code such that it interacts with your collector and runtime, not a description of the collector itself.

Next, mark your generated functions as using your chosen collector strategy. From c++, you can call:

```
F.setGC(<collector description name>);
```

This will produce IR like the following fragment:

```
define void @foo() gc "<collector description name>" { ... }
```

When generating LLVM IR for your functions, you will need to:

- Use `@llvm.gcread` and/or `@llvm.gfwrite` in place of standard load and store instructions. These intrinsics are used to represent load and store barriers. If your collector does not require such barriers, you can skip this step.
- Use the memory allocation routines provided by your garbage collector's runtime library.
- If your collector requires them, generate type maps according to your runtime's binary interface. LLVM is not involved in the process. In particular, the LLVM type system is not suitable for conveying such information through the compiler.
- Insert any coordination code required for interacting with your collector. Many collectors require running application code to periodically check a flag and conditionally call a runtime function. This is often referred to as a safepoint poll.

You will need to identify roots (i.e. references to heap objects your collector needs to know about) in your generated IR, so that LLVM can encode them into your final stack maps. Depending on the collector strategy chosen, this is accomplished by using either the `@llvm.gcroot` intrinsics or an `gc.statepoint` relocation sequence.

Don't forget to create a root for each intermediate value that is generated when evaluating an expression. In `h(f(), g())`, the result of `f()` could easily be collected if evaluating `g()` triggers a collection.

Finally, you need to link your runtime library with the generated program executable (for a static compiler) or ensure the appropriate symbols are available for the runtime linker (for a JIT compiler).

4.20.3 Introduction

What is Garbage Collection?

Garbage collection is a widely used technique that frees the programmer from having to know the lifetimes of heap objects, making software easier to produce and maintain. Many programming languages rely on garbage collection for automatic memory management. There are two primary forms of garbage collection: conservative and accurate.

Conservative garbage collection often does not require any special support from either the language or the compiler: it can handle non-type-safe programming languages (such as C/C++) and does not require any special information from the compiler. The [Boehm collector](#) is an example of a state-of-the-art conservative collector.

Accurate garbage collection requires the ability to identify all pointers in the program at run-time (which requires that the source-language be type-safe in most cases). Identifying pointers at run-time requires compiler support to locate all places that hold live pointer variables at run-time, including the *processor stack and registers*.

Conservative garbage collection is attractive because it does not require any special compiler support, but it does have problems. In particular, because the conservative garbage collector cannot *know* that a particular word in the machine is a pointer, it cannot move live objects in the heap (preventing the use of compacting and generational GC algorithms) and it can occasionally suffer from memory leaks due to integer values that happen to point to objects in the program. In addition, some aggressive compiler transformations can break conservative garbage collectors (though these seem rare in practice).

Accurate garbage collectors do not suffer from any of these problems, but they can suffer from degraded scalar optimization of the program. In particular, because the runtime must be able to identify and update all pointers active in the program, some optimizations are less effective. In practice, however, the locality and performance benefits of using aggressive garbage collection techniques dominates any low-level losses.

This document describes the mechanisms and interfaces provided by LLVM to support accurate garbage collection.

Goals and non-goals

LLVM's intermediate representation provides *garbage collection intrinsics* that offer support for a broad class of collector models. For instance, the intrinsics permit:

- semi-space collectors
- mark-sweep collectors
- generational collectors
- incremental collectors
- concurrent collectors
- cooperative collectors
- reference counting

We hope that the support built into the LLVM IR is sufficient to support a broad class of garbage collected languages including Scheme, ML, Java, C#, Perl, Python, Lua, Ruby, other scripting languages, and more.

Note that LLVM **does not itself provide a garbage collector** — this should be part of your language's runtime library. LLVM provides a framework for describing the garbage collectors requirements to the compiler. In particular, LLVM provides support for generating stack maps at call sites, polling for a safepoint, and emitting load and store barriers. You can also extend LLVM - possibly through a loadable *code generation plugins* - to generate code and data structures which conforms to the *binary interface* specified by the *runtime library*. This is similar to the relationship between LLVM and DWARF debugging info, for example. The difference primarily lies in the lack of an established standard in the domain of garbage collection — thus the need for a flexible extension mechanism.

The aspects of the binary interface with which LLVM's GC support is concerned are:

- Creation of GC safepoints within code where collection is allowed to execute safely.
- Computation of the stack map. For each safe point in the code, object references within the stack frame must be identified so that the collector may traverse and perhaps update them.
- Write barriers when storing object references to the heap. These are commonly used to optimize incremental scans in generational collectors.
- Emission of read barriers when loading object references. These are useful for interoperating with concurrent collectors.

There are additional areas that LLVM does not directly address:

- Registration of global roots with the runtime.
- Registration of stack map entries with the runtime.
- The functions used by the program to allocate memory, trigger a collection, etc.
- Computation or compilation of type maps, or registration of them with the runtime. These are used to crawl the heap for object references.

In general, LLVM's support for GC does not include features which can be adequately addressed with other features of the IR and does not specify a particular binary interface. On the plus side, this means that you should be able to integrate LLVM with an existing runtime. On the other hand, it can have the effect of leaving a lot of work for the developer of a novel language. We try to mitigate this by providing built in collector strategy descriptions that can work with many common collector designs and easy extension points. If you don't already have a specific binary interface you need to support, we recommend trying to use one of these built in collector strategies.

4.20.4 LLVM IR Features

This section describes the garbage collection facilities provided by the *LLVM intermediate representation*. The exact behavior of these IR features is specified by the selected *GC strategy description*.

Specifying GC code generation: `gc "..."`

```
define <returntype> @name(...) gc "name" { ... }
```

The `gc` function attribute is used to specify the desired GC strategy to the compiler. Its programmatic equivalent is the `setGC` method of `Function`.

Setting `gc "name"` on a function triggers a search for a matching subclass of `GCStrategy`. Some collector strategies are built in. You can add others using either the loadable plugin mechanism, or by patching your copy of LLVM. It is the selected GC strategy which defines the exact nature of the code generated to support GC. If none is found, the compiler will raise an error.

Specifying the GC style on a per-function basis allows LLVM to link together programs that use different garbage collection algorithms (or none at all).

Identifying GC roots on the stack

LLVM currently supports two different mechanisms for describing references in compiled code at safepoints. `llvm.gcroot` is the older mechanism; `gc.statepoint` has been added more recently. At the moment, you can choose either implementation (on a per *GC strategy* basis). Longer term, we will probably either migrate away from `llvm.gcroot` entirely, or substantially merge their implementations. Note that most new development work is focused on `gc.statepoint`.

Using `gc.statepoint`

This page contains detailed documentation for `gc.statepoint`.

Using `llvm.gcwrite`

```
void @llvm.gcroot(i8** %ptrloc, i8* %metadata)
```

The `llvm.gcroot` intrinsic is used to inform LLVM that a stack variable references an object on the heap and is to be tracked for garbage collection. The exact impact on generated code is specified by the Function's selected *GC strategy*. All calls to `llvm.gcroot` **must** reside inside the first basic block.

The first argument **must** be a value referring to an `alloca` instruction or a bitcast of an `alloca`. The second contains a pointer to metadata that should be associated with the pointer, and **must** be a constant or global value address. If your target collector uses tags, use a null pointer for metadata.

A compiler which performs manual SSA construction **must** ensure that SSA values representing GC references are stored in to the `alloca` passed to the respective `gcroot` before every call site and reloaded after every call. A compiler which uses `mem2reg` to raise imperative code using `alloca` into SSA form need only add a call to `@llvm.gcroot` for those variables which are pointers into the GC heap.

It is also important to mark intermediate values with `llvm.gcroot`. For example, consider `h(f(), g())`. Beware leaking the result of `f()` in the case that `g()` triggers a collection. Note, that stack variables must be initialized and marked with `llvm.gcroot` in function's prologue.

The `%metadata` argument can be used to avoid requiring heap objects to have 'isa' pointers or tag bits. [Appel89, Goldberg91, Tolmach94] If specified, its value will be tracked along with the location of the pointer in the stack frame.

Consider the following fragment of Java code:

```
{
    Object X;    // A null-initialized reference to an object
    ...
}
```

This block (which may be located in the middle of a function or in a loop nest), could be compiled to this LLVM code:

```
Entry:
    ;; In the entry block for the function, allocate the
    ;; stack space for X, which is an LLVM pointer.
    %X = alloca %Object*

    ;; Tell LLVM that the stack space is a stack root.
    ;; Java has type-tags on objects, so we pass null as metadata.
    %tmp = bitcast %Object** %X to i8**
    call void @llvm.gcroot(i8** %tmp, i8* null)
    ...

    ;; "CodeBlock" is the block corresponding to the start
    ;; of the scope above.
CodeBlock:
    ;; Java null-initializes pointers.
    store %Object* null, %Object** %X

    ...

    ;; As the pointer goes out of scope, store a null value into
    ;; it, to indicate that the value is no longer live.
    store %Object* null, %Object** %X
    ...
```

Reading and writing references in the heap

Some collectors need to be informed when the mutator (the program that needs garbage collection) either reads a pointer from or writes a pointer to a field of a heap object. The code fragments inserted at these points are called *read*

barriers and *write barriers*, respectively. The amount of code that needs to be executed is usually quite small and not on the critical path of any computation, so the overall performance impact of the barrier is tolerable.

Barriers often require access to the *object pointer* rather than the *derived pointer* (which is a pointer to the field within the object). Accordingly, these intrinsics take both pointers as separate arguments for completeness. In this snippet, `%object` is the object pointer, and `%derived` is the derived pointer:

```
;; An array type.
@class.Array = type { %class.Object, i32, [0 x %class.Object*] }
...

;; Load the object pointer from a gcroot.
%object = load %class.Array** %object_addr

;; Compute the derived pointer.
%derived = getelementptr %object, i32 0, i32 2, i32 %n
```

LLVM does not enforce this relationship between the object and derived pointer (although a particular *collector strategy* might). However, it would be an unusual collector that violated it.

The use of these intrinsics is naturally optional if the target GC does not require the corresponding barrier. The GC strategy used with such a collector should replace the intrinsic calls with the corresponding `load` or `store` instruction if they are used.

One known deficiency with the current design is that the barrier intrinsics do not include the size or alignment of the underlying operation performed. It is currently assumed that the operation is of pointer size and the alignment is assumed to be the target machine's default alignment.

Write barrier: `llvm.gcwrite`

```
void @llvm.gcwrite(i8* %value, i8* %object, i8** %derived)
```

For write barriers, LLVM provides the `llvm.gcwrite` intrinsic function. It has exactly the same semantics as a non-volatile `store` to the derived pointer (the third argument). The exact code generated is specified by the Function's selected *GC strategy*.

Many important algorithms require write barriers, including generational and concurrent collectors. Additionally, write barriers could be used to implement reference counting.

Read barrier: `llvm.gcread`

```
i8* @llvm.gcread(i8* %object, i8** %derived)
```

For read barriers, LLVM provides the `llvm.gcread` intrinsic function. It has exactly the same semantics as a non-volatile `load` from the derived pointer (the second argument). The exact code generated is specified by the Function's selected *GC strategy*.

Read barriers are needed by fewer algorithms than write barriers, and may have a greater performance impact since pointer reads are more frequent than writes.

4.20.5 Built In GC Strategies

LLVM includes built in support for several varieties of garbage collectors.

The Shadow Stack GC

To use this collector strategy, mark your functions with:

```
F.setGC("shadow-stack");
```

Unlike many GC algorithms which rely on a cooperative code generator to compile stack maps, this algorithm carefully maintains a linked list of stack roots [Henderson2002]. This so-called “shadow stack” mirrors the machine stack. Maintaining this data structure is slower than using a stack map compiled into the executable as constant data, but has a significant portability advantage because it requires no special support from the target code generator, and does not require tricky platform-specific code to crawl the machine stack.

The tradeoff for this simplicity and portability is:

- High overhead per function call.
- Not thread-safe.

Still, it’s an easy way to get started. After your compiler and runtime are up and running, writing a *plugin* will allow you to take advantage of *more advanced GC features* of LLVM in order to improve performance.

The shadow stack doesn’t imply a memory allocation algorithm. A semispace collector or building atop `malloc` are great places to start, and can be implemented with very little code.

When it comes time to collect, however, your runtime needs to traverse the stack roots, and for this it needs to integrate with the shadow stack. Luckily, doing so is very simple. (This code is heavily commented to help you understand the data structure, but there are only 20 lines of meaningful code.)

```
/// @brief The map for a single function's stack frame. One of these is
///         compiled as constant data into the executable for each function.
///
/// Storage of metadata values is elided if the %metadata parameter to
/// @llvm.gcroot is null.
struct FrameMap {
    int32_t NumRoots;    ///< Number of roots in stack frame.
    int32_t NumMeta;     ///< Number of metadata entries. May be < NumRoots.
    const void *Meta[0]; ///< Metadata for each root.
};

/// @brief A link in the dynamic shadow stack. One of these is embedded in
///         the stack frame of each function on the call stack.
struct StackEntry {
    StackEntry *Next;    ///< Link to next stack entry (the caller's).
    const FrameMap *Map; ///< Pointer to constant FrameMap.
    void *Roots[0];     ///< Stack roots (in-place array).
};

/// @brief The head of the singly-linked list of StackEntries. Functions push
///         and pop onto this in their prologue and epilogue.
///
/// Since there is only a global list, this technique is not threadsafe.
StackEntry *llvm_gc_root_chain;

/// @brief Calls Visitor(root, meta) for each GC root on the stack.
///         root and meta are exactly the values passed to
///         @llvm.gcroot.
///
/// Visitor could be a function to recursively mark live objects. Or it
/// might copy them to another heap or generation.
///
```

```
/// @param Visitor A function to invoke for every GC root on the stack.
void visitGCRoots(void (*Visitor)(void **Root, const void *Meta)) {
    for (StackEntry *R = llvm_gc_root_chain; R; R = R->Next) {
        unsigned i = 0;

        // For roots [0, NumMeta), the metadata pointer is in the FrameMap.
        for (unsigned e = R->Map->NumMeta; i != e; ++i)
            Visitor(&R->Roots[i], R->Map->Meta[i]);

        // For roots [NumMeta, NumRoots), the metadata pointer is null.
        for (unsigned e = R->Map->NumRoots; i != e; ++i)
            Visitor(&R->Roots[i], NULL);
    }
}
```

The ‘Erlang’ and ‘Ocaml’ GCs

LLVM ships with two example collectors which leverage the `gcroot` mechanisms. To our knowledge, these are not actually used by any language runtime, but they do provide a reasonable starting point for someone interested in writing an `gcroot` compatible GC plugin. In particular, these are the only in tree examples of how to produce a custom binary stack map format using a `gcroot` strategy.

As their names imply, the binary format produced is intended to model that used by the Erlang and OCaml compilers respectively.

The Statepoint Example GC

```
F.setGC("statepoint-example");
```

This GC provides an example of how one might use the infrastructure provided by `gc.statepoint`. This example GC is compatible with the [PlaceSafepoints](#) and [RewriteStatepointsForGC](#) utility passes which simplify `gc.statepoint` sequence insertion. If you need to build a custom GC strategy around the `gc.statepoints` mechanisms, it is recommended that you use this one as a starting point.

This GC strategy does not support read or write barriers. As a result, these intrinsics are lowered to normal loads and stores.

The stack map format generated by this GC strategy can be found in the [Stack Map Section](#) using a format documented [here](#). This format is intended to be the standard format supported by LLVM going forward.

The CoreCLR GC

```
F.setGC("coreclr");
```

This GC leverages the `gc.statepoint` mechanism to support the [CoreCLR](#) runtime.

Support for this GC strategy is a work in progress. This strategy will differ from *statepoint-example GC* strategy in certain aspects like:

- Base-pointers of interior pointers are not explicitly tracked and reported.
- A different format is used for encoding stack maps.
- Safe-point polls are only needed before loop-back edges and before tail-calls (not needed at function-entry).

4.20.6 Custom GC Strategies

If none of the built in GC strategy descriptions met your needs above, you will need to define a custom GCStrategy and possibly, a custom LLVM pass to perform lowering. Your best example of where to start defining a custom GCStrategy would be to look at one of the built in strategies.

You may be able to structure this additional code as a loadable plugin library. Loadable plugins are sufficient if all you need is to enable a different combination of built in functionality, but if you need to provide a custom lowering pass, you will need to build a patched version of LLVM. If you think you need a patched build, please ask for advice on `llvm-dev`. There may be an easy way we can extend the support to make it work for your use case without requiring a custom build.

Collector Requirements

You should be able to leverage any existing collector library that includes the following elements:

1. A memory allocator which exposes an allocation function your compiled code can call.
2. A binary format for the stack map. A stack map describes the location of references at a safepoint and is used by precise collectors to identify references within a stack frame on the machine stack. Note that collectors which conservatively scan the stack don't require such a structure.
3. A stack crawler to discover functions on the call stack, and enumerate the references listed in the stack map for each call site.
4. A mechanism for identifying references in global locations (e.g. global variables).
5. If your collector requires them, an LLVM IR implementation of your collectors load and store barriers. Note that since many collectors don't require barriers at all, LLVM defaults to lowering such barriers to normal loads and stores unless you arrange otherwise.

Implementing a collector plugin

User code specifies which GC code generation to use with the `gc` function attribute or, equivalently, with the `setGC` method of `Function`.

To implement a GC plugin, it is necessary to subclass `llvm::GCStrategy`, which can be accomplished in a few lines of boilerplate code. LLVM's infrastructure provides access to several important algorithms. For an uncontroversial collector, all that remains may be to compile LLVM's computed stack map to assembly code (using the binary representation expected by the runtime library). This can be accomplished in about 100 lines of code.

This is not the appropriate place to implement a garbage collected heap or a garbage collector itself. That code should exist in the language's runtime library. The compiler plugin is responsible for generating code which conforms to the binary interface defined by library, most essentially the *stack map*.

To subclass `llvm::GCStrategy` and register it with the compiler:

```
// lib/MyGC/MyGC.cpp - Example LLVM GC plugin

#include "llvm/CodeGen/GCStrategy.h"
#include "llvm/CodeGen/GCMetadata.h"
#include "llvm/Support/Compiler.h"

using namespace llvm;

namespace {
  class LLVM_LIBRARY_VISIBILITY MyGC : public GCStrategy {
  public:
```

```
MyGC() {}  
};  
  
GCRegistry::Add<MyGC>  
X("mygc", "My bespoke garbage collector.");  
}
```

This boilerplate collector does nothing. More specifically:

- `llvm.gcread` calls are replaced with the corresponding `load` instruction.
- `llvm.gcwrite` calls are replaced with the corresponding `store` instruction.
- No safe points are added to the code.
- The stack map is not compiled into the executable.

Using the LLVM makefiles, this code can be compiled as a plugin using a simple makefile:

```
# lib/MyGC/Makefile  
  
LEVEL := ../..  
LIBRARYNAME = MyGC  
LOADABLE_MODULE = 1  
  
include $(LEVEL)/Makefile.common
```

Once the plugin is compiled, code using it may be compiled using `llc -load=MyGC.so` (though `MyGC.so` may have some other platform-specific extension):

```
$ cat sample.ll  
define void @f() gc "mygc" {  
entry:  
    ret void  
}  
$ llvm-as < sample.ll | llc -load=MyGC.so
```

It is also possible to statically link the collector plugin into tools, such as a language-specific compiler front-end.

Overview of available features

`GCStrategy` provides a range of features through which a plugin may do useful work. Some of these are callbacks, some are algorithms that can be enabled, disabled, or customized. This matrix summarizes the supported (and planned) features and correlates them with the collection techniques which typically require them.

Algorithm	Done	Shadow stack	ref-count	mark-sweep	copying	incremental	threaded	concurrent
stack map								
initialize roots								
derived pointers	NO						N*	N*
custom lowering								
<i>gcroot</i>								
<i>gcwrite</i>								
<i>gcread</i>								
safe points								
<i>in calls</i>								
<i>before calls</i>								
<i>for loops</i>	NO						N	N
<i>before escape</i>								
emit code at safe points	NO						N	N
output								
<i>assembly</i>								
<i>JIT</i>	NO			?	?	?	?	?
<i>obj</i>	NO			?	?	?	?	?
live analysis	NO			?	?	?	?	?
register map	NO			?	?	?	?	?
* Derived pointers only pose a hazard to copying collections.								
? denotes a feature which could be utilized if available.								

To be clear, the collection techniques above are defined as:

Shadow Stack The mutator carefully maintains a linked list of stack roots.

Reference Counting The mutator maintains a reference count for each object and frees an object when its count falls to zero.

Mark-Sweep When the heap is exhausted, the collector marks reachable objects starting from the roots, then deallocates unreachable objects in a sweep phase.

Copying As reachability analysis proceeds, the collector copies objects from one heap area to another, compacting them in the process. Copying collectors enable highly efficient “bump pointer” allocation and can improve locality of reference.

Incremental (Including generational collectors.) Incremental collectors generally have all the properties of a copying collector (regardless of whether the mature heap is compacting), but bring the added complexity of requiring write barriers.

Threaded Denotes a multithreaded mutator; the collector must still stop the mutator (“stop the world”) before beginning reachability analysis. Stopping a multithreaded mutator is a complicated problem. It generally requires highly platform-specific code in the runtime, and the production of carefully designed machine code at safe points.

Concurrent In this technique, the mutator and the collector run concurrently, with the goal of eliminating pause times. In a *cooperative* collector, the mutator further aids with collection should a pause occur, allowing collection to take advantage of multiprocessor hosts. The “stop the world” problem of threaded collectors is generally still present to a limited extent. Sophisticated marking algorithms are necessary. Read barriers may be necessary.

As the matrix indicates, LLVM’s garbage collection infrastructure is already suitable for a wide variety of collectors, but does not currently extend to multithreaded programs. This will be added in the future as there is interest.

Computing stack maps

LLVM automatically computes a stack map. One of the most important features of a `GCStrategy` is to compile this information into the executable in the binary representation expected by the runtime library.

The stack map consists of the location and identity of each GC root in the each function in the module. For each root:

- `RootNum`: The index of the root.
- `StackOffset`: The offset of the object relative to the frame pointer.
- `RootMetadata`: The value passed as the `%metadata` parameter to the `@llvm.gcroot` intrinsic.

Also, for the function as a whole:

- **`getFrameSize()`**: The overall size of the function's initial stack frame, not accounting for any dynamic allocation.
- `roots_size()`: The count of roots in the function.

To access the stack map, use `GCFUNCTIONMetadata::roots_begin()` and `-end()` from the *GCMetadataPrinter*:

```
for (iterator I = begin(), E = end(); I != E; ++I) {
    GCFUNCTIONInfo *FI = *I;
    unsigned FrameSize = FI->getFrameSize();
    size_t RootCount = FI->roots_size();

    for (GCFUNCTIONInfo::roots_iterator RI = FI->roots_begin(),
         RE = FI->roots_end();
         RI != RE; ++RI) {
        int RootNum = RI->Num;
        int RootStackOffset = RI->StackOffset;
        Constant *RootMetadata = RI->Metadata;
    }
}
```

If the `llvm.gcroot` intrinsic is eliminated before code generation by a custom lowering pass, LLVM will compute an empty stack map. This may be useful for collector plugins which implement reference counting or a shadow stack.

Initializing roots to null: `InitRoots`

```
MyGC::MyGC() {
    InitRoots = true;
}
```

When set, LLVM will automatically initialize each root to `null` upon entry to the function. This prevents the GC's sweep phase from visiting uninitialized pointers, which will almost certainly cause it to crash. This initialization occurs before custom lowering, so the two may be used together.

Since LLVM does not yet compute liveness information, there is no means of distinguishing an uninitialized stack root from an initialized one. Therefore, this feature should be used by all GC plugins. It is enabled by default.

Custom lowering of intrinsics: `CustomRoots`, `CustomReadBarriers`, and `CustomWriteBarriers`

For GCs which use barriers or unusual treatment of stack roots, these flags allow the collector to perform arbitrary transformations of the LLVM IR:

```
class MyGC : public GCStrategy {
public:
    MyGC() {
        CustomRoots = true;
        CustomReadBarriers = true;
        CustomWriteBarriers = true;
    }
};
```

If any of these flags are set, LLVM suppresses its default lowering for the corresponding intrinsics. Instead, you must provide a custom Pass which lowers the intrinsics as desired. If you have opted in to custom lowering of a particular intrinsic your pass **must** eliminate all instances of the corresponding intrinsic in functions which opt in to your GC. The best example of such a pass is the ShadowStackGC and its ShadowStackGCLowering pass.

There is currently no way to register such a custom lowering pass without building a custom copy of LLVM.

Generating safe points: NeededSafePoints

LLVM can compute four kinds of safe points:

```
namespace GC {
    /// PointKind - The type of a collector-safe point.
    ///
    enum PointKind {
        Loop,    ///< Instr is a loop (backwards branch).
        Return,  ///< Instr is a return instruction.
        PreCall, ///< Instr is a call instruction.
        PostCall ///< Instr is the return address of a call.
    };
}
```

A collector can request any combination of the four by setting the NeededSafePoints mask:

```
MyGC::MyGC() {
    NeededSafePoints = 1 << GC::Loop
                    | 1 << GC::Return
                    | 1 << GC::PreCall
                    | 1 << GC::PostCall;
}
```

It can then use the following routines to access safe points.

```
for (iterator I = begin(), E = end(); I != E; ++I) {
    GCFunctionInfo *MD = *I;
    size_t PointCount = MD->size();

    for (GCFunctionInfo::iterator PI = MD->begin(),
        PE = MD->end(); PI != PE; ++PI) {
        GC::PointKind PointKind = PI->Kind;
        unsigned PointNum = PI->Num;
    }
}
```

Almost every collector requires PostCall safe points, since these correspond to the moments when the function is suspended during a call to a subroutine.

Threaded programs generally require `Loop` safe points to guarantee that the application will reach a safe point within a bounded amount of time, even if it is executing a long-running loop which contains no function calls.

Threaded collectors may also require `Return` and `PreCall` safe points to implement “stop the world” techniques using self-modifying code, where it is important that the program not exit the function without reaching a safe point (because only the topmost function has been patched).

Emitting assembly code: `GCMetadataPrinter`

LLVM allows a plugin to print arbitrary assembly code before and after the rest of a module’s assembly code. At the end of the module, the GC can compile the LLVM stack map into assembly code. (At the beginning, this information is not yet computed.)

Since `AsmWriter` and `CodeGen` are separate components of LLVM, a separate abstract base class and registry is provided for printing assembly code, the `GCMetadataPrinter` and `GCMetadataPrinterRegistry`. The `AsmWriter` will look for such a subclass if the `GCStrategy` sets `UsesMetadata`:

```
MyGC::MyGC() {
    UsesMetadata = true;
}
```

This separation allows JIT-only clients to be smaller.

Note that LLVM does not currently have analogous APIs to support code generation in the JIT, nor using the object writers.

```
// lib/MyGC/MyGCPrinter.cpp - Example LLVM GC printer

#include "llvm/CodeGen/GCMetadataPrinter.h"
#include "llvm/Support/Compiler.h"

using namespace llvm;

namespace {
    class LLVM_LIBRARY_VISIBILITY MyGCPrinter : public GCMetadataPrinter {
    public:
        virtual void beginAssembly(AsmPrinter &AP);

        virtual void finishAssembly(AsmPrinter &AP);
    };

    GCMetadataPrinterRegistry::Add<MyGCPrinter>
    X("mygc", "My bespoke garbage collector.");
}
```

The collector should use `AsmPrinter` to print portable assembly code. The collector itself contains the stack map for the entire module, and may access the `GCFunctionInfo` using its own `begin()` and `end()` methods. Here’s a realistic example:

```
#include "llvm/CodeGen/AsmPrinter.h"
#include "llvm/IR/Function.h"
#include "llvm/IR/DataLayout.h"
#include "llvm/Target/TargetAsmInfo.h"
#include "llvm/Target/TargetMachine.h"

void MyGCPrinter::beginAssembly(AsmPrinter &AP) {
    // Nothing to do.
}
```

```

void MyGCPrinter::finishAssembly(AsmPrinter &AP) {
    MCStreamer &OS = AP.OutStreamer;
    unsigned IntPtrSize = AP.getPointerSize();

    // Put this in the data section.
    OS.SwitchSection(AP.getObjFileLowering().getDataSection());

    // For each function...
    for (iterator FI = begin(), FE = end(); FI != FE; ++FI) {
        GCFunctionInfo &MD = **FI;

        // A compact GC layout. Emit this data structure:
        //
        // struct {
        //     int32_t PointCount;
        //     void *SafePointAddress[PointCount];
        //     int32_t StackFrameSize; // in words
        //     int32_t StackArity;
        //     int32_t LiveCount;
        //     int32_t LiveOffsets[LiveCount];
        // } __gcmmap_<FUNCTIONNAME>;

        // Align to address width.
        AP.EmitAlignment(IntPtrSize == 4 ? 2 : 3);

        // Emit PointCount.
        OS.AddComment("safe point count");
        AP.EmitInt32(MD.size());

        // And each safe point...
        for (GCFunctionInfo::iterator PI = MD.begin(),
            PE = MD.end(); PI != PE; ++PI) {
            // Emit the address of the safe point.
            OS.AddComment("safe point address");
            MCSymbol *Label = PI->Label;
            AP.EmitLabelPlusOffset(Label/*Hi*/, 0/*Offset*/, 4/*Size*/);
        }

        // Stack information never change in safe points! Only print info from the
        // first call-site.
        GCFunctionInfo::iterator PI = MD.begin();

        // Emit the stack frame size.
        OS.AddComment("stack frame size (in words)");
        AP.EmitInt32(MD.getFrameSize() / IntPtrSize);

        // Emit stack arity, i.e. the number of stacked arguments.
        unsigned RegisteredArgs = IntPtrSize == 4 ? 5 : 6;
        unsigned StackArity = MD.getFunction().arg_size() > RegisteredArgs ?
            MD.getFunction().arg_size() - RegisteredArgs : 0;
        OS.AddComment("stack arity");
        AP.EmitInt32(StackArity);

        // Emit the number of live roots in the function.
        OS.AddComment("live root count");
        AP.EmitInt32(MD.live_size(PI));
    }
}

```

```
// And for each live root...
for (GCFunctionInfo::live_iterator LI = MD.live_begin(PI),
      LE = MD.live_end(PI);
      LI != LE; ++LI) {
    // Emit live root's offset within the stack frame.
    OS.AddComment("stack index (offset / wordsize)");
    AP.EmitInt32(LI->StackOffset);
}
}
```

4.20.7 References

[Appel89] Runtime Tags Aren't Necessary. Andrew W. Appel. Lisp and Symbolic Computation 19(7):703-705, July 1989. [Goldberg91] Tag-free garbage collection for strongly typed programming languages. Benjamin Goldberg. ACM SIGPLAN PLDI'91. [Tolmach94] Tag-free garbage collection using explicit type parameters. Andrew Tolmach. Proceedings of the 1994 ACM conference on LISP and functional programming. [Henderson2002] [Accurate Garbage Collection in an Uncooperative Environment](#)

4.21 Writing an LLVM Pass

- *Introduction — What is a pass?*
- *Quick Start — Writing hello world*
 - *Setting up the build environment*
 - *Basic code required*
 - *Running a pass with opt*
- *Pass classes and requirements*
 - *The ImmutablePass class*
 - *The ModulePass class*
 - * *The runOnModule method*
 - *The CallGraphSCCPass class*
 - * *The doInitialization(CallGraph &) method*
 - * *The runOnSCC method*
 - * *The doFinalization(CallGraph &) method*
 - *The FunctionPass class*
 - * *The doInitialization(Module &) method*
 - * *The runOnFunction method*
 - * *The doFinalization(Module &) method*
 - *The LoopPass class*
 - * *The doInitialization(Loop *, LPPassManager &) method*

- * *The runOnLoop method*
- * *The doFinalization() method*
- *The RegionPass class*
 - * *The doInitialization(Region *, RGPassManager &) method*
 - * *The runOnRegion method*
 - * *The doFinalization() method*
- *The BasicBlockPass class*
 - * *The doInitialization(Function &) method*
 - * *The runOnBasicBlock method*
 - * *The doFinalization(Function &) method*
- *The MachineFunctionPass class*
 - * *The runOnMachineFunction(MachineFunction &MF) method*
- *Pass registration*
 - * *The print method*
- *Specifying interactions between passes*
 - * *The getAnalysisUsage method*
 - * *The AnalysisUsage::addRequired<> and AnalysisUsage::addRequiredTransitive<> methods*
 - * *The AnalysisUsage::addPreserved<> method*
 - * *Example implementations of getAnalysisUsage*
 - * *The getAnalysis<> and getAnalysisIfAvailable<> methods*
- *Implementing Analysis Groups*
 - * *Analysis Group Concepts*
 - * *Using RegisterAnalysisGroup*
- *Pass Statistics*
 - *What PassManager does*
 - * *The releaseMemory method*
- *Registering dynamically loaded passes*
 - *Using existing registries*
 - *Creating new registries*
 - *Using GDB with dynamically loaded passes*
 - * *Setting a breakpoint in your pass*
 - * *Miscellaneous Problems*
 - *Future extensions planned*
 - * *Multithreaded LLVM*

4.21.1 Introduction — What is a pass?

The LLVM Pass Framework is an important part of the LLVM system, because LLVM passes are where most of the interesting parts of the compiler exist. Passes perform the transformations and optimizations that make up the compiler, they build the analysis results that are used by these transformations, and they are, above all, a structuring technique for compiler code.

All LLVM passes are subclasses of the `Pass` class, which implement functionality by overriding virtual methods inherited from `Pass`. Depending on how your pass works, you should inherit from the `ModulePass`, `CallGraphSCCPass`, `FunctionPass`, or `LoopPass`, or `RegionPass`, or `BasicBlockPass` classes, which gives the system more information about what your pass does, and how it can be combined with other passes. One of the main features of the LLVM Pass Framework is that it schedules passes to run in an efficient way based on the constraints that your pass meets (which are indicated by which class they derive from).

We start by showing you how to construct a pass, everything from setting up the code, to compiling, loading, and executing it. After the basics are down, more advanced features are discussed.

4.21.2 Quick Start — Writing hello world

Here we describe how to write the “hello world” of passes. The “Hello” pass is designed to simply print out the name of non-external functions that exist in the program being compiled. It does not modify the program at all, it just inspects it. The source code and files for this pass are available in the LLVM source tree in the `lib/Transforms/Hello` directory.

Setting up the build environment

First, configure and build LLVM. Next, you need to create a new directory somewhere in the LLVM source base. For this example, we’ll assume that you made `lib/Transforms/Hello`. Finally, you must set up a build script that will compile the source code for the new pass. To do this, copy the following into `CMakeLists.txt`:

```
add_llvm_loadable_module( LLVMHello
    Hello.cpp

    PLUGIN_TOOL
    opt
)
```

and the following line into `lib/Transforms/CMakeLists.txt`:

```
add_subdirectory(Hello)
```

(Note that there is already a directory named `Hello` with a sample “Hello” pass; you may play with it – in which case you don’t need to modify any `CMakeLists.txt` files – or, if you want to create everything from scratch, use another name.)

This build script specifies that `Hello.cpp` file in the current directory is to be compiled and linked into a shared object `$(LEVEL)/lib/LLVMHello.so` that can be dynamically loaded by the `opt` tool via its `-load` option. If your operating system uses a suffix other than `.so` (such as Windows or Mac OS X), the appropriate extension will be used.

Now that we have the build scripts set up, we just need to write the code for the pass itself.

Basic code required

Now that we have a way to compile our new pass, we just have to write it. Start out with:

```
#include "llvm/Pass.h"
#include "llvm/IR/Function.h"
#include "llvm/Support/raw_ostream.h"
```

Which are needed because we are writing a `Pass`, we are operating on `Functions`, and we will be doing some printing.

Next we have:

```
using namespace llvm;
```

... which is required because the functions from the include files live in the `llvm` namespace.

Next we have:

```
namespace {
```

... which starts out an anonymous namespace. Anonymous namespaces are to C++ what the “`static`” keyword is to C (at global scope). It makes the things declared inside of the anonymous namespace visible only to the current file. If you’re not familiar with them, consult a decent C++ book for more information.

Next, we declare our pass itself:

```
struct Hello : public FunctionPass {
```

This declares a “`Hello`” class that is a subclass of `FunctionPass`. The different builtin pass subclasses are described in detail *later*, but for now, know that `FunctionPass` operates on a function at a time.

```
static char ID;
Hello() : FunctionPass(ID) {}
```

This declares pass identifier used by LLVM to identify pass. This allows LLVM to avoid using expensive C++ runtime information.

```
bool runOnFunction(Function &F) override {
    errs() << "Hello: ";
    errs().write_escaped(F.getName()) << '\n';
    return false;
}
}; // end of struct Hello
} // end of anonymous namespace
```

We declare a `runOnFunction` method, which overrides an abstract virtual method inherited from `FunctionPass`. This is where we are supposed to do our thing, so we just print out our message with the name of each function.

```
char Hello::ID = 0;
```

We initialize pass ID here. LLVM uses ID’s address to identify a pass, so initialization value is not important.

```
static RegisterPass<Hello> X("hello", "Hello World Pass",
                             false /* Only looks at CFG */,
                             false /* Analysis Pass */);
```

Lastly, we *register our class* `Hello`, giving it a command line argument “`hello`”, and a name “`Hello World Pass`”. The last two arguments describe its behavior: if a pass walks CFG without modifying it then the third argument is set to `true`; if a pass is an analysis pass, for example dominator tree pass, then `true` is supplied as the fourth argument.

As a whole, the `.cpp` file looks like:

```
#include "llvm/Pass.h"
#include "llvm/IR/Function.h"
#include "llvm/Support/raw_ostream.h"

using namespace llvm;

namespace {
struct Hello : public FunctionPass {
    static char ID;
    Hello() : FunctionPass(ID) {}

    bool runOnFunction(Function &F) override {
        errs() << "Hello: ";
        errs().write_escaped(F.getName()) << '\n';
        return false;
    }
}; // end of struct Hello
} // end of anonymous namespace

char Hello::ID = 0;
static RegisterPass<Hello> X("hello", "Hello World Pass",
                             false /* Only looks at CFG */,
                             false /* Analysis Pass */);
```

Now that it's all together, compile the file with a simple “gmake” command from the top level of your build directory and you should get a new file “lib/LLVMHello.so”. Note that everything in this file is contained in an anonymous namespace — this reflects the fact that passes are self contained units that do not need external interfaces (although they can have them) to be useful.

Running a pass with **opt**

Now that you have a brand new shiny shared object file, we can use the **opt** command to run an LLVM program through your pass. Because you registered your pass with `RegisterPass`, you will be able to use the **opt** tool to access it, once loaded.

To test it, follow the example at the end of the *Getting Started with the LLVM System* to compile “Hello World” to LLVM. We can now run the bitcode file (hello.bc) for the program through our transformation like this (or course, any bitcode file will work):

```
$ opt -load lib/LLVMHello.so -hello < hello.bc > /dev/null
Hello: __main
Hello: puts
Hello: main
```

The `-load` option specifies that **opt** should load your pass as a shared object, which makes “-hello” a valid command line argument (which is one reason you need to *register your pass*). Because the Hello pass does not modify the program in any interesting way, we just throw away the result of **opt** (sending it to `/dev/null`).

To see what happened to the other string you registered, try running **opt** with the `-help` option:

```
$ opt -load lib/LLVMHello.so -help
OVERVIEW: llvm .bc -> .bc modular optimizer and analysis printer

USAGE: opt [subcommand] [options] <input bitcode file>

OPTIONS:
    Optimizations available:
```

```
...
-guard-widening      - Widen guards
-gvn                  - Global Value Numbering
-gvn-hoist            - Early GVN Hoisting of Expressions
-hello                - Hello World Pass
-indvars              - Induction Variable Simplification
-inferattrs           - Infer set function attributes
...
```

The pass name gets added as the information string for your pass, giving some documentation to users of `opt`. Now that you have a working pass, you would go ahead and make it do the cool transformations you want. Once you get it all working and tested, it may become useful to find out how fast your pass is. The *PassManager* provides a nice command line option (`--time-passes`) that allows you to get information about the execution time of your pass along with the other passes you queue up. For example:

```
$ opt -load lib/LLVMHello.so -hello -time-passes < hello.bc > /dev/null
Hello: __main
Hello: puts
Hello: main
=====
... Pass execution timing report ...
=====
Total Execution Time: 0.0007 seconds (0.0005 wall clock)

---User Time---  --User+System--  ---Wall Time---  --- Name ---
0.0004 ( 55.3%)  0.0004 ( 55.3%)  0.0004 ( 75.7%)  Bitcode Writer
0.0003 ( 44.7%)  0.0003 ( 44.7%)  0.0001 ( 13.6%)  Hello World Pass
0.0000 (  0.0%)  0.0000 (  0.0%)  0.0001 ( 10.7%)  Module Verifier
0.0007 (100.0%)  0.0007 (100.0%)  0.0005 (100.0%)  Total
```

As you can see, our implementation above is pretty fast. The additional passes listed are automatically inserted by the `opt` tool to verify that the LLVM emitted by your pass is still valid and well formed LLVM, which hasn't been broken somehow.

Now that you have seen the basics of the mechanics behind passes, we can talk about some more details of how they work and how to use them.

4.21.3 Pass classes and requirements

One of the first things that you should do when designing a new pass is to decide what class you should subclass for your pass. The *Hello World* example uses the *FunctionPass* class for its implementation, but we did not discuss why or when this should occur. Here we talk about the classes available, from the most general to the most specific.

When choosing a superclass for your Pass, you should choose the **most specific** class possible, while still being able to meet the requirements listed. This gives the LLVM Pass Infrastructure information necessary to optimize how passes are run, so that the resultant compiler isn't unnecessarily slow.

The `ImmutablePass` class

The most plain and boring type of pass is the “`ImmutablePass`” class. This pass type is used for passes that do not have to be run, do not change state, and never need to be updated. This is not a normal type of transformation or analysis, but can provide information about the current compiler configuration.

Although this pass class is very infrequently used, it is important for providing information about the current target machine being compiled for, and other static information that can affect the various transformations.

`ImmutablePasses` never invalidate other transformations, are never invalidated, and are never “run”.

The `ModulePass` class

The `ModulePass` class is the most general of all superclasses that you can use. Deriving from `ModulePass` indicates that your pass uses the entire program as a unit, referring to function bodies in no predictable order, or adding and removing functions. Because nothing is known about the behavior of `ModulePass` subclasses, no optimization can be done for their execution.

A module pass can use function level passes (e.g. dominators) using the `getAnalysis` interface `getAnalysis<DominatorTree>(llvm::Function *)` to provide the function to retrieve analysis result for, if the function pass does not require any module or immutable passes. Note that this can only be done for functions for which the analysis ran, e.g. in the case of dominators you should only ask for the `DominatorTree` for function definitions, not declarations.

To write a correct `ModulePass` subclass, derive from `ModulePass` and overload the `runOnModule` method with the following signature:

The `runOnModule` method

```
virtual bool runOnModule(Module &M) = 0;
```

The `runOnModule` method performs the interesting work of the pass. It should return `true` if the module was modified by the transformation and `false` otherwise.

The `CallGraphSCCPass` class

The `CallGraphSCCPass` is used by passes that need to traverse the program bottom-up on the call graph (callees before callers). Deriving from `CallGraphSCCPass` provides some mechanics for building and traversing the `CallGraph`, but also allows the system to optimize execution of `CallGraphSCCPasses`. If your pass meets the requirements outlined below, and doesn’t meet the requirements of a *FunctionPass* or *BasicBlockPass*, you should derive from `CallGraphSCCPass`.

TODO: explain briefly what SCC, Tarjan’s algo, and B-U mean.

To be explicit, `CallGraphSCCPass` subclasses are:

1. ... *not allowed* to inspect or modify any `Functions` other than those in the current SCC and the direct callers and direct callees of the SCC.
2. ... *required* to preserve the current `CallGraph` object, updating it to reflect any changes made to the program.
3. ... *not allowed* to add or remove SCC’s from the current `Module`, though they may change the contents of an SCC.
4. ... *allowed* to add or remove global variables from the current `Module`.
5. ... *allowed* to maintain state across invocations of *runOnSCC* (including global data).

Implementing a `CallGraphSCCPass` is slightly tricky in some cases because it has to handle SCCs with more than one node in it. All of the virtual methods described below should return `true` if they modified the program, or `false` if they didn’t.

The `doInitialization(CallGraph &)` method

```
virtual bool doInitialization(CallGraph &CG);
```

The `doInitialization` method is allowed to do most of the things that `CallGraphSCCPasses` are not allowed to do. They can add and remove functions, get pointers to functions, etc. The `doInitialization` method is designed to do simple initialization type of stuff that does not depend on the SCCs being processed. The `doInitialization` method call is not scheduled to overlap with any other pass executions (thus it should be very fast).

The `runOnSCC` method

```
virtual bool runOnSCC(CallGraphSCC &SCC) = 0;
```

The `runOnSCC` method performs the interesting work of the pass, and should return `true` if the module was modified by the transformation, `false` otherwise.

The `doFinalization(CallGraph &)` method

```
virtual bool doFinalization(CallGraph &CG);
```

The `doFinalization` method is an infrequently used method that is called when the pass framework has finished calling `runOnSCC` for every SCC in the program being compiled.

The `FunctionPass` class

In contrast to `ModulePass` subclasses, `FunctionPass` subclasses do have a predictable, local behavior that can be expected by the system. All `FunctionPass` execute on each function in the program independent of all of the other functions in the program. `FunctionPasses` do not require that they are executed in a particular order, and `FunctionPasses` do not modify external functions.

To be explicit, `FunctionPass` subclasses are not allowed to:

1. Inspect or modify a `Function` other than the one currently being processed.
2. Add or remove `Functions` from the current `Module`.
3. Add or remove global variables from the current `Module`.
4. Maintain state across invocations of `runOnFunction` (including global data).

Implementing a `FunctionPass` is usually straightforward (See the *Hello World* pass for example). `FunctionPasses` may overload three virtual methods to do their work. All of these methods should return `true` if they modified the program, or `false` if they didn't.

The `doInitialization(Module &)` method

```
virtual bool doInitialization(Module &M);
```

The `doInitialization` method is allowed to do most of the things that `FunctionPasses` are not allowed to do. They can add and remove functions, get pointers to functions, etc. The `doInitialization` method is designed to do simple initialization type of stuff that does not depend on the functions being processed. The

`doInitialization` method call is not scheduled to overlap with any other pass executions (thus it should be very fast).

A good example of how this method should be used is the [LowerAllocations](#) pass. This pass converts `malloc` and `free` instructions into platform dependent `malloc()` and `free()` function calls. It uses the `doInitialization` method to get a reference to the `malloc` and `free` functions that it needs, adding prototypes to the module if necessary.

The `runOnFunction` method

```
virtual bool runOnFunction(Function &F) = 0;
```

The `runOnFunction` method must be implemented by your subclass to do the transformation or analysis work of your pass. As usual, a `true` value should be returned if the function is modified.

The `doFinalization(Module &)` method

```
virtual bool doFinalization(Module &M);
```

The `doFinalization` method is an infrequently used method that is called when the pass framework has finished calling *`runOnFunction`* for every function in the program being compiled.

The `LoopPass` class

All `LoopPass` execute on each loop in the function independent of all of the other loops in the function. `LoopPass` processes loops in loop nest order such that outer most loop is processed last.

`LoopPass` subclasses are allowed to update loop nest using `LPPassManager` interface. Implementing a loop pass is usually straightforward. `LoopPasses` may overload three virtual methods to do their work. All these methods should return `true` if they modified the program, or `false` if they didn't.

A `LoopPass` subclass which is intended to run as part of the main loop pass pipeline needs to preserve all of the same *function* analyses that the other loop passes in its pipeline require. To make that easier, a `getLoopAnalysisUsage` function is provided by `LoopUtils.h`. It can be called within the subclass's `getAnalysisUsage` override to get consistent and correct behavior. Analogously, `INITIALIZE_PASS_DEPENDENCY(LoopPass)` will initialize this set of function analyses.

The `doInitialization(Loop *, LPPassManager &)` method

```
virtual bool doInitialization(Loop *, LPPassManager &LPM);
```

The `doInitialization` method is designed to do simple initialization type of stuff that does not depend on the functions being processed. The `doInitialization` method call is not scheduled to overlap with any other pass executions (thus it should be very fast). `LPPassManager` interface should be used to access `Function` or `Module` level analysis information.

The `runOnLoop` method

```
virtual bool runOnLoop(Loop *, LPPassManager &LPM) = 0;
```


The `runOnLoop` method must be implemented by your subclass to do the transformation or analysis work of your pass. As usual, a `true` value should be returned if the function is modified. `LPPassManager` interface should be used to update loop nest.

The `doFinalization()` method

```
virtual bool doFinalization();
```

The `doFinalization` method is an infrequently used method that is called when the pass framework has finished calling *runOnLoop* for every loop in the program being compiled.

The `RegionPass` class

`RegionPass` is similar to *LoopPass*, but executes on each single entry single exit region in the function. `RegionPass` processes regions in nested order such that the outer most region is processed last.

`RegionPass` subclasses are allowed to update the region tree by using the `RGPassManager` interface. You may overload three virtual methods of `RegionPass` to implement your own region pass. All these methods should return `true` if they modified the program, or `false` if they did not.

The `doInitialization(Region *, RGPassManager &)` method

```
virtual bool doInitialization(Region *, RGPassManager &RGM);
```

The `doInitialization` method is designed to do simple initialization type of stuff that does not depend on the functions being processed. The `doInitialization` method call is not scheduled to overlap with any other pass executions (thus it should be very fast). `RPPassManager` interface should be used to access `Function` or `Module` level analysis information.

The `runOnRegion` method

```
virtual bool runOnRegion(Region *, RGPassManager &RGM) = 0;
```

The `runOnRegion` method must be implemented by your subclass to do the transformation or analysis work of your pass. As usual, a `true` value should be returned if the region is modified. `RGPassManager` interface should be used to update region tree.

The `doFinalization()` method

```
virtual bool doFinalization();
```

The `doFinalization` method is an infrequently used method that is called when the pass framework has finished calling *runOnRegion* for every region in the program being compiled.

The `BasicBlockPass` class

`BasicBlockPasses` are just like *FunctionPass's*, except that they must limit their scope of inspection and modification to a single basic block at a time. As such, they are **not** allowed to do any of the following:

1. Modify or inspect any basic blocks outside of the current one.
2. Maintain state across invocations of *runOnBasicBlock*.
3. Modify the control flow graph (by altering terminator instructions)
4. Any of the things forbidden for *FunctionPasses*.

BasicBlockPasses are useful for traditional local and “peephole” optimizations. They may override the same *doInitialization(Module &)* and *doFinalization(Module &)* methods that *FunctionPass*’s have, but also have the following virtual methods that may also be implemented:

The *doInitialization(Function &)* method

```
virtual bool doInitialization(Function &F);
```

The *doInitialization* method is allowed to do most of the things that *BasicBlockPasses* are not allowed to do, but that *FunctionPasses* can. The *doInitialization* method is designed to do simple initialization that does not depend on the *BasicBlocks* being processed. The *doInitialization* method call is not scheduled to overlap with any other pass executions (thus it should be very fast).

The *runOnBasicBlock* method

```
virtual bool runOnBasicBlock(BasicBlock &BB) = 0;
```

Override this function to do the work of the *BasicBlockPass*. This function is not allowed to inspect or modify basic blocks other than the parameter, and are not allowed to modify the CFG. A `true` value must be returned if the basic block is modified.

The *doFinalization(Function &)* method

```
virtual bool doFinalization(Function &F);
```

The *doFinalization* method is an infrequently used method that is called when the pass framework has finished calling *runOnBasicBlock* for every *BasicBlock* in the program being compiled. This can be used to perform per-function finalization.

The *MachineFunctionPass* class

A *MachineFunctionPass* is a part of the LLVM code generator that executes on the machine-dependent representation of each LLVM function in the program.

Code generator passes are registered and initialized specially by `TargetMachine::addPassesToEmitFile` and similar routines, so they cannot generally be run from the **opt** or **bugpoint** commands.

A *MachineFunctionPass* is also a *FunctionPass*, so all the restrictions that apply to a *FunctionPass* also apply to it. *MachineFunctionPasses* also have additional restrictions. In particular, *MachineFunctionPasses* are not allowed to do any of the following:

1. Modify or create any LLVM IR Instructions, BasicBlocks, Arguments, Functions, GlobalVariables, GlobalAliases, or Modules.
2. Modify a *MachineFunction* other than the one currently being processed.

3. Maintain state across invocations of *runOnMachineFunction* (including global data).

The `runOnMachineFunction(MachineFunction &MF)` method

```
virtual bool runOnMachineFunction(MachineFunction &MF) = 0;
```

`runOnMachineFunction` can be considered the main entry point of a `MachineFunctionPass`; that is, you should override this method to do the work of your `MachineFunctionPass`.

The `runOnMachineFunction` method is called on every `MachineFunction` in a `Module`, so that the `MachineFunctionPass` may perform optimizations on the machine-dependent representation of the function. If you want to get at the LLVM Function for the `MachineFunction` you're working on, use `MachineFunction`'s `getFunction()` accessor method — but remember, you may not modify the LLVM Function or its contents from a `MachineFunctionPass`.

Pass registration

In the *Hello World* example pass we illustrated how pass registration works, and discussed some of the reasons that it is used and what it does. Here we discuss how and why passes are registered.

As we saw above, passes are registered with the `RegisterPass` template. The template parameter is the name of the pass that is to be used on the command line to specify that the pass should be added to a program (for example, with `opt` or `bugpoint`). The first argument is the name of the pass, which is to be used for the `-help` output of programs, as well as for debug output generated by the `-debug-pass` option.

If you want your pass to be easily dumpable, you should implement the virtual print method:

The print method

```
virtual void print(llvm::raw_ostream &O, const Module *M) const;
```

The `print` method must be implemented by “analyses” in order to print a human readable version of the analysis results. This is useful for debugging an analysis itself, as well as for other people to figure out how an analysis works. Use the `opt -analyze` argument to invoke this method.

The `llvm::raw_ostream` parameter specifies the stream to write the results on, and the `Module` parameter gives a pointer to the top level module of the program that has been analyzed. Note however that this pointer may be `NULL` in certain circumstances (such as calling the `Pass::dump()` from a debugger), so it should only be used to enhance debug output, it should not be depended on.

Specifying interactions between passes

One of the main responsibilities of the `PassManager` is to make sure that passes interact with each other correctly. Because `PassManager` tries to *optimize the execution of passes* it must know how the passes interact with each other and what dependencies exist between the various passes. To track this, each pass can declare the set of passes that are required to be executed before the current pass, and the passes which are invalidated by the current pass.

Typically this functionality is used to require that analysis results are computed before your pass is run. Running arbitrary transformation passes can invalidate the computed analysis results, which is what the invalidation set specifies. If a pass does not implement the *getAnalysisUsage* method, it defaults to not having any prerequisite passes, and invalidating **all** other passes.

The `getAnalysisUsage` method

```
virtual void getAnalysisUsage(AnalysisUsage &Info) const;
```

By implementing the `getAnalysisUsage` method, the required and invalidated sets may be specified for your transformation. The implementation should fill in the `AnalysisUsage` object with information about which passes are required and not invalidated. To do this, a pass may call any of the following methods on the `AnalysisUsage` object:

The `AnalysisUsage::addRequired<>` and `AnalysisUsage::addRequiredTransitive<>` methods

If your pass requires a previous pass to be executed (an analysis for example), it can use one of these methods to arrange for it to be run before your pass. LLVM has many different types of analyses and passes that can be required, spanning the range from `DominatorSet` to `BreakCriticalEdges`. Requiring `BreakCriticalEdges`, for example, guarantees that there will be no critical edges in the CFG when your pass has been run.

Some analyses chain to other analyses to do their job. For example, an *AliasAnalysis* <*AliasAnalysis*> implementation is required to *chain* to other alias analysis passes. In cases where analyses chain, the `addRequiredTransitive` method should be used instead of the `addRequired` method. This informs the `PassManager` that the transitively required pass should be alive as long as the requiring pass is.

The `AnalysisUsage::addPreserved<>` method

One of the jobs of the `PassManager` is to optimize how and when analyses are run. In particular, it attempts to avoid recomputing data unless it needs to. For this reason, passes are allowed to declare that they preserve (i.e., they don't invalidate) an existing analysis if it's available. For example, a simple constant folding pass would not modify the CFG, so it can't possibly affect the results of dominator analysis. By default, all passes are assumed to invalidate all others.

The `AnalysisUsage` class provides several methods which are useful in certain circumstances that are related to `addPreserved`. In particular, the `setPreservesAll` method can be called to indicate that the pass does not modify the LLVM program at all (which is true for analyses), and the `setPreservesCFG` method can be used by transformations that change instructions in the program but do not modify the CFG or terminator instructions (note that this property is implicitly set for *BasicBlockPasses*).

`addPreserved` is particularly useful for transformations like `BreakCriticalEdges`. This pass knows how to update a small set of loop and dominator related analyses if they exist, so it can preserve them, despite the fact that it hacks on the CFG.

Example implementations of `getAnalysisUsage`

```
// This example modifies the program, but does not modify the CFG
void LICM::getAnalysisUsage(AnalysisUsage &AU) const {
    AU.setPreservesCFG();
    AU.addRequired<LoopInfoWrapperPass>();
}
```

The `getAnalysis<>` and `getAnalysisIfAvailable<>` methods

The `Pass::getAnalysis<>` method is automatically inherited by your class, providing you with access to the passes that you declared that you required with the `getAnalysisUsage` method. It takes a single template argument that specifies which pass class you want, and returns a reference to that pass. For example:

```
bool LICM::runOnFunction(Function &F) {
    LoopInfo &LI = getAnalysis<LoopInfoWrapperPass>().getLoopInfo();
    //...
}
```

This method call returns a reference to the pass desired. You may get a runtime assertion failure if you attempt to get an analysis that you did not declare as required in your `getAnalysisUsage` implementation. This method can be called by your `run*` method implementation, or by any other local method invoked by your `run*` method.

A module level pass can use function level analysis info using this interface. For example:

```
bool ModuleLevelPass::runOnModule(Module &M) {
    //...
    DominatorTree &DT = getAnalysis<DominatorTree>(Func);
    //...
}
```

In above example, `runOnFunction` for `DominatorTree` is called by pass manager before returning a reference to the desired pass.

If your pass is capable of updating analyses if they exist (e.g., `BreakCriticalEdges`, as described above), you can use the `getAnalysisIfAvailable` method, which returns a pointer to the analysis if it is active. For example:

```
if (DominatorSet *DS = getAnalysisIfAvailable<DominatorSet>()) {
    // A DominatorSet is active. This code will update it.
}
```

Implementing Analysis Groups

Now that we understand the basics of how passes are defined, how they are used, and how they are required from other passes, it's time to get a little bit fancier. All of the pass relationships that we have seen so far are very simple: one pass depends on one other specific pass to be run before it can run. For many applications, this is great, for others, more flexibility is required.

In particular, some analyses are defined such that there is a single simple interface to the analysis results, but multiple ways of calculating them. Consider alias analysis for example. The most trivial alias analysis returns “may alias” for any alias query. The most sophisticated analysis a flow-sensitive, context-sensitive interprocedural analysis that can take a significant amount of time to execute (and obviously, there is a lot of room between these two extremes for other implementations). To cleanly support situations like this, the LLVM Pass Infrastructure supports the notion of Analysis Groups.

Analysis Group Concepts

An Analysis Group is a single simple interface that may be implemented by multiple different passes. Analysis Groups can be given human readable names just like passes, but unlike passes, they need not derive from the `Pass` class. An analysis group may have one or more implementations, one of which is the “default” implementation.

Analysis groups are used by client passes just like other passes are: the `AnalysisUsage::addRequired()` and `Pass::getAnalysis()` methods. In order to resolve this requirement, the `PassManager` scans the available

passes to see if any implementations of the analysis group are available. If none is available, the default implementation is created for the pass to use. All standard rules for *interaction between passes* still apply.

Although *Pass Registration* is optional for normal passes, all analysis group implementations must be registered, and must use the `INITIALIZE_AG_PASS` template to join the implementation pool. Also, a default implementation of the interface **must** be registered with *RegisterAnalysisGroup*.

As a concrete example of an Analysis Group in action, consider the *AliasAnalysis* analysis group. The default implementation of the alias analysis interface (the *basicaa* pass) just does a few simple checks that don't require significant analysis to compute (such as: two different globals can never alias each other, etc). Passes that use the *AliasAnalysis* interface (for example the *gvn* pass), do not care which implementation of alias analysis is actually provided, they just use the designated interface.

From the user's perspective, commands work just like normal. Issuing the command `opt -gvn ...` will cause the *basicaa* class to be instantiated and added to the pass sequence. Issuing the command `opt -somefancyaa -gvn ...` will cause the *gvn* pass to use the *somefancyaa* alias analysis (which doesn't actually exist, it's just a hypothetical example) instead.

Using RegisterAnalysisGroup

The *RegisterAnalysisGroup* template is used to register the analysis group itself, while the `INITIALIZE_AG_PASS` is used to add pass implementations to the analysis group. First, an analysis group should be registered, with a human readable name provided for it. Unlike registration of passes, there is no command line argument to be specified for the Analysis Group Interface itself, because it is "abstract":

```
static RegisterAnalysisGroup<AliasAnalysis> A("Alias Analysis");
```

Once the analysis is registered, passes can declare that they are valid implementations of the interface by using the following code:

```
namespace {
  // Declare that we implement the AliasAnalysis interface
  INITIALIZE_AG_PASS(FancyAA, AliasAnalysis, "somefancyaa",
    "A more complex alias analysis implementation",
    false, // Is CFG Only?
    true,  // Is Analysis?
    false); // Is default Analysis Group implementation?
}
```

This just shows a class *FancyAA* that uses the `INITIALIZE_AG_PASS` macro both to register and to "join" the *AliasAnalysis* analysis group. Every implementation of an analysis group should join using this macro.

```
namespace {
  // Declare that we implement the AliasAnalysis interface
  INITIALIZE_AG_PASS(BasicAA, AliasAnalysis, "basicaa",
    "Basic Alias Analysis (default AA impl)",
    false, // Is CFG Only?
    true,  // Is Analysis?
    true); // Is default Analysis Group implementation?
}
```

Here we show how the default implementation is specified (using the final argument to the `INITIALIZE_AG_PASS` template). There must be exactly one default implementation available at all times for an Analysis Group to be used. Only default implementation can derive from *ImmutablePass*. Here we declare that the *BasicAliasAnalysis* pass is the default implementation for the interface.

4.21.4 Pass Statistics

The `Statistic` class is designed to be an easy way to expose various success metrics from passes. These statistics are printed at the end of a run, when the `-stats` command line option is enabled on the command line. See the [Statistics section](#) in the Programmer’s Manual for details.

What PassManager does

The `PassManager` class takes a list of passes, ensures their *prerequisites* are set up correctly, and then schedules passes to run efficiently. All of the LLVM tools that run passes use the `PassManager` for execution of these passes.

The `PassManager` does two main things to try to reduce the execution time of a series of passes:

1. **Share analysis results.** The `PassManager` attempts to avoid recomputing analysis results as much as possible. This means keeping track of which analyses are available already, which analyses get invalidated, and which analyses are needed to be run for a pass. An important part of work is that the `PassManager` tracks the exact lifetime of all analysis results, allowing it to *free memory* allocated to holding analysis results as soon as they are no longer needed.
2. **Pipeline the execution of passes on the program.** The `PassManager` attempts to get better cache and memory usage behavior out of a series of passes by pipelining the passes together. This means that, given a series of consecutive `FunctionPass`, it will execute all of the `FunctionPass` on the first function, then all of the `FunctionPasses` on the second function, etc. . . until the entire program has been run through the passes.

This improves the cache behavior of the compiler, because it is only touching the LLVM program representation for a single function at a time, instead of traversing the entire program. It reduces the memory consumption of compiler, because, for example, only one `DominatorSet` needs to be calculated at a time. This also makes it possible to implement some *interesting enhancements* in the future.

The effectiveness of the `PassManager` is influenced directly by how much information it has about the behaviors of the passes it is scheduling. For example, the “preserved” set is intentionally conservative in the face of an unimplemented `getAnalysisUsage` method. Not implementing when it should be implemented will have the effect of not allowing any analysis results to live across the execution of your pass.

The `PassManager` class exposes a `--debug-pass` command line options that is useful for debugging pass execution, seeing how things work, and diagnosing when you should be preserving more analyses than you currently are. (To get information about all of the variants of the `--debug-pass` option, just type “`opt -help-hidden`”).

By using the `--debug-pass=Structure` option, for example, we can see how our *Hello World* pass interacts with other passes. Lets try it out with the `gvn` and `licm` passes:

```
$ opt -load lib/LLVMHello.so -gvn -licm --debug-pass=Structure < hello.bc > /dev/null
ModulePass Manager
  FunctionPass Manager
    Dominator Tree Construction
    Basic Alias Analysis (stateless AA impl)
    Function Alias Analysis Results
    Memory Dependence Analysis
    Global Value Numbering
    Natural Loop Information
    Canonicalize natural loops
    Loop-Closed SSA Form Pass
    Basic Alias Analysis (stateless AA impl)
    Function Alias Analysis Results
    Scalar Evolution Analysis
    Loop Pass Manager
      Loop Invariant Code Motion
```

```
Module Verifier
Bitcode Writer
```

This output shows us when passes are constructed. Here we see that GVN uses dominator tree information to do its job. The LICM pass uses natural loop information, which uses dominator tree as well.

After the LICM pass, the module verifier runs (which is automatically added by the **opt** tool), which uses the dominator tree to check that the resultant LLVM code is well formed. Note that the dominator tree is computed once, and shared by three passes.

Lets see how this changes when we run the *Hello World* pass in between the two passes:

```
$ opt -load lib/LLVMHello.so -gvn -hello -licm --debug-pass=Structure < hello.bc > /
↳dev/null
ModulePass Manager
  FunctionPass Manager
    Dominator Tree Construction
    Basic Alias Analysis (stateless AA impl)
    Function Alias Analysis Results
    Memory Dependence Analysis
    Global Value Numbering
    Hello World Pass
    Dominator Tree Construction
    Natural Loop Information
    Canonicalize natural loops
    Loop-Closed SSA Form Pass
    Basic Alias Analysis (stateless AA impl)
    Function Alias Analysis Results
    Scalar Evolution Analysis
    Loop Pass Manager
      Loop Invariant Code Motion
    Module Verifier
    Bitcode Writer
Hello: __main
Hello: puts
Hello: main
```

Here we see that the *Hello World* pass has killed the Dominator Tree pass, even though it doesn't modify the code at all! To fix this, we need to add the following *getAnalysisUsage* method to our pass:

```
// We don't modify the program, so we preserve all analyses
void getAnalysisUsage(AnalysisUsage &AU) const override {
    AU.setPreservesAll();
}
```

Now when we run our pass, we get this output:

```
$ opt -load lib/LLVMHello.so -gvn -hello -licm --debug-pass=Structure < hello.bc > /
↳dev/null
Pass Arguments: -gvn -hello -licm
ModulePass Manager
  FunctionPass Manager
    Dominator Tree Construction
    Basic Alias Analysis (stateless AA impl)
    Function Alias Analysis Results
    Memory Dependence Analysis
    Global Value Numbering
    Hello World Pass
```



```

    Natural Loop Information
    Canonicalize natural loops
    Loop-Closed SSA Form Pass
    Basic Alias Analysis (stateless AA impl)
    Function Alias Analysis Results
    Scalar Evolution Analysis
    Loop Pass Manager
      Loop Invariant Code Motion
    Module Verifier
    Bitcode Writer
Hello: __main
Hello: puts
Hello: main

```

Which shows that we don't accidentally invalidate dominator information anymore, and therefore do not have to compute it twice.

The `releaseMemory` method

```
virtual void releaseMemory();
```

The `PassManager` automatically determines when to compute analysis results, and how long to keep them around for. Because the lifetime of the pass object itself is effectively the entire duration of the compilation process, we need some way to free analysis results when they are no longer useful. The `releaseMemory` virtual method is the way to do this.

If you are writing an analysis or any other pass that retains a significant amount of state (for use by another pass which “requires” your pass and uses the `getAnalysis` method) you should implement `releaseMemory` to, well, release the memory allocated to maintain this internal state. This method is called after the `run*` method for the class, before the next call of `run*` in your pass.

4.21.5 Registering dynamically loaded passes

Size matters when constructing production quality tools using LLVM, both for the purposes of distribution, and for regulating the resident code size when running on the target system. Therefore, it becomes desirable to selectively use some passes, while omitting others and maintain the flexibility to change configurations later on. You want to be able to do all this, and, provide feedback to the user. This is where pass registration comes into play.

The fundamental mechanisms for pass registration are the `MachinePassRegistry` class and subclasses of `MachinePassRegistryNode`.

An instance of `MachinePassRegistry` is used to maintain a list of `MachinePassRegistryNode` objects. This instance maintains the list and communicates additions and deletions to the command line interface.

An instance of `MachinePassRegistryNode` subclass is used to maintain information provided about a particular pass. This information includes the command line name, the command help string and the address of the function used to create an instance of the pass. A global static constructor of one of these instances *registers* with a corresponding `MachinePassRegistry`, the static destructor *unregisters*. Thus a pass that is statically linked in the tool will be registered at start up. A dynamically loaded pass will register on load and unregister at unload.

Using existing registries

There are predefined registries to track instruction scheduling (`RegisterScheduler`) and register allocation (`RegisterRegAlloc`) machine passes. Here we will describe how to *register* a register allocator machine pass.

Implement your register allocator machine pass. In your register allocator .cpp file add the following include:

```
#include "llvm/CodeGen/RegAllocRegistry.h"
```

Also in your register allocator .cpp file, define a creator function in the form:

```
FunctionPass *createMyRegisterAllocator() {  
    return new MyRegisterAllocator();  
}
```

Note that the signature of this function should match the type of `RegisterRegAlloc::FunctionPassCtor`. In the same file add the “installing” declaration, in the form:

```
static RegisterRegAlloc myRegAlloc("myregalloc",  
                                   "my register allocator help string",  
                                   createMyRegisterAllocator);
```

Note the two spaces prior to the help string produces a tidy result on the `-help` query.

```
$ llc -help  
...  
-regalloc                - Register allocator to use (default=linearscan)  
  =linearscan            - linear scan register allocator  
  =local                  - local register allocator  
  =simple                  - simple register allocator  
  =myregalloc             - my register allocator help string  
...
```

And that’s it. The user is now free to use `-regalloc=myregalloc` as an option. Registering instruction schedulers is similar except use the `RegisterScheduler` class. Note that the `RegisterScheduler::FunctionPassCtor` is significantly different from `RegisterRegAlloc::FunctionPassCtor`.

To force the load/linking of your register allocator into the `llc/lli` tools, add your creator function’s global declaration to `Passes.h` and add a “pseudo” call line to `llvm/CodeGen/LinkAllCodegenComponents.h`.

Creating new registries

The easiest way to get started is to clone one of the existing registries; we recommend `llvm/CodeGen/RegAllocRegistry.h`. The key things to modify are the class name and the `FunctionPassCtor` type.

Then you need to declare the registry. Example: if your pass registry is `RegisterMyPasses` then define:

```
MachinePassRegistry RegisterMyPasses::Registry;
```

And finally, declare the command line option for your passes. Example:

```
cl::opt<RegisterMyPasses::FunctionPassCtor, false,  
      RegisterPassParser<RegisterMyPasses> >  
MyPassOpt("mypass",  
          cl::init(&createDefaultMyPass),  
          cl::desc("my pass option help"));
```

Here the command option is “mypass”, with `createDefaultMyPass` as the default creator.

Using GDB with dynamically loaded passes

Unfortunately, using GDB with dynamically loaded passes is not as easy as it should be. First of all, you can't set a breakpoint in a shared object that has not been loaded yet, and second of all there are problems with inlined functions in shared objects. Here are some suggestions to debugging your pass with GDB.

For sake of discussion, I'm going to assume that you are debugging a transformation invoked by `opt`, although nothing described here depends on that.

Setting a breakpoint in your pass

First thing you do is start gdb on the `opt` process:

```
$ gdb opt
GNU gdb 5.0
Copyright 2000 Free Software Foundation, Inc.
GDB is free software, covered by the GNU General Public License, and you are
welcome to change it and/or distribute copies of it under certain conditions.
Type "show copying" to see the conditions.
There is absolutely no warranty for GDB.  Type "show warranty" for details.
This GDB was configured as "sparc-sun-solaris2.6"...
(gdb)
```

Note that `opt` has a lot of debugging information in it, so it takes time to load. Be patient. Since we cannot set a breakpoint in our pass yet (the shared object isn't loaded until runtime), we must execute the process, and have it stop before it invokes our pass, but after it has loaded the shared object. The most foolproof way of doing this is to set a breakpoint in `PassManager::run` and then run the process with the arguments you want:

```
$ (gdb) break llvm::PassManager::run
Breakpoint 1 at 0x2413bc: file Pass.cpp, line 70.
(gdb) run test.bc -load $(LLVMTOP)/llvm/Debug+Asserts/lib/[libname].so -[passoption]
Starting program: opt test.bc -load $(LLVMTOP)/llvm/Debug+Asserts/lib/[libname].so -
↪[passoption]
Breakpoint 1, PassManager::run (this=0xffbef174, M=@0x70b298) at Pass.cpp:70
70      bool PassManager::run(Module &M) { return PM->run(M); }
(gdb)
```

Once the `opt` stops in the `PassManager::run` method you are now free to set breakpoints in your pass so that you can trace through execution or do other standard debugging stuff.

Miscellaneous Problems

Once you have the basics down, there are a couple of problems that GDB has, some with solutions, some without.

- Inline functions have bogus stack information. In general, GDB does a pretty good job getting stack traces and stepping through inline functions. When a pass is dynamically loaded however, it somehow completely loses this capability. The only solution I know of is to de-inline a function (move it from the body of a class to a `.cpp` file).
- Restarting the program breaks breakpoints. After following the information above, you have succeeded in getting some breakpoints planted in your pass. Next thing you know, you restart the program (i.e., you type "run" again), and you start getting errors about breakpoints being unsettable. The only way I have found to "fix" this problem is to delete the breakpoints that are already set in your pass, run the program, and re-set the breakpoints once execution stops in `PassManager::run`.

Hopefully these tips will help with common case debugging situations. If you'd like to contribute some tips of your own, just contact [Chris](#).

Future extensions planned

Although the LLVM Pass Infrastructure is very capable as it stands, and does some nifty stuff, there are things we'd like to add in the future. Here is where we are going:

Multithreaded LLVM

Multiple CPU machines are becoming more common and compilation can never be fast enough: obviously we should allow for a multithreaded compiler. Because of the semantics defined for passes above (specifically they cannot maintain state across invocations of their `run*` methods), a nice clean way to implement a multithreaded compiler would be for the `PassManager` class to create multiple instances of each pass object, and allow the separate instances to be hacking on different parts of the program at the same time.

This implementation would prevent each of the passes from having to implement multithreaded constructs, requiring only the LLVM core to have locking in a few places (for global resources). Although this is a simple extension, we simply haven't had time (or multiprocessor machines, thus a reason) to implement this. Despite that, we have kept the LLVM passes SMP ready, and you should too.

4.22 How To Use Attributes

- [Introduction](#)
- [Attribute](#)
- [AttributeList](#)
- [AttrBuilder](#)

4.22.1 Introduction

Attributes in LLVM have changed in some fundamental ways. It was necessary to do this to support expanding the attributes to encompass more than a handful of attributes — e.g. command line options. The old way of handling attributes consisted of representing them as a bit mask of values. This bit mask was stored in a “list” structure that was reference counted. The advantage of this was that attributes could be manipulated with ‘or’s and ‘and’s. The disadvantage of this was that there was limited room for expansion, and virtually no support for attribute-value pairs other than alignment.

In the new scheme, an `Attribute` object represents a single attribute that's unique. You use the `Attribute::get` methods to create a new `Attribute` object. An attribute can be a single “enum” value (the enum being the `Attribute::AttrKind` enum), a string representing a target-dependent attribute, or an attribute-value pair. Some examples:

- Target-independent: `noinline`, `zext`
- Target-dependent: `"no-sse"`, `"thumb2"`
- Attribute-value pair: `"cpu" = "cortex-a8", align = 4`

Note: for an attribute value pair, we expect a target-dependent attribute to have a string for the value.

4.22.2 Attribute

An `Attribute` object is designed to be passed around by value.

Because attributes are no longer represented as a bit mask, you will need to convert any code which does treat them as a bit mask to use the new query methods on the `Attribute` class.

4.22.3 AttributeList

The `AttributeList` stores a collection of `Attribute` objects for each kind of object that may have an attribute associated with it: the function as a whole, the return type, or the function's parameters. A function's attributes are at index `AttributeList::FunctionIndex`; the return type's attributes are at index `AttributeList::ReturnIndex`; and the function's parameters' attributes are at indices 1, ..., n (where 'n' is the number of parameters). Most methods on the `AttributeList` class take an index parameter.

An `AttributeList` is also a unique and immutable object. You create an `AttributeList` through the `AttributeList::get` methods. You can add and remove attributes, which result in the creation of a new `AttributeList`.

An `AttributeList` object is designed to be passed around by value.

Note: It is advised that you do *not* use the `AttributeList` “introspection” methods (e.g. `Raw`, `getRawPointer`, etc.). These methods break encapsulation, and may be removed in a future release (i.e. LLVM 4.0).

4.22.4 AttrBuilder

Lastly, we have a “builder” class to help create the `AttributeList` object without having to create several different intermediate unique `AttributeList` objects. The `AttrBuilder` class allows you to add and remove attributes at will. The attributes won't be unique until you call the appropriate `AttributeList::get` method.

An `AttrBuilder` object is *not* designed to be passed around by value. It should be passed by reference.

Note: It is advised that you do *not* use the `AttrBuilder::addRawValue()` method or the `AttrBuilder(uint64_t Val)` constructor. These are for backwards compatibility and may be removed in a future release (i.e. LLVM 4.0).

And that's basically it! A lot of functionality is hidden behind these classes, but the interfaces are pretty straight forward.

4.23 User Guide for NVPTX Back-end

- *Introduction*
- *Conventions*
 - *Marking Functions as Kernels*
 - *Address Spaces*
 - *Triples*
- *NVPTX Intrinsics*
 - *Address Space Conversion*

- * `'llvm.nvvm.ptr.*.to.gen'` Intrinsic
 - * `'llvm.nvvm.ptr.gen.to.*'` Intrinsic
 - *Reading PTX Special Registers*
 - * `'llvm.nvvm.read.ptx.sreg.*'`
 - *Barriers*
 - * `'llvm.nvvm.barrier0'`
 - *Other Intrinsic*
- *Linking with Libdevice*
 - *Reflection Parameters*
- *Executing PTX*
- *Common Issues*
 - *ptxas complains of undefined function: __nvvm_reflect*
- *Tutorial: A Simple Compute Kernel*
 - *The Kernel*
 - *Dissecting the Kernel*
 - * *Data Layout*
 - * *Target Intrinsic*
 - * *Address Spaces*
 - * *Kernel Metadata*
 - *Running the Kernel*
- *Tutorial: Linking with Libdevice*

4.23.1 Introduction

To support GPU programming, the NVPTX back-end supports a subset of LLVM IR along with a defined set of conventions used to represent GPU programming concepts. This document provides an overview of the general usage of the back-end, including a description of the conventions used and the set of accepted LLVM IR.

Note: This document assumes a basic familiarity with CUDA and the PTX assembly language. Information about the CUDA Driver API and the PTX assembly language can be found in the [CUDA documentation](#).

4.23.2 Conventions

Marking Functions as Kernels

In PTX, there are two types of functions: *device functions*, which are only callable by device code, and *kernel functions*, which are callable by host code. By default, the back-end will emit device functions. Metadata is used to declare a function as a kernel function. This metadata is attached to the `nvvm.annotations` named metadata object, and has the following format:

```
!0 = !{<function-ref>, metadata !"kernel", i32 1}
```

The first parameter is a reference to the kernel function. The following example shows a kernel function calling a device function in LLVM IR. The function `@my_kernel` is callable from host code, but `@my_fmadd` is not.

```
define float @my_fmadd(float %x, float %y, float %z) {
  %mul = fmul float %x, %y
  %add = fadd float %mul, %z
  ret float %add
}

define void @my_kernel(float* %ptr) {
  %val = load float, float* %ptr
  %ret = call float @my_fmadd(float %val, float %val, float %val)
  store float %ret, float* %ptr
  ret void
}

!nvvm.annotations = !{!1}
!1 = !{void (float*)* @my_kernel, !"kernel", i32 1}
```

When compiled, the PTX kernel functions are callable by host-side code.

Address Spaces

The NVPTX back-end uses the following address space mapping:

Address Space	Memory Space
0	Generic
1	Global
2	Internal Use
3	Shared
4	Constant
5	Local

Every global variable and pointer type is assigned to one of these address spaces, with 0 being the default address space. Intrinsic are provided which can be used to convert pointers between the generic and non-generic address spaces.

As an example, the following IR will define an array `@g` that resides in global device memory.

```
@g = internal addrspace(1) global [4 x i32] [ i32 0, i32 1, i32 2, i32 3 ]
```

LLVM IR functions can read and write to this array, and host-side code can copy data to it by name with the CUDA Driver API.

Note that since address space 0 is the generic space, it is illegal to have global variables in address space 0. Address space 0 is the default address space in LLVM, so the `addrspace(N)` annotation is *required* for global variables.

Triples

The NVPTX target uses the module triple to select between 32/64-bit code generation and the driver-compiler interface to use. The triple architecture can be one of `nvptx` (32-bit PTX) or `nvptx64` (64-bit PTX). The operating system should be one of `cuda` or `nvcl`, which determines the interface used by the generated code to communicate with the

driver. Most users will want to use `cuda` as the operating system, which makes the generated PTX compatible with the CUDA Driver API.

Example: 32-bit PTX for CUDA Driver API: `nvptx-nvidia-cuda`

Example: 64-bit PTX for CUDA Driver API: `nvptx64-nvidia-cuda`

4.23.3 NVPTX Intrinsics

Address Space Conversion

`'llvm.nvvm.ptr.*.to.gen'` Intrinsics

Syntax:

These are overloaded intrinsics. You can use these on any pointer types.

```
declare i8* @llvm.nvvm.ptr.global.to.gen.p0i8.pli8(i8 addrspace(1) *)
declare i8* @llvm.nvvm.ptr.shared.to.gen.p0i8.p3i8(i8 addrspace(3) *)
declare i8* @llvm.nvvm.ptr.constant.to.gen.p0i8.p4i8(i8 addrspace(4) *)
declare i8* @llvm.nvvm.ptr.local.to.gen.p0i8.p5i8(i8 addrspace(5) *)
```

Overview:

The `'llvm.nvvm.ptr.*.to.gen'` intrinsics convert a pointer in a non-generic address space to a generic address space pointer.

Semantics:

These intrinsics modify the pointer value to be a valid generic address space pointer.

`'llvm.nvvm.ptr.gen.to.*'` Intrinsics

Syntax:

These are overloaded intrinsics. You can use these on any pointer types.

```
declare i8 addrspace(1) * @llvm.nvvm.ptr.gen.to.global.pli8.p0i8(i8*)
declare i8 addrspace(3) * @llvm.nvvm.ptr.gen.to.shared.p3i8.p0i8(i8*)
declare i8 addrspace(4) * @llvm.nvvm.ptr.gen.to.constant.p4i8.p0i8(i8*)
declare i8 addrspace(5) * @llvm.nvvm.ptr.gen.to.local.p5i8.p0i8(i8*)
```

Overview:

The `'llvm.nvvm.ptr.gen.to.*'` intrinsics convert a pointer in the generic address space to a pointer in the target address space. Note that these intrinsics are only useful if the address space of the target address space of the pointer is known. It is not legal to use address space conversion intrinsics to convert a pointer from one non-generic address space to another non-generic address space.

Semantics:

These intrinsics modify the pointer value to be a valid pointer in the target non-generic address space.

Reading PTX Special Registers

`'llvm.nvvm.read.ptx.sreg.*'`

Syntax:

```
declare i32 @llvm.nvvm.read.ptx.sreg.tid.x()
declare i32 @llvm.nvvm.read.ptx.sreg.tid.y()
declare i32 @llvm.nvvm.read.ptx.sreg.tid.z()
declare i32 @llvm.nvvm.read.ptx.sreg.ntid.x()
declare i32 @llvm.nvvm.read.ptx.sreg.ntid.y()
declare i32 @llvm.nvvm.read.ptx.sreg.ntid.z()
declare i32 @llvm.nvvm.read.ptx.sreg.ctaid.x()
declare i32 @llvm.nvvm.read.ptx.sreg.ctaid.y()
declare i32 @llvm.nvvm.read.ptx.sreg.ctaid.z()
declare i32 @llvm.nvvm.read.ptx.sreg.nctaid.x()
declare i32 @llvm.nvvm.read.ptx.sreg.nctaid.y()
declare i32 @llvm.nvvm.read.ptx.sreg.nctaid.z()
declare i32 @llvm.nvvm.read.ptx.sreg.warpSize()
```

Overview:

The `@llvm.nvvm.read.ptx.sreg.*` intrinsics provide access to the PTX special registers, in particular the kernel launch bounds. These registers map in the following way to CUDA builtins:

CUDA Builtin	PTX Special Register Intrinsic
<code>threadId</code>	<code>@llvm.nvvm.read.ptx.sreg.tid.*</code>
<code>blockIdx</code>	<code>@llvm.nvvm.read.ptx.sreg.ctaid.*</code>
<code>blockDim</code>	<code>@llvm.nvvm.read.ptx.sreg.ntid.*</code>
<code>gridDim</code>	<code>@llvm.nvvm.read.ptx.sreg.nctaid.*</code>

Barriers

`'llvm.nvvm.barrier0'`

Syntax:

```
declare void @llvm.nvvm.barrier0()
```

Overview:

The `@llvm.nvvm.barrier0()` intrinsic emits a PTX `bar.sync 0` instruction, equivalent to the `__syncthreads()` call in CUDA.

Other Intrinsics

For the full set of NVPTX intrinsics, please see the `include/llvm/IR/IntrinsicsNVVM.td` file in the LLVM source tree.

4.23.4 Linking with Libdevice

The CUDA Toolkit comes with an LLVM bitcode library called `libdevice` that implements many common mathematical functions. This library can be used as a high-performance math library for any compilers using the LLVM NVPTX target. The library can be found under `nvvm/libdevice/` in the CUDA Toolkit and there is a separate version for each compute architecture.

For a list of all math functions implemented in `libdevice`, see [libdevice Users Guide](#).

To accommodate various math-related compiler flags that can affect code generation of `libdevice` code, the library code depends on a special LLVM IR pass (`NVVMReflect`) to handle conditional compilation within LLVM IR. This pass looks for calls to the `@__nvvm_reflect` function and replaces them with constants based on the defined reflection parameters. Such conditional code often follows a pattern:

```
float my_function(float a) {  
    if (__nvvm_reflect("FASTMATH"))  
        return my_function_fast(a);  
    else  
        return my_function_precise(a);  
}
```

The default value for all unspecified reflection parameters is zero.

The `NVVMReflect` pass should be executed early in the optimization pipeline, immediately after the link stage. The `internalize` pass is also recommended to remove unused math functions from the resulting PTX. For an input IR module `module.bc`, the following compilation flow is recommended:

1. Save list of external functions in `module.bc`
2. Link `module.bc` with `libdevice.compute_XX.YY.bc`
3. Internalize all functions not in list from (1)
4. Eliminate all unused internal functions
5. Run `NVVMReflect` pass
6. Run standard optimization pipeline

Note: `linkonce` and `linkonce_odr` linkage types are not suitable for the `libdevice` functions. It is possible to link two IR modules that have been linked against `libdevice` using different reflection variables.

Since the `NVVMReflect` pass replaces conditionals with constants, it will often leave behind dead code of the form:

```
entry:  
    ..  
    br i1 true, label %foo, label %bar  
foo:  
    ..  
bar:  
    ; Dead code  
    ..
```

Therefore, it is recommended that `NVVMReflect` is executed early in the optimization pipeline before dead-code elimination.

The NVPTX `TargetMachine` knows how to schedule `NVVMReflect` at the beginning of your pass manager; just use the following code when setting up your pass manager:

```
std::unique_ptr<TargetMachine> TM = ...;
PassManagerBuilder PMBuilder(...);
if (TM)
    TM->adjustPassManager(PMBuilder);
```

Reflection Parameters

The `libdevice` library currently uses the following reflection parameters to control code generation:

Flag	Description
<code>__CUDA_FTZ=[0,1]</code>	Use optimized code paths that flush subnormals to zero

The value of this flag is determined by the “`nvvm-reflect-ftz`” module flag. The following sets the `ftz` flag to 1.

```
!llvm.module.flag = !{!0}
!0 = !{i32 4, !"nvvm-reflect-ftz", i32 1}
```

(`i32 4` indicates that the value set here overrides the value in another module we link with. See the *LangRef* <[LangRef.html#module-flags-metadata](#)> for details.)

4.23.5 Executing PTX

The most common way to execute PTX assembly on a GPU device is to use the CUDA Driver API. This API is a low-level interface to the GPU driver and allows for JIT compilation of PTX code to native GPU machine code.

Initializing the Driver API:

```
CUdevice device;
CUcontext context;

// Initialize the driver API
cuInit(0);
// Get a handle to the first compute device
cuDeviceGet(&device, 0);
// Create a compute device context
cuCtxCreate(&context, 0, device);
```

JIT compiling a PTX string to a device binary:

```
CUmodule module;
CUfunction function;

// JIT compile a null-terminated PTX string
cuModuleLoadData(&module, (void*)PTXString);

// Get a handle to the "myfunction" kernel function
cuModuleGetFunction(&function, module, "myfunction");
```

For full examples of executing PTX assembly, please see the [CUDA Samples](#) distribution.

4.23.6 Common Issues

ptxas complains of undefined function: __nvvm_reflect

When linking with libdevice, the NVVMReflect pass must be used. See [Linking with Libdevice](#) for more information.

4.23.7 Tutorial: A Simple Compute Kernel

To start, let us take a look at a simple compute kernel written directly in LLVM IR. The kernel implements vector addition, where each thread computes one element of the output vector C from the input vectors A and B. To make this easier, we also assume that only a single CTA (thread block) will be launched, and that it will be one dimensional.

The Kernel

```
target datalayout = "e-p:64:64:64-i1:8:8-i8:8:8-i16:16:16-i32:32:32-i64:64:64-
↪f32:32:32-f64:64:64-v16:16:16-v32:32:32-v64:64:64-v128:128:128-n16:32:64"
target triple = "nvptx64-nvidia-cuda"

; Intrinsic to read X component of thread ID
declare i32 @llvm.nvvm.read.ptx.sreg.tid.x() readnone nounwind

define void @kernel(float addrspace(1)* %A,
                    float addrspace(1)* %B,
                    float addrspace(1)* %C) {
entry:
  ; What is my ID?
  %id = tail call i32 @llvm.nvvm.read.ptx.sreg.tid.x() readnone nounwind

  ; Compute pointers into A, B, and C
  %ptrA = getelementptr float, float addrspace(1)* %A, i32 %id
  %ptrB = getelementptr float, float addrspace(1)* %B, i32 %id
  %ptrC = getelementptr float, float addrspace(1)* %C, i32 %id

  ; Read A, B
  %valA = load float, float addrspace(1)* %ptrA, align 4
  %valB = load float, float addrspace(1)* %ptrB, align 4

  ; Compute C = A + B
  %valC = fadd float %valA, %valB

  ; Store back to C
  store float %valC, float addrspace(1)* %ptrC, align 4

  ret void
}

!nvvm.annotations = !{!0}
!0 = !{void (float addrspace(1)*,
              float addrspace(1)*,
              float addrspace(1)*)* @kernel, !"kernel", i32 1}
```

We can use the LLVM `llc` tool to directly run the NVPTX code generator:

```
# llc -mcpu=sm_20 kernel.ll -o kernel.ptx
```

Note: If you want to generate 32-bit code, change `p:64:64:64` to `p:32:32:32` in the module data layout string and use `nvptx-nvidia-cuda` as the target triple.

The output we get from `llc` (as of LLVM 3.4):

```
//
// Generated by LLVM NVPTX Back-End
//

.version 3.1
.target sm_20
.address_size 64

    // .globl kernel
                                // @kernel
.visible .entry kernel(
    .param .u64 kernel_param_0,
    .param .u64 kernel_param_1,
    .param .u64 kernel_param_2
)
{
    .reg .f32    %f<4>;
    .reg .s32    %r<2>;
    .reg .s64    %r<8>;

// BB#0:                                // %entry
    ld.param.u64    %r11, [kernel_param_0];
    mov.u32         %r1, %tid.x;
    mul.wide.s32     %r12, %r1, 4;
    add.s64         %r13, %r11, %r12;
    ld.param.u64     %r14, [kernel_param_1];
    add.s64         %r15, %r14, %r12;
    ld.param.u64     %r16, [kernel_param_2];
    add.s64         %r17, %r16, %r12;
    ld.global.f32     %f1, [%r13];
    ld.global.f32     %f2, [%r15];
    add.f32          %f3, %f1, %f2;
    st.global.f32     [%r17], %f3;
    ret;
}
```

Dissecting the Kernel

Now let us dissect the LLVM IR that makes up this kernel.

Data Layout

The data layout string determines the size in bits of common data types, their ABI alignment, and their storage size. For NVPTX, you should use one of the following:

32-bit PTX:

```
target datalayout = "e-p:32:32:32-i1:8:8-i8:8:8-i16:16:16-i32:32:32-i64:64:64-
↪f32:32:32-f64:64:64-v16:16:16-v32:32:32-v64:64:64-v128:128:128-n16:32:64"
```

64-bit PTX:

```
target datalayout = "e-p:64:64:64-i1:8:8-i8:8:8-i16:16:16-i32:32:32-i64:64:64-
↪f32:32:32-f64:64:64-v16:16:16-v32:32:32-v64:64:64-v128:128:128-n16:32:64"
```

Target Intrinsics

In this example, we use the `@llvm.nvvm.read.ptx.sreg.tid.x` intrinsic to read the X component of the current thread's ID, which corresponds to a read of register `%tid.x` in PTX. The NVPTX back-end supports a large set of intrinsics. A short list is shown below; please see `include/llvm/IR/IntrinsicsNVVM.td` for the full list.

Intrinsic	CUDA Equivalent
<code>i32 @llvm.nvvm.read.ptx.sreg.tid.{x,y,z}</code>	<code>threadIdx.{x,y,z}</code>
<code>i32 @llvm.nvvm.read.ptx.sreg.ctaid.{x,y,z}</code>	<code>blockIdx.{x,y,z}</code>
<code>i32 @llvm.nvvm.read.ptx.sreg.ntid.{x,y,z}</code>	<code>blockDim.{x,y,z}</code>
<code>i32 @llvm.nvvm.read.ptx.sreg.nctaid.{x,y,z}</code>	<code>gridDim.{x,y,z}</code>
<code>void @llvm.nvvm.barrier0()</code>	<code>__syncthreads()</code>

Address Spaces

You may have noticed that all of the pointer types in the LLVM IR example had an explicit address space specifier. What is address space 1? NVIDIA GPU devices (generally) have four types of memory:

- Global: Large, off-chip memory
- Shared: Small, on-chip memory shared among all threads in a CTA
- Local: Per-thread, private memory
- Constant: Read-only memory shared across all threads

These different types of memory are represented in LLVM IR as address spaces. There is also a fifth address space used by the NVPTX code generator that corresponds to the “generic” address space. This address space can represent addresses in any other address space (with a few exceptions). This allows users to write IR functions that can load/store memory using the same instructions. Intrinsics are provided to convert pointers between the generic and non-generic address spaces.

See [Address Spaces](#) and [NVPTX Intrinsics](#) for more information.

Kernel Metadata

In PTX, a function can be either a *kernel* function (callable from the host program), or a *device* function (callable only from GPU code). You can think of *kernel* functions as entry-points in the GPU program. To mark an LLVM IR function as a *kernel* function, we make use of special LLVM metadata. The NVPTX back-end will look for a named metadata node called `nvvm.annotations`. This named metadata must contain a list of metadata that describe the IR. For our purposes, we need to declare a metadata node that assigns the “kernel” attribute to the LLVM IR function that should be emitted as a PTX *kernel* function. These metadata nodes take the form:

```
!{<function ref>, metadata !"kernel", i32 1}
```

For the previous example, we have:

```
!nvvm.annotations = !{!0}
!0 = !{void (float addrspace(1) *,
            float addrspace(1) *,
            float addrspace(1) *)* @kernel, !"kernel", i32 1}
```

Here, we have a single metadata declaration in `nvvm.annotations`. This metadata annotates our `@kernel` function with the `kernel` attribute.

Running the Kernel

Generating PTX from LLVM IR is all well and good, but how do we execute it on a real GPU device? The CUDA Driver API provides a convenient mechanism for loading and JIT compiling PTX to a native GPU device, and launching a kernel. The API is similar to OpenCL. A simple example showing how to load and execute our vector addition code is shown below. Note that for brevity this code does not perform much error checking!

Note: You can also use the `ptxas` tool provided by the CUDA Toolkit to offline compile PTX to machine code (SASS) for a specific GPU architecture. Such binaries can be loaded by the CUDA Driver API in the same way as PTX. This can be useful for reducing startup time by precompiling the PTX kernels.

```
#include <iostream>
#include <fstream>
#include <cassert>
#include "cuda.h"

void checkCudaErrors(CUresult err) {
    assert(err == CUDA_SUCCESS);
}

/// main - Program entry point
int main(int argc, char **argv) {
    CUdevice    device;
    CUmodule    cudaModule;
    CUcontext    context;
    CUfunction    function;
    CUlinkState    linker;
    int          devCount;

    // CUDA initialization
    checkCudaErrors(cuInit(0));
    checkCudaErrors(cuDeviceGetCount(&devCount));
    checkCudaErrors(cuDeviceGet(&device, 0));

    char name[128];
    checkCudaErrors(cuDeviceGetName(name, 128, device));
    std::cout << "Using CUDA Device [0]: " << name << "\n";

    int devMajor, devMinor;
    checkCudaErrors(cuDeviceComputeCapability(&devMajor, &devMinor, device));
    std::cout << "Device Compute Capability: "
              << devMajor << "." << devMinor << "\n";
    if (devMajor < 2) {
        std::cerr << "ERROR: Device 0 is not SM 2.0 or greater\n";
        return 1;
    }
}
```

```
}

std::ifstream t("kernel.ptx");
if (!t.is_open()) {
    std::cerr << "kernel.ptx not found\n";
    return 1;
}
std::string str((std::istreambuf_iterator<char>(t),
               std::istreambuf_iterator<char>()));

// Create driver context
checkCudaErrors(cuCtxCreate(&context, 0, device));

// Create module for object
checkCudaErrors(cuModuleLoadDataEx(&cudaModule, str.c_str(), 0, 0, 0));

// Get kernel function
checkCudaErrors(cuModuleGetFunction(&function, cudaModule, "kernel"));

// Device data
CUdeviceptr devBufferA;
CUdeviceptr devBufferB;
CUdeviceptr devBufferC;

checkCudaErrors(cuMemAlloc(&devBufferA, sizeof(float)*16));
checkCudaErrors(cuMemAlloc(&devBufferB, sizeof(float)*16));
checkCudaErrors(cuMemAlloc(&devBufferC, sizeof(float)*16));

float* hostA = new float[16];
float* hostB = new float[16];
float* hostC = new float[16];

// Populate input
for (unsigned i = 0; i != 16; ++i) {
    hostA[i] = (float)i;
    hostB[i] = (float)(2*i);
    hostC[i] = 0.0f;
}

checkCudaErrors(cuMemcpyHtoD(devBufferA, &hostA[0], sizeof(float)*16));
checkCudaErrors(cuMemcpyHtoD(devBufferB, &hostB[0], sizeof(float)*16));

unsigned blockSizeX = 16;
unsigned blockSizeY = 1;
unsigned blockSizeZ = 1;
unsigned gridSizeX = 1;
unsigned gridSizeY = 1;
unsigned gridSizeZ = 1;

// Kernel parameters
void *KernelParams[] = { &devBufferA, &devBufferB, &devBufferC };

std::cout << "Launching kernel\n";

// Kernel launch
checkCudaErrors(cuLaunchKernel(function, gridSizeX, gridSizeY, gridSizeZ,
                               blockSizeX, blockSizeY, blockSizeZ,
```



```

        0, NULL, KernelParams, NULL));

// Retrieve device data
checkCudaErrors(cuMemcpyDtoH(&hostC[0], devBufferC, sizeof(float)*16));

std::cout << "Results:\n";
for (unsigned i = 0; i != 16; ++i) {
    std::cout << hostA[i] << " + " << hostB[i] << " = " << hostC[i] << "\n";
}

// Clean up after ourselves
delete [] hostA;
delete [] hostB;
delete [] hostC;

// Clean-up
checkCudaErrors(cuMemFree(devBufferA));
checkCudaErrors(cuMemFree(devBufferB));
checkCudaErrors(cuMemFree(devBufferC));
checkCudaErrors(cuModuleUnload(cudaModule));
checkCudaErrors(cuCtxDestroy(context));

return 0;
}

```

You will need to link with the CUDA driver and specify the path to `cuda.h`.

```
# clang++ sample.cpp -o sample -O2 -g -I/usr/local/cuda-5.5/include -lcuda
```

We don't need to specify a path to `libcuda.so` since this is installed in a system location by the driver, not the CUDA toolkit.

If everything goes as planned, you should see the following output when running the compiled program:

```

Using CUDA Device [0]: GeForce GTX 680
Device Compute Capability: 3.0
Launching kernel
Results:
0 + 0 = 0
1 + 2 = 3
2 + 4 = 6
3 + 6 = 9
4 + 8 = 12
5 + 10 = 15
6 + 12 = 18
7 + 14 = 21
8 + 16 = 24
9 + 18 = 27
10 + 20 = 30
11 + 22 = 33
12 + 24 = 36
13 + 26 = 39
14 + 28 = 42
15 + 30 = 45

```

Note: You will likely see a different device identifier based on your hardware

4.23.8 Tutorial: Linking with Libdevice

In this tutorial, we show a simple example of linking LLVM IR with the libdevice library. We will use the same kernel as the previous tutorial, except that we will compute $C = \text{pow}(A, B)$ instead of $C = A + B$. Libdevice provides an `__nv_powf` function that we will use.

```
target datalayout = "e-p:64:64:64-i1:8:8-i8:8:8-i16:16:16-i32:32:32-i64:64:64-
→f32:32:32-f64:64:64-v16:16:16-v32:32:32-v64:64:64-v128:128:128-n16:32:64"
target triple = "nvptx64-nvidia-cuda"

; Intrinsic to read X component of thread ID
declare i32 @llvm.nvvm.read.ptx.sreg.tid.x() readnone nounwind
; libdevice function
declare float @__nv_powf(float, float)

define void @kernel(float addrspace(1)* %A,
                    float addrspace(1)* %B,
                    float addrspace(1)* %C) {
entry:
    ; What is my ID?
    %id = tail call i32 @llvm.nvvm.read.ptx.sreg.tid.x() readnone nounwind

    ; Compute pointers into A, B, and C
    %ptrA = getelementptr float, float addrspace(1)* %A, i32 %id
    %ptrB = getelementptr float, float addrspace(1)* %B, i32 %id
    %ptrC = getelementptr float, float addrspace(1)* %C, i32 %id

    ; Read A, B
    %valA = load float, float addrspace(1)* %ptrA, align 4
    %valB = load float, float addrspace(1)* %ptrB, align 4

    ; Compute C = pow(A, B)
    %valC = call float @__nv_powf(float %valA, float %valB)

    ; Store back to C
    store float %valC, float addrspace(1)* %ptrC, align 4

    ret void
}

!nvvm.annotations = !{!0}
!0 = !{void (float addrspace(1)*,
              float addrspace(1)*,
              float addrspace(1)*)* @kernel, !"kernel", i32 1}
```

To compile this kernel, we perform the following steps:

1. Link with libdevice
2. Internalize all but the public kernel function
3. Run NVVMReflect and set `__CUDA_FTZ` to 0
4. Optimize the linked module

5. Codegen the module

These steps can be performed by the LLVM `llvm-link`, `opt`, and `llc` tools. In a complete compiler, these steps can also be performed entirely programmatically by setting up an appropriate pass configuration (see [Linking with Libdevice](#)).

```
# llvm-link t2.bc libdevice.compute_20.10.bc -o t2.linked.bc
# opt -internalize -internalize-public-api-list=kernel -nvvm-reflect-list=__CUDA_
  ↪FTZ=0 -nvvm-reflect -O3 t2.linked.bc -o t2.opt.bc
# llc -mcpu=sm_20 t2.opt.bc -o t2.ptx
```

Note: The `-nvvm-reflect-list=__CUDA_FTZ=0` is not strictly required, as any undefined variables will default to zero. It is shown here for evaluation purposes.

This gives us the following PTX (excerpt):

```
//
// Generated by LLVM NVPTX Back-End
//

.version 3.1
.target sm_20
.address_size 64

    // .globl kernel
                                // @kernel
.visible .entry kernel(
    .param .u64 kernel_param_0,
    .param .u64 kernel_param_1,
    .param .u64 kernel_param_2
)
{
    .reg .pred    %p<30>;
    .reg .f32     %f<111>;
    .reg .s32     %r<21>;
    .reg .s64     %r1<8>;

// BB#0:                                // %entry
    ld.param.u64    %r12, [kernel_param_0];
    mov.u32        %r3, %tid.x;
    ld.param.u64    %r13, [kernel_param_1];
    mul.wide.s32    %r14, %r3, 4;
    add.s64        %r15, %r12, %r14;
    ld.param.u64    %r16, [kernel_param_2];
    add.s64        %r17, %r13, %r14;
    add.s64        %r11, %r16, %r14;
    ld.global.f32   %f1, [%r15];
    ld.global.f32   %f2, [%r17];
    setp.eq.f32     %p1, %f1, 0f3F800000;
    setp.eq.f32     %p2, %f2, 0f00000000;
    or.pred        %p3, %p1, %p2;
    @%p3 bra       BB0_1;
    bra.uni        BB0_2;
BB0_1:
    mov.f32        %f110, 0f3F800000;
    st.global.f32   [%r11], %f110;
    ret;
```

```
BB0_2:                                     // __nv_isnanf.exit.i
    abs.f32    %f4, %f1;
    setp.gtu.f32 %p4, %f4, 0f7F800000;
    @%p4 bra   BB0_4;
// BB#3:                                     // __nv_isnanf.exit5.i
    abs.f32    %f5, %f2;
    setp.le.f32 %p5, %f5, 0f7F800000;
    @%p5 bra   BB0_5;
BB0_4:                                     // %.crtedgel.i
    add.f32    %f110, %f1, %f2;
    st.global.f32 [%r11], %f110;
    ret;
BB0_5:                                     // __nv_isinff.exit.i
    ...

BB0_26:                                    // __nv_truncf.exit.i.i.i.i.i
    mul.f32    %f90, %f107, 0f3FB8AA3B;
    cvt.rzi.f32.f32 %f91, %f90;
    mov.f32    %f92, 0fBF317200;
    fma.rn.f32 %f93, %f91, %f92, %f107;
    mov.f32    %f94, 0fB5BFBE8E;
    fma.rn.f32 %f95, %f91, %f94, %f93;
    mul.f32    %f89, %f95, 0f3FB8AA3B;
    // inline asm
    ex2.approx.ftz.f32 %f88,%f89;
    // inline asm
    add.f32    %f96, %f91, 0f00000000;
    ex2.approx.f32 %f97, %f96;
    mul.f32    %f98, %f88, %f97;
    setp.lt.f32 %p15, %f107, 0fC2D20000;
    selp.f32    %f99, 0f00000000, %f98, %p15;
    setp.gt.f32 %p16, %f107, 0f42D20000;
    selp.f32    %f110, 0f7F800000, %f99, %p16;
    setp.eq.f32 %p17, %f110, 0f7F800000;
    @%p17 bra   BB0_28;
// BB#27:
    fma.rn.f32 %f110, %f110, %f108, %f110;
BB0_28:                                    // __internal_accurate_powf.exit.i
    setp.lt.f32 %p18, %f1, 0f00000000;
    setp.eq.f32 %p19, %f3, 0f3F800000;
    and.pred    %p20, %p18, %p19;
    @!%p20 bra  BB0_30;
    bra.uni     BB0_29;
BB0_29:
    mov.b32     %r9, %f110;
    xor.b32     %r10, %r9, -2147483648;
    mov.b32     %f110, %r10;
BB0_30:                                    // __nv_powf.exit
    st.global.f32 [%r11], %f110;
    ret;
}
```

4.24 User Guide for AMDGPU Backend

- *Introduction*
- *LLVM*
 - *Target Triples*
 - *Processors*
 - *Address Spaces*
 - *Memory Scopes*
 - *AMDGPU Intrinsic*s
- *Code Object*
 - *Header*
 - *Sections*
 - *Note Records*
 - *Code Object Metadata*
 - *Symbols*
 - *Relocation Records*
 - *DWARF*
 - * *Address Space Mapping*
 - * *Register Mapping*
 - * *Source Text*
- *Code Conventions*
 - *AMDHSA*
 - * *Kernel Dispatch*
 - * *Memory Spaces*
 - * *HSA Image and Samplers*
 - * *HSA Signals*
 - * *HSA AQL Queue*
 - * *Kernel Descriptor*
 - *Kernel Descriptor for GFX6-GFX9*
 - * *Initial Kernel Execution State*
 - * *Kernel Prolog*
 - *M0*
 - *Flat Scratch*
 - * *Memory Model*
 - * *Trap Handler ABI*
 - *Non-AMDHSA*
 - * *Trap Handler ABI*

- *Source Languages*
 - *OpenCL*
 - *HCC*
 - *Assembler*
 - * *Operands*
 - * *Instruction Examples*
 - * *DS*
 - *FLAT*
 - *MUBUF*
 - *SMRD/SMEM*
 - *SOP1*
 - *SOP2*
 - *SOPC*
 - *SOPP*
 - *VALU*
 - * *HSA Code Object Directives*
 - *.hsa_code_object_version major, minor*
 - *.hsa_code_object_isa [major, minor, stepping, vendor, arch]*
 - *.amdgpu_hsa_kernel (name)*
 - *.amd_kernel_code_t*
- *Additional Documentation*

4.24.1 Introduction

The AMDGPU backend provides ISA code generation for AMD GPUs, starting with the R600 family up until the current GCN families. It lives in the `lib/Target/AMDGPU` directory.

4.24.2 LLVM

Target Triples

Use the `clang -target <Architecture>-<Vendor>-<OS>-<Environment>` option to specify the target triple:

Table 4.29: AMDGPU Target Triples

Architecture	Vendor	OS	Environment
r600	amd	<empty>	<empty>
amdgc	amd	<empty>	<empty>
amdgc	amd	amdhsa	<empty>
amdgc	amd	amdhsa	openc
amdgc	amd	amdhsa	amdgizcl
amdgc	amd	amdhsa	amdgiz
amdgc	amd	amdhsa	hcc

r600-amd-- Supports AMD GPUs HD2XXX-HD6XXX for graphics and compute shaders executed on the MESA runtime.

amdgc-amd-- Supports AMD GPUs GCN 6 onwards for graphics and compute shaders executed on the MESA runtime.

amdgc-amd-amdhsa- Supports AMD GCN GPUs GFX6 onwards for compute kernels executed on HSA [\[HSA\]](#) compatible runtimes such as AMD's ROCm [\[AMD-ROCm\]](#).

amdgc-amd-amdhsa-openc Supports AMD GCN GPUs GFX6 onwards for OpenCL compute kernels executed on HSA [\[HSA\]](#) compatible runtimes such as AMD's ROCm [\[AMD-ROCm\]](#). See [OpenCL](#).

amdgc-amd-amdhsa-amdgizcl Same as `amdgc-amd-amdhsa-openc` except a different address space mapping is used (see [Address Spaces](#)).

amdgc-amd-amdhsa-amdgiz Same as `amdgc-amd-amdhsa-` except a different address space mapping is used (see [Address Spaces](#)).

amdgc-amd-amdhsa-hcc Supports AMD GCN GPUs GFX6 onwards for AMD HC language compute kernels executed on HSA [\[HSA\]](#) compatible runtimes such as AMD's ROCm [\[AMD-ROCm\]](#). See [HCC](#).

Processors

Use the `clang -mcpu <Processor>` option to specify the AMD GPU processor. The names from both the *Processor* and *Alternative Processor* can be used.

Table 4.30: AMDGPU Processors

Processor	Alternative Processor	Target Triple Architecture	dGPU/ APU	Runtime Support	Example Products
R600 [AMD-R6xx]					
r600		r600	dGPU		
r630		r600	dGPU		
rs880		r600	dGPU		
rv670		r600	dGPU		
R700 [AMD-R7xx]					
rv710		r600	dGPU		
rv730		r600	dGPU		
rv770		r600	dGPU		
Evergreen [AMD-Evergreen]					
cedar		r600	dGPU		
redwood		r600	dGPU		
sumo		r600	dGPU		
juniper		r600	dGPU		

Continued on next page

Table 4.30 – continued from previous page

Processor	Alternative Processor	Target Triple Architecture	dGPU/ APU	Runtime Support	Example Products
cypress		r600	dGPU		
Northern Islands <i>[AMD-Cayman-Trinity]</i>					
barts		r600	dGPU		
turks		r600	dGPU		
caicos		r600	dGPU		
cayman		r600	dGPU		
GCN GFX6 (Southern Islands (SI)) <i>[AMD-Souther-Islands]</i>					
gfx600	<ul style="list-style-type: none"> • SI • tahiti 	amdgc	dGPU		
gfx601	<ul style="list-style-type: none"> • pitcairn • verde • oland • hainan 	amdgc	dGPU		
GCN GFX7 (Sea Islands (CI)) <i>[AMD-Sea-Islands]</i>					
gfx700	<ul style="list-style-type: none"> • bonaire 	amdgc	dGPU		<ul style="list-style-type: none"> • Radeon HD 7790 • Radeon HD 8770 • R7 260 • R7 260X

Continued on next page

Table 4.30 – continued from previous page

Processor	Alternative Processor	Target Triple Architecture	dGPU/ APU	Runtime Support	Example Products
	<ul style="list-style-type: none"> • kaveri 	amdgc	APU		<ul style="list-style-type: none"> • A6-7000 • A6 Pro-7050B • A8-7100 • A8 Pro-7150B • A10-7300 • A10 Pro-7350B • FX-7500 • A8-7200P • A10-7400P • FX-7600P
gfx701	<ul style="list-style-type: none"> • hawaii 	amdgc	dGPU	ROCm	<ul style="list-style-type: none"> • FirePro W8100 • FirePro W9100 • FirePro S9150 • FirePro S9170
gfx702			dGPU	ROCm	<ul style="list-style-type: none"> • Radeon R9 290 • Radeon R9 290x • Radeon R390 • Radeon R390x

Continued on next page

Table 4.30 – continued from previous page

Processor	Alternative Processor	Target Triple Architecture	dGPU/ APU	Runtime Support	Example Products
gfx703	<ul style="list-style-type: none"> • kabini • mullins 	amdgc	APU		<ul style="list-style-type: none"> • E1-2100 • E1-2200 • E1-2500 • E2-3000 • E2-3800 • A4-5000 • A4-5100 • A6-5200 • A4 Pro-3340B
GCN GFX8 (Volcanic Islands (VI)) <i>[AMD-Volcanic-Islands]</i>					
gfx800	<ul style="list-style-type: none"> • iceland 	amdgc	dGPU		<ul style="list-style-type: none"> • FirePro S7150 • FirePro S7100 • FirePro W7100 • Radeon R285 • Radeon R9 380 • Radeon R9 385 • Mobile FirePro M7170
gfx801	<ul style="list-style-type: none"> • carrizo 	amdgc	APU		<ul style="list-style-type: none"> • A6-8500P • Pro A6-8500B • A8-8600P • Pro A8-8600B • FX-8800P • Pro A12-8800B

Continued on next page

Table 4.30 – continued from previous page

Processor	Alternative Processor	Target Triple Architecture	dGPU/ APU	Runtime Support	Example Products
		amdgc	APU	ROCm	<ul style="list-style-type: none"> • A10-8700P • Pro A10-8700B • A10-8780P
		amdgc	APU		<ul style="list-style-type: none"> • A10-9600P • A10-9630P • A12-9700P • A12-9730P • FX-9800P • FX-9830P
		amdgc	APU		<ul style="list-style-type: none"> • E2-9010 • A6-9210 • A9-9410
gfx802	<ul style="list-style-type: none"> • tonga 	amdgc	dGPU	ROCm	Same as gfx800
gfx803	<ul style="list-style-type: none"> • fiji 	amdgc	dGPU	ROCm	<ul style="list-style-type: none"> • Radeon R9 Nano • Radeon R9 Fury • Radeon R9 FuryX • Radeon Pro Duo • FirePro S9300x2

Continued on next page

Table 4.30 – continued from previous page

Processor	Alternative Processor	Target Triple Architecture	dGPU/ APU	Runtime Support	Example Products
	<ul style="list-style-type: none">• polaris10	amdgc	dGPU	ROCm	<ul style="list-style-type: none">• Radeon RX 470• Radeon RX 480
	<ul style="list-style-type: none">• polaris11	amdgc	dGPU	ROCm	<ul style="list-style-type: none">• Radeon RX 460
gfx804		amdgc	dGPU		Same as gfx803
gfx810	<ul style="list-style-type: none">• stoney	amdgc	APU		
GCN GFX9					
gfx900		amdgc	dGPU		<ul style="list-style-type: none">• Radeon Vega Frontier Edition
gfx901		amdgc	dGPU	ROCm	Same as gfx900 except XNACK is enabled
gfx902		amdgc	APU		<i>TBA</i>
gfx903		amdgc	APU		Same as gfx902 except XNACK is enabled

Address Spaces

The AMDGPU backend uses the following address space mappings.

The memory space names used in the table, aside from the region memory space, is from the OpenCL standard.

LLVM Address Space number is used throughout LLVM (for example, in LLVM IR).

Table 4.31: Address Space Mapping

LLVM Space	Address	Memory Space			
		Current Default	amd-giz/amdgizcl	hcc	Future Default
0		Private (Scratch)	Generic (Flat)	Generic (Flat)	Generic (Flat)
1		Global	Global	Global	Global
2		Constant	Constant	Constant	Region (GDS)
3		Local (group/LDS)	Local (group/LDS)	Local (group/LDS)	Local (group/LDS)
4		Generic (Flat)	Region (GDS)	Region (GDS)	Constant
5		Region (GDS)	Private (Scratch)	Private (Scratch)	Private (Scratch)

Current Default This is the current default address space mapping used for all languages except `hcc`. This will shortly be deprecated.

amd-giz/amdgizcl This is the current address space mapping used when `amd-giz` or `amd-gizcl` is specified as the target triple environment value.

hcc This is the current address space mapping used when `hcc` is specified as the target triple environment value. This will shortly be deprecated.

Future Default This will shortly be the only address space mapping for all languages using AMDGPU backend.

Memory Scopes

This section provides LLVM memory synchronization scopes supported by the AMDGPU backend memory model when the target triple OS is `amdhsa` (see [Memory Model](#) and [Target Triples](#)).

The memory model supported is based on the HSA memory model [\[HSA\]](#) which is based in turn on HRF-indirect with scope inclusion [\[HRF\]](#). The happens-before relation is transitive over the synchronizes-with relation independent of scope, and synchronizes-with allows the memory scope instances to be inclusive (see table [AMDHSA LLVM Sync Scopes for AMDHSA](#)).

This is different to the OpenCL [\[OpenCL\]](#) memory model which does not have scope inclusion and requires the memory scopes to exactly match. However, this is conservatively correct for OpenCL.

Table 4.32: AMDHSA LLVM Sync Scopes for AMDHSA

LLVM Sync Scope	Description
<i>none</i>	<p>The default: <code>system</code>.</p> <p>Synchronizes with, and participates in modification and <code>seq_cst</code> total orderings with, other operations (except image operations) for all address spaces (except <code>private</code>, or <code>generic</code> that accesses <code>private</code>) provided the other operation's sync scope is:</p> <ul style="list-style-type: none"> • <code>system</code>. • <code>agent</code> and executed by a thread on the same agent. • <code>workgroup</code> and executed by a thread in the same workgroup. • <code>wavefront</code> and executed by a thread in the same wavefront.
<code>agent</code>	<p>Synchronizes with, and participates in modification and <code>seq_cst</code> total orderings with, other operations (except image operations) for all address spaces (except <code>private</code>, or <code>generic</code> that accesses <code>private</code>) provided the other operation's sync scope is:</p> <ul style="list-style-type: none"> • <code>system</code> or <code>agent</code> and executed by a thread on the same agent. • <code>workgroup</code> and executed by a thread in the same workgroup. • <code>wavefront</code> and executed by a thread in the same wavefront.
<code>workgroup</code>	<p>Synchronizes with, and participates in modification and <code>seq_cst</code> total orderings with, other operations (except image operations) for all address spaces (except <code>private</code>, or <code>generic</code> that accesses <code>private</code>) provided the other operation's sync scope is:</p> <ul style="list-style-type: none"> • <code>system</code>, <code>agent</code> or <code>workgroup</code> and executed by a thread in the same workgroup. • <code>wavefront</code> and executed by a thread in the same wavefront.
<code>wavefront</code>	<p>Synchronizes with, and participates in modification and <code>seq_cst</code> total orderings with, other operations (except image operations) for all address spaces (except <code>private</code>, or <code>generic</code> that accesses <code>private</code>) provided the other operation's sync scope is:</p> <ul style="list-style-type: none"> • <code>system</code>, <code>agent</code>, <code>workgroup</code> or <code>wavefront</code> and executed by a thread in the same wavefront.
<code>singlethread</code>	<p>Only synchronizes with, and participates in modification and <code>seq_cst</code> total orderings with, other operations (except image operations) running in the same thread for all address spaces (for example, in signal handlers).</p>

AMDGPU Intrinsics

The AMDGPU backend implements the following intrinsics.

This section is WIP.

4.24.3 Code Object

The AMDGPU backend generates a standard ELF [\[ELF\]](#) relocatable code object that can be linked by `lld` to produce a standard ELF shared code object which can be loaded and executed on an AMDGPU target.

Header

The AMDGPU backend uses the following ELF header:

Table 4.33: AMDGPU ELF Header

Field	Value
<code>e_ident[EI_CLASS]</code>	<code>ELFCLASS64</code>
<code>e_ident[EI_DATA]</code>	<code>ELFDATA2LSB</code>
<code>e_ident[EI_OSABI]</code>	<code>ELFOSABI_AMDGPU_HSA</code>
<code>e_ident[EI_ABIVERSION]</code>	<code>ELFABIVERSION_AMDGPU_HSA</code>
<code>e_type</code>	<code>ET_REL</code> or <code>ET_DYN</code>
<code>e_machine</code>	<code>EM_AMDGPU</code>
<code>e_entry</code>	<code>0</code>
<code>e_flags</code>	<code>0</code>

Table 4.34: AMDGPU ELF Header Enumeration Values

Name	Value
<code>EM_AMDGPU</code>	<code>224</code>
<code>ELFOSABI_AMDGPU_HSA</code>	<code>64</code>
<code>ELFABIVERSION_AMDGPU_HSA</code>	<code>1</code>

`e_ident[EI_CLASS]` The ELF class is always `ELFCLASS64`. The AMDGPU backend only supports 64 bit applications.

`e_ident[EI_DATA]` All AMDGPU targets use `ELFDATA2LSB` for little-endian byte ordering.

`e_ident[EI_OSABI]` The AMD GPU architecture specific OS ABI of `ELFOSABI_AMDGPU_HSA` is used to specify that the code object conforms to the AMD HSA runtime ABI [\[HSA\]](#).

`e_ident[EI_ABIVERSION]` The AMD GPU architecture specific OS ABI version of `ELFABIVERSION_AMDGPU_HSA` is used to specify the version of AMD HSA runtime ABI to which the code object conforms.

`e_type` Can be one of the following values:

`ET_REL` The type produced by the AMD GPU backend compiler as it is relocatable code object.

`ET_DYN` The type produced by the linker as it is a shared code object.

The AMD HSA runtime loader requires a `ET_DYN` code object.

`e_machine` The value `EM_AMDGPU` is used for the machine for all members of the AMD GPU architecture family. The specific member is specified in the `NT_AMD_AMDGPU_ISA` entry in the `.note` section (see [Note Records](#)).

e_entry The entry point is 0 as the entry points for individual kernels must be selected in order to invoke them through AQL packets.

e_flags The value is 0 as no flags are used.

Sections

An AMDGPU target ELF code object has the standard ELF sections which include:

Table 4.35: AMDGPU ELF Sections

Name	Type	Attributes
.bss	SHT_NOBITS	SHF_ALLOC + SHF_WRITE
.data	SHT_PROGBITS	SHF_ALLOC + SHF_WRITE
.debug_*	SHT_PROGBITS	<i>none</i>
.dynamic	SHT_DYNAMIC	SHF_ALLOC
.dynstr	SHT_PROGBITS	SHF_ALLOC
.dynsym	SHT_PROGBITS	SHF_ALLOC
.got	SHT_PROGBITS	SHF_ALLOC + SHF_WRITE
.hash	SHT_HASH	SHF_ALLOC
.note	SHT_NOTE	<i>none</i>
.relaname	SHT_RELA	<i>none</i>
.rela.dyn	SHT_RELA	<i>none</i>
.rodata	SHT_PROGBITS	SHF_ALLOC
.shstrtab	SHT_STRTAB	<i>none</i>
.strtab	SHT_STRTAB	<i>none</i>
.symtab	SHT_SYMTAB	<i>none</i>
.text	SHT_PROGBITS	SHF_ALLOC + SHF_EXECINSTR

These sections have their standard meanings (see [\[ELF\]](#)) and are only generated if needed.

.debug* The standard DWARF sections. See [DWARF](#) for information on the DWARF produced by the AMDGPU backend.

.dynamic, .dynstr, .dynstr, .hash The standard sections used by a dynamic loader.

.note See [Note Records](#) for the note records supported by the AMDGPU backend.

.relaname, .rela.dyn For relocatable code objects, *name* is the name of the section that the relocation records apply. For example, `.rela.text` is the section name for relocation records associated with the `.text` section.

For linked shared code objects, `.rela.dyn` contains all the relocation records from each of the relocatable code object's `.relaname` sections.

See [Relocation Records](#) for the relocation records supported by the AMDGPU backend.

.text The executable machine code for the kernels and functions they call. Generated as position independent code. See [Code Conventions](#) for information on conventions used in the isa generation.

Note Records

As required by ELFCLASS64, minimal zero byte padding must be generated after the `name` field to ensure the `desc` field is 4 byte aligned. In addition, minimal zero byte padding must be generated to ensure the `desc` field size is a multiple of 4 bytes. The `sh_addralign` field of the `.note` section must be at least 4 to indicate at least 8 byte alignment.

The AMDGPU backend code object uses the following ELF note records in the `.note` section. The *Description* column specifies the layout of the note record's `desc` field. All fields are consecutive bytes. Note records with variable size strings have a corresponding `*_size` field that specifies the number of bytes, including the terminating null character, in the string. The string(s) come immediately after the preceding fields.

Additional note records can be present.

Table 4.36: AMDGPU ELF Note Records

Name	Type	Description
"AMD"	NT_AMD_AMDGPU_METADATA	<metadata null terminated string>
"AMD"	NT_AMD_AMDGPU_ISA	<isa name null terminated string>

Table 4.37: AMDGPU ELF Note Record Enumeration Values

Name	Value
<i>reserved</i>	0-9
NT_AMD_AMDGPU_METADATA	10
NT_AMD_AMDGPU_ISA	11

NT_AMD_AMDGPU_ISA Specifies the instruction set architecture used by the machine code contained in the code object.

This note record is required for code objects containing machine code for processors matching the `amdgc`n architecture in table [Processors](#).

The null terminated string has the following syntax:

architecture-vendor-os-environment-processor

where:

architecture The architecture from table [AMDGPU Target Triples](#).

This is always `amdgc`n when the target triple OS is `amdhsa` (see [Target Triples](#)).

vendor The vendor from table [AMDGPU Target Triples](#).

For the AMDGPU backend this is always `amd`.

os The OS from table [AMDGPU Target Triples](#).

environment An environment from table [AMDGPU Target Triples](#), or blank if the environment has no affect on the execution of the code object.

For the AMDGPU backend this is currently always blank.

processor The processor from table [AMDGPU Processors](#).

For example:

`amdgc`n-amd-amdhsa--gfx901

NT_AMD_AMDGPU_METADATA Specifies extensible metadata associated with the code object. See [Code Object Metadata](#) for the syntax of the code object metadata string.

This note record is required and must contain the minimum information necessary to support the ROCM kernel queries. For example, the segment sizes needed in a dispatch packet. In addition, a high level language runtime may require other information to be included. For example, the AMD OpenCL runtime records kernel argument information.

Code Object Metadata

The code object metadata is specified by the `NT_AMD_AMDHSA_METADATA` note record (see *Note Records*).

The metadata is specified as a YAML formatted string (see *[YAML]* and *YAML I/O*).

The metadata is represented as a single YAML document comprised of the mapping defined in table *AMDHSA Code Object Metadata Mapping* and referenced tables.

For boolean values, the string values of `false` and `true` are used for false and true respectively.

Additional information can be added to the mappings. To avoid conflicts, any non-AMD key names should be prefixed by “*vendor-name*”.

Table 4.38: AMDHSA Code Object Metadata Mapping

String Key	Value Type	Required?	Description
"Version"	sequence of 2 integers	Required	<ul style="list-style-type: none"> The first integer is the major version. Currently 1. The second integer is the minor version. Currently 0.
"Printf"	sequence of strings		<p>Each string is encoded information about a printf function call. The encoded information is organized as fields separated by colon (':'):</p> <p>ID:N:S[0]:S[1]:. .. :S[N-1]:FormatString</p> <p>where:</p> <p>ID A 32 bit integer as a unique id for each printf function call</p> <p>N A 32 bit integer equal to the number of arguments of printf function call minus 1</p> <p>S[i] (where i = 0, 1, ..., N-1) 32 bit integers for the size in bytes of the i-th FormatString argument of the printf function call</p> <p>FormatString The format string passed to the printf function call.</p>
"Kernels"	sequence of mapping	Required	Sequence of the mappings for each kernel in the code object. See AMDHSA Code Object Kernel Metadata Mapping for the definition of the mapping.

Table 4.39: AMDHSA Code Object Kernel Metadata Mapping

String Key	Value Type	Required?	Description
“Name”	string	Required	Source name of the kernel.
“SymbolName”	string	Required	Name of the kernel descriptor ELF symbol.
“Language”	string		Source language of the kernel. Values include: <ul style="list-style-type: none"> • “OpenCL C” • “OpenCL C++” • “HCC” • “OpenMP”
“LanguageVersion”	sequence of 2 integers		<ul style="list-style-type: none"> • The first integer is the major version. • The second integer is the minor version.
“Attrs”	mapping		Mapping of kernel attributes. See <i>AMDHSA Code Object Kernel Attribute Metadata Mapping</i> for the mapping definition.
“Arguments”	sequence of mapping		Sequence of mappings of the kernel arguments. See <i>AMDHSA Code Object Kernel Argument Metadata Mapping</i> for the definition of the mapping.
“CodeProps”	mapping		Mapping of properties related to the kernel code. See <i>AMDHSA Code Object Kernel Code Properties Metadata Mapping</i> for the mapping definition.
“DebugProps”	mapping		Mapping of properties related to the kernel debugging. See <i>AMDHSA Code Object Kernel Debug Properties Metadata Mapping</i> for the mapping definition.

Table 4.40: AMDHSA Code Object Kernel Attribute Metadata Mapping

String Key	Value Type	Re-quired?	Description
“Reqd-Work-Group-Size”	sequence of 3 integers		The dispatch work-group size X, Y, Z must correspond to the specified values. Corresponds to the OpenCL <code>reqd_work_group_size</code> attribute.
“Work-Group-SizeHint”	sequence of 3 integers		The dispatch work-group size X, Y, Z is likely to be the specified values. Corresponds to the OpenCL <code>work_group_size_hint</code> attribute.
“VecType-Hint”	string		The name of a scalar or vector type. Corresponds to the OpenCL <code>vec_type_hint</code> attribute.

Table 4.41: AMDHSA Code Object Kernel Argument Metadata Mapping

String Key	Value Type	Required?	Description
“Name”	string		Kernel argument name.
“TypeName”	string		Kernel argument type name.
“Size”	integer	Required	Kernel argument size in bytes.
“Align”	integer	Required	Kernel argument alignment in bytes. Must be a power of two.
“ValueKind”	string	Required	<p>Kernel argument kind that specifies how to set up the corresponding argument. Values include:</p> <p>“ByValue” The argument is copied directly into the kernarg.</p> <p>“GlobalBuffer” A global address space pointer to the buffer data is passed in the kernarg.</p> <p>“DynamicSharedPointer” A group address space pointer to dynamically allocated LDS is passed in the kernarg.</p> <p>“Sampler” A global address space pointer to a S# is passed in the kernarg.</p> <p>“Image” A global address space pointer to a T# is passed in the kernarg.</p> <p>“Pipe” A global address space pointer to an OpenCL pipe is passed in the kernarg.</p> <p>“Queue” A global address space pointer to an OpenCL device enqueue queue is passed in the kernarg.</p> <p>“HiddenGlobalOffsetX” The OpenCL grid</p>

Table 4.42: AMDHSA Code Object Kernel Code Properties Metadata Mapping

String Key	Value Type	Required?	Description
“KernargSegment-Size”	integer	Required	The size in bytes of the kernarg segment that holds the values of the arguments to the kernel.
“GroupSegment-Fixed-Size”	integer	Required	The amount of group segment memory required by a work-group in bytes. This does not include any dynamically allocated group segment memory that may be added when the kernel is dispatched.
“Private-Segment-Fixed-Size”	integer	Required	The amount of fixed private address space memory required for a work-item in bytes. If IsDynamicCallstack is 1 then additional space must be added to this value for the call stack.
“KernargSegment-Align”	integer	Required	The maximum byte alignment of arguments in the kernarg segment. Must be a power of 2.
“Wavefront-Size”	integer	Required	Wavefront size. Must be a power of 2.
“NumSGPRs”	integer		Number of scalar registers used by a wavefront for GFX6-GFX9. This includes the special SGPRs for VCC, Flat Scratch (GFX7-GFX9) and XNACK (for GFX8-GFX9). It does not include the 16 SGPR added if a trap handler is enabled. It is not rounded up to the allocation granularity.
“NumVGPRs”	integer		Number of vector registers used by each work-item for GFX6-GFX9
“MaxFlat-Work-group-Size”	integer		Maximum flat work-group size supported by the kernel in work-items.
“IsDynamicCall-Stack”	boolean		Indicates if the generated machine code is using a dynamically sized call stack.
“IsX-NACK-Enabled”	boolean		Indicates if the generated machine code is capable of supporting XNACK.

Table 4.43: AMDHSA Code Object Kernel Debug Properties Metadata Mapping

String Key	Value Type	Required?	Description
“DebuggerABIVersion”	string		
“ReservedNumVGPRs”	integer		
“ReservedFirstVGPR”	integer		
“PrivateSegmentBufferSGPR”	integer		
“WavefrontPrivateSegmentOffsetSGPR”	integer		

Symbols

Symbols include the following:

Table 4.44: AMDGPU ELF Symbols

Name	Type	Section	Description
<i>link-name</i>	STT_OBJECT	<ul style="list-style-type: none">• .data• .rodata• .bss	Global variable
<i>link-name</i> @kd	STT_OBJECT	<ul style="list-style-type: none">• .rodata	Kernel descriptor
<i>link-name</i>	STT_FUNC	<ul style="list-style-type: none">• .text	Kernel entry point

Global variable Global variables both used and defined by the compilation unit.

If the symbol is defined in the compilation unit then it is allocated in the appropriate section according to if it has initialized data or is readonly.

If the symbol is external then its section is `STN_UNDEF` and the loader will resolve relocations using the definition provided by another code object or explicitly defined by the runtime.

All global symbols, whether defined in the compilation unit or external, are accessed by the machine code indirectly through a GOT table entry. This allows them to be preemptable. The GOT table is only supported when the target triple OS is `amdhsa` (see [Target Triples](#)).

Kernel descriptor Every HSA kernel has an associated kernel descriptor. It is the address of the kernel descriptor that is used in the AQL dispatch packet used to invoke the kernel, not the kernel entry point. The layout of the HSA kernel descriptor is defined in [Kernel Descriptor](#).

Kernel entry point Every HSA kernel also has a symbol for its machine code entry point.

Relocation Records

AMDGPU backend generates `Elf64_Rela` relocation records. Supported relocatable fields are:

word32 This specifies a 32-bit field occupying 4 bytes with arbitrary byte alignment. These values use the same byte order as other word values in the AMD GPU architecture.

word64 This specifies a 64-bit field occupying 8 bytes with arbitrary byte alignment. These values use the same byte order as other word values in the AMD GPU architecture.

Following notations are used for specifying relocation calculations:

A Represents the addend used to compute the value of the relocatable field.

G Represents the offset into the global offset table at which the relocation entry's symbol will reside during execution.

GOT Represents the address of the global offset table.

P Represents the place (section offset for `et_rel` or address for `et_dyn`) of the storage unit being relocated (computed using `r_offset`).

S Represents the value of the symbol whose index resides in the relocation entry.

The following relocation types are supported:

Table 4.45: AMDGPU ELF Relocation Records

Relocation Type	Value	Field	Calculation
R_AMDGPU_NONE	0	<i>none</i>	<i>none</i>
R_AMDGPU_ABS32_LO	1	word32	$(S + A) \& 0xFFFFFFFF$
R_AMDGPU_ABS32_HI	2	word32	$(S + A) \gg 32$
R_AMDGPU_ABS64	3	word64	$S + A$
R_AMDGPU_REL32	4	word32	$S + A - P$
R_AMDGPU_REL64	5	word64	$S + A - P$
R_AMDGPU_ABS32	6	word32	$S + A$
R_AMDGPU_GOTPCREL	7	word32	$G + GOT + A - P$
R_AMDGPU_GOTPCREL32_LO	8	word32	$(G + GOT + A - P) \& 0xFFFFFFFF$
R_AMDGPU_GOTPCREL32_HI	9	word32	$(G + GOT + A - P) \gg 32$
R_AMDGPU_REL32_LO	10	word32	$(S + A - P) \& 0xFFFFFFFF$
R_AMDGPU_REL32_HI	11	word32	$(S + A - P) \gg 32$

DWARF

Standard DWARF [\[DWARF\]](#) Version 2 sections can be generated. These contain information that maps the code object executable code and data to the source language constructs. It can be used by tools such as debuggers and profilers.

Address Space Mapping

The following address space mapping is used:

Table 4.46: AMDGPU DWARF Address Space Mapping

DWARF Address Space	Memory Space
1	Private (Scratch)
2	Local (group/LDS)
<i>omitted</i>	Global
<i>omitted</i>	Constant
<i>omitted</i>	Generic (Flat)
<i>not supported</i>	Region (GDS)

See [Address Spaces](#) for information on the memory space terminology used in the table.

An `address_class` attribute is generated on pointer type DIEs to specify the DWARF address space of the value of the pointer when it is in the *private* or *local* address space. Otherwise the attribute is omitted.

An `XDEREF` operation is generated in location list expressions for variables that are allocated in the *private* and *local* address space. Otherwise no `XDEREF` is omitted.

Register Mapping

This section is WIP.

Source Text

This section is WIP.

4.24.4 Code Conventions

AMDHSA

This section provides code conventions used when the target triple OS is `amdhsa` (see [Target Triples](#)).

Kernel Dispatch

The HSA architected queuing language (AQL) defines a user space memory interface that can be used to control the dispatch of kernels, in an agent independent way. An agent can have zero or more AQL queues created for it using the ROCm runtime, in which AQL packets (all of which are 64 bytes) can be placed. See the *HSA Platform System Architecture Specification* [\[HSA\]](#) for the AQL queue mechanics and packet layouts.

The packet processor of a kernel agent is responsible for detecting and dispatching HSA kernels from the AQL queues associated with it. For AMD GPUs the packet processor is implemented by the hardware command processor (CP), asynchronous dispatch controller (ADC) and shader processor input controller (SPI).

The ROCm runtime can be used to allocate an AQL queue object. It uses the kernel mode driver to initialize and register the AQL queue with CP.

To dispatch a kernel the following actions are performed. This can occur in the CPU host program, or from an HSA kernel executing on a GPU.

1. A pointer to an AQL queue for the kernel agent on which the kernel is to be executed is obtained.
2. A pointer to the kernel descriptor (see [Kernel Descriptor](#)) of the kernel to execute is obtained. It must be for a kernel that is contained in a code object that that was loaded by the ROCm runtime on the kernel agent with which the AQL queue is associated.
3. Space is allocated for the kernel arguments using the ROCm runtime allocator for a memory region with the `kernarg` property for the kernel agent that will execute the kernel. It must be at least 16 byte aligned.
4. Kernel argument values are assigned to the kernel argument memory allocation. The layout is defined in the *HSA Programmer's Language Reference* [\[HSA\]](#). For AMDGPU the kernel execution directly accesses the kernel argument memory in the same way constant memory is accessed. (Note that the HSA specification allows an implementation to copy the kernel argument contents to another location that is accessed by the kernel.)
5. An AQL kernel dispatch packet is created on the AQL queue. The ROCm runtime api uses 64 bit atomic operations to reserve space in the AQL queue for the packet. The packet must be set up, and the final write must use an atomic store release to set the packet kind to ensure the packet contents are visible to the kernel agent. AQL defines a doorbell signal mechanism to notify the kernel agent that the AQL queue has been updated. These rules, and the layout of the AQL queue and kernel dispatch packet is defined in the *HSA System Architecture Specification* [\[HSA\]](#).
6. A kernel dispatch packet includes information about the actual dispatch, such as grid and work-group size, together with information from the code object about the kernel, such as segment sizes. The ROCm runtime queries on the kernel symbol can be used to obtain the code object values which are recorded in the [Code Object Metadata](#).
7. CP executes micro-code and is responsible for detecting and setting up the GPU to execute the wavefronts of a kernel dispatch.
8. CP ensures that when the a wavefront starts executing the kernel machine code, the scalar general purpose registers (SGPR) and vector general purpose registers (VGPR) are set up as required by the machine code. The required setup is defined in the [Kernel Descriptor](#). The initial register state is defined in [Initial Kernel Execution State](#).
9. The prolog of the kernel machine code (see [Kernel Prolog](#)) sets up the machine state as necessary before continuing executing the machine code that corresponds to the kernel.

10. When the kernel dispatch has completed execution, CP signals the completion signal specified in the kernel dispatch packet if not 0.

Memory Spaces

The memory space properties are:

Table 4.47: AMDHSA Memory Spaces

Memory Name	Space	HSA Segment Name	Hardware Name	Address Size	NULL Value
Private		private	scratch	32	0x00000000
Local		group	LDS	32	0xFFFFFFFF
Global		global	global	64	0x0000000000000000
Constant		constant	<i>same as global</i>	64	0x0000000000000000
Generic		flat	flat	64	0x0000000000000000
Region		N/A	GDS	32	<i>not implemented for AMDHSA</i>

The global and constant memory spaces both use global virtual addresses, which are the same virtual address space used by the CPU. However, some virtual addresses may only be accessible to the CPU, some only accessible by the GPU, and some by both.

Using the constant memory space indicates that the data will not change during the execution of the kernel. This allows scalar read instructions to be used. The vector and scalar L1 caches are invalidated of volatile data before each kernel dispatch execution to allow constant memory to change values between kernel dispatches.

The local memory space uses the hardware Local Data Store (LDS) which is automatically allocated when the hardware creates work-groups of wavefronts, and freed when all the wavefronts of a work-group have terminated. The data store (DS) instructions can be used to access it.

The private memory space uses the hardware scratch memory support. If the kernel uses scratch, then the hardware allocates memory that is accessed using wavefront lane dword (4 byte) interleaving. The mapping used from private address to physical address is:

$$\text{wavefront-scratch-base} + (\text{private-address} * \text{wavefront-size} * 4) + (\text{wavefront-lane-id} * 4)$$

There are different ways that the wavefront scratch base address is determined by a wavefront (see [Initial Kernel Execution State](#)). This memory can be accessed in an interleaved manner using buffer instruction with the scratch buffer descriptor and per wave scratch offset, by the scratch instructions, or by flat instructions. If each lane of a wavefront accesses the same private address, the interleaving results in adjacent dwords being accessed and hence requires fewer cache lines to be fetched. Multi-dword access is not supported except by flat and scratch instructions in GFX9.

The generic address space uses the hardware flat address support available in GFX7-GFX9. This uses two fixed ranges of virtual addresses (the private and local apertures), that are outside the range of addressible global memory, to map from a flat address to a private or local address.

FLAT instructions can take a flat address and access global, private (scratch) and group (LDS) memory depending in if the address is within one of the aperture ranges. Flat access to scratch requires hardware aperture setup and setup in the kernel prologue (see [Flat Scratch](#)). Flat access to LDS requires hardware aperture setup and M0 (GFX7-GFX8) register setup (see [M0](#)).

To convert between a segment address and a flat address the base address of the apertures address can be used. For GFX7-GFX8 these are available in the [HSA AQL Queue](#) the address of which can be obtained with Queue Ptr SGPR

(see *Initial Kernel Execution State*). For GFX9 the appature base addresses are directly available as inline constant registers `SRC_SHARED_BASE/LIMIT` and `SRC_PRIVATE_BASE/LIMIT`. In 64 bit address mode the apperture sizes are 2^{32} bytes and the base is aligned to 2^{32} which makes it easier to convert from flat to segment or segment to flat.

HSA Image and Samplers

Image and sample handles created by the ROCm runtime are 64 bit addresses of a hardware 32 byte V# and 48 byte S# object respectively. In order to support the HSA `query_sampler` operations two extra dwords are used to store the HSA BRIG enumeration values for the queries that are not trivially deducible from the S# representation.

HSA Signals

Signal handles created by the ROCm runtime are 64 bit addresses of a structure allocated in memory accessible from both the CPU and GPU. The structure is defined by the ROCm runtime and subject to change between releases (see [\[AMD-ROCm-github\]](#)).

HSA AQL Queue

The AQL queue structure is defined by the ROCm runtime and subject to change between releases (see [\[AMD-ROCm-github\]](#)). For some processors it contains fields needed to implement certain language features such as the flat address aperture bases. It also contains fields used by CP such as managing the allocation of scratch memory.

Kernel Descriptor

A kernel descriptor consists of the information needed by CP to initiate the execution of a kernel, including the entry point address of the machine code that implements the kernel.

Kernel Descriptor for GFX6-GFX9

CP microcode requires the Kernel descritor to be allocated on 64 byte alignment.

Table 4.48: Kernel Descriptor for GFX6-GFX9

Bits	Size	Field Name	Description
31:0	4 bytes	group_segment_size	The amount of fixed local address space memory required for a work-group in bytes. This does not include any dynamically allocated local address space memory that may be added when the kernel is dispatched.
63:32	4 bytes	private_segment_size	The amount of fixed private address space memory required for a work-item in bytes. If <code>is_dynamic_callstack</code> is 1 then additional space must be added to this value for the call stack.
95:64	4 bytes	max_flat_work_group_size	Maximum flat work-group size supported by the kernel in work-items.
96	1 bit	is_dynamic_callstack	Indicates if the generated machine code is using a dynamically sized call stack.
97	1 bit	is_xnack_enabled	Indicates if the generated machine code is capable of supporting XNACK.
127:98	30 bits		Reserved. Must be 0.
191:128	64 bytes	kernel_code_entry_byte_offset	Byte offset (possibly negative) from base address of kernel descriptor to the byte offset point instruction which must be 256 byte aligned.
383:192	192 bytes		Reserved. Must be 0.
415:384	32 bytes	compute_pgm_rsrc1	Compute Shader (CS) program settings used by CP to set up COMPUTE_PGM_RSRC1 configuration register. See compute_pgm_rsrc1 for GFX6-GFX9 .
447:416	32 bytes	compute_pgm_rsrc2	Compute Shader (CS) program settings used by CP to set up COMPUTE_PGM_RSRC2 configuration register. See compute_pgm_rsrc2 for GFX6-GFX9 .
448	1 bit	enable_sgpr_private_segment_buffer	Enable the setup of the SGPR user data registers (see Initial Kernel Execution). The total number of SGPR user data registers requested must not exceed 16 and match value in <code>compute_pgm_rsrc2.user_sgpr.user_sgpr_count</code> . Any requests beyond 16 will be ignored.
449	1 bit	enable_sgpr_dispatch_ptr	<i>see above</i>
450	1 bit	enable_sgpr_queue_ptr	<i>see above</i>
451	1 bit	enable_sgpr_kernarg_segment_ptr	<i>see above</i>
452	1 bit	enable_sgpr_dispatch_id	<i>see above</i>
453	1 bit	enable_sgpr_flat_scratch_init	<i>see above</i>
454	1 bit	enable_sgpr_private_segment_size	<i>see above</i>
455	1 bit	enable_sgpr_grid_workgroup_count_X	Not implemented in CP and should always be 0.
456	1 bit	enable_sgpr_grid_workgroup_count_Y	Not implemented in CP and should always be 0.
457	1 bit	enable_sgpr_grid_workgroup_count_Z	Not implemented in CP and should always be 0.
463:458	6 bits		Reserved. Must be 0.
511:464	48 bytes		Reserved. Must be 0.
512	4 bytes		Total size 64 bytes

Table 4.49: compute_pgm_rsrc1 for GFX6-GFX9

Bits	Size	Field Name	Description
5:0	6 bits	granulated_workitem_vgpr_number	Number of vector registers used by each work-item, granularity is device specific: GFX6-9 $\text{roundup}((\text{max_vgpg} + 1) / 4) - 1$ Used by CP to set up COMPUTE_PGM_RSRC1.VGPRS.
9:6	4 bits	granulated_wavefront_sgpr_number	Number of scalar registers used by a wavefront, granularity is device specific: GFX6-8 $\text{roundup}((\text{max_sgpg} + 1) / 8) - 1$ GFX9 $\text{roundup}((\text{max_sgpg} + 1) / 16) - 1$ Includes the special SGPRs for VCC, Flat Scratch (for GFX7 onwards) and XNACK (for GFX8 onwards). It does not include the 16 SGPR added if a trap handler is enabled. Used by CP to set up COMPUTE_PGM_RSRC1.SGPRS.
11:10	2 bits	priority	Must be 0. Start executing wavefront at the specified priority. CP is responsible for filling in COMPUTE_PGM_RSRC1.PRIORITY.
13:12	2 bits	float_mode_round_32	Wavefront starts execution with specified rounding mode for single (32 bit) floating point precision floating point operations. Floating point rounding mode values are defined in <i>Floating Point Rounding Mode Enumeration Values</i> . Used by CP to set up COMPUTE_PGM_RSRC1.FLOAT_MODE.
15:14	2 bits	float_mode_round_16_64	Wavefront starts execution with specified

Table 4.50: compute_pgm_rsrc2 for GFX6-GFX9

Bits	Size	Field Name	Description
0	1 bit	enable_sgpr_private_segment_wave_offset	Enable the setup of the SGPR wave scratch offset system register (see <i>Initial Kernel Execution State</i>). Used by CP to set up COMPUTE_PGM_RSRC2.SCRATCH_EN.
5:1	5 bits	user_sgpr_count	The total number of SGPR user data registers requested. This number must match the number of user data registers enabled. Used by CP to set up COMPUTE_PGM_RSRC2.USER_SGPR.
6	1 bit	enable_trap_handler	Set to 1 if code contains a TRAP instruction which requires a trap handler to be enabled. CP sets COMPUTE_PGM_RSRC2.TRAP_PRESENT if the runtime has installed a trap handler regardless of the setting of this field.
7	1 bit	enable_sgpr_workgroup_id_x	Enable the setup of the system SGPR register for the work-group id in the X dimension (see <i>Initial Kernel Execution State</i>). Used by CP to set up COMPUTE_PGM_RSRC2.TGID_X_EN.
8	1 bit	enable_sgpr_workgroup_id_y	Enable the setup of the system SGPR register for the work-group id in the Y dimension (see <i>Initial Kernel Execution State</i>). Used by CP to set up COMPUTE_PGM_RSRC2.TGID_Y_EN.
9	1 bit	enable_sgpr_workgroup_id_z	Enable the setup of the system SGPR register for the work-group id in the Z dimension (see <i>Initial Kernel Execution State</i>). Used by CP to set up COMPUTE_PGM_RSRC2.TGID_Z_EN.
4.24.	User Guide for AMDGPU Backend		1249
10	1 bit	enable_sgpr_workgroup_id_w	Enable the setup of the

Table 4.51: Floating Point Rounding Mode Enumeration Values

Enumeration Name	Value	Description
AMD_FLOAT_ROUND_MODE_NEAR_EVEN	0	Round Ties To Even
AMD_FLOAT_ROUND_MODE_PLUS_INFINITY	1	Round Toward +infinity
AMD_FLOAT_ROUND_MODE_MINUS_INFINITY	2	Round Toward -infinity
AMD_FLOAT_ROUND_MODE_ZERO	3	Round Toward 0

Table 4.52: Floating Point Denorm Mode Enumeration Values

Enumeration Name	Value	Description
AMD_FLOAT_DENORM_MODE_FLUSH_SRC_DST	0	Flush Source and Destination Denorms
AMD_FLOAT_DENORM_MODE_FLUSH_DST	1	Flush Output Denorms
AMD_FLOAT_DENORM_MODE_FLUSH_SRC	2	Flush Source Denorms
AMD_FLOAT_DENORM_MODE_FLUSH_NONE	3	No Flush

Table 4.53: System VGPR Work-Item ID Enumeration Values

Enumeration Name	Value	Description
AMD_SYSTEM_VGPR_WORKITEM_ID_X	0	Set work-item X dimension ID.
AMD_SYSTEM_VGPR_WORKITEM_ID_X_Y	1	Set work-item X and Y dimensions ID.
AMD_SYSTEM_VGPR_WORKITEM_ID_X_Y_Z	2	Set work-item X, Y and Z dimensions ID.
AMD_SYSTEM_VGPR_WORKITEM_ID_UNDEFINED	3	Undefined.

Initial Kernel Execution State

This section defines the register state that will be set up by the packet processor prior to the start of execution of every wavefront. This is limited by the constraints of the hardware controllers of CP/ADC/SPI.

The order of the SGPR registers is defined, but the compiler can specify which ones are actually setup in the kernel descriptor using the `enable_sgpr_*` bit fields (see [Kernel Descriptor](#)). The register numbers used for enabled registers are dense starting at SGPR0: the first enabled register is SGPR0, the next enabled register is SGPR1 etc.; disabled registers do not have an SGPR number.

The initial SGPRs comprise up to 16 User SRGPs that are set by CP and apply to all waves of the grid. It is possible to specify more than 16 User SGPRs using the `enable_sgpr_*` bit fields, in which case only the first 16 are actually initialized. These are then immediately followed by the System SGPRs that are set up by ADC/SPI and can have different values for each wave of the grid dispatch.

SGPR register initial state is defined in [SGPR Register Set Up Order](#).

Table 4.54: SGPR Register Set Up Order

SGPR Order	Name (kernel descriptor enable field)	Number of SGPRs	Description
First	Private Segment Buffer (enable_sgpr_private_segment_buffer)	4	V# that can be used, together with Scratch Wave Offset as an offset, to access the private memory space using a segment address. CP uses the value provided by the runtime.
then	Dispatch Ptr (enable_sgpr_dispatch_ptr)	2	64 bit address of AQL dispatch packet for kernel dispatch actually executing.
then	Queue Ptr (enable_sgpr_queue_ptr)	2	64 bit address of amd_queue_t object for AQL queue on which the dispatch packet was queued.
then	Kernarg Segment Ptr (enable_sgpr_kernarg_segment_ptr)	2	64 bit address of Kernarg segment. This is directly copied from the kernarg_address in the kernel dispatch packet. Having CP load it once avoids loading it at the beginning of every wavefront.
then	Dispatch Id (enable_sgpr_dispatch_id)	2	64 bit Dispatch ID of the dispatch packet being executed.
then	Flat Scratch Init (enable_sgpr_flat_scratch_init)	2	This is 2 SGPRs: GFX6 Not supported. GFX7-GFX8 The first SGPR is a 32 bit byte offset from SH_HIDDEN_PRIVATE_BASE_VIMID to per SPI base of memory for scratch for the queue executing the kernel dispatch. CP obtains this from the runtime. This is the same offset used in computing the Scratch Segment Buffer base address. The value of Scratch Wave Offset must be added by the kernel machine code and moved

The order of the VGPR registers is defined, but the compiler can specify which ones are actually setup in the kernel descriptor using the `enable_vgpr*` bit fields (see [Kernel Descriptor](#)). The register numbers used for enabled registers are dense starting at VGPR0: the first enabled register is VGPR0, the next enabled register is VGPR1 etc.; disabled registers do not have a VGPR number.

VGPR register initial state is defined in [VGPR Register Set Up Order](#).

Table 4.55: VGPR Register Set Up Order

VGPR Order	Name (kernel descriptor enable field)	Number of VGPRs	Description
First	Work-Item Id X (Always initialized)	1	32 bit work item id in X dimension of work-group for wavefront lane.
then	Work-Item Id Y (<code>enable_vgpr_workitem_id > 0</code>)	1	32 bit work item id in Y dimension of work-group for wavefront lane.
then	Work-Item Id Z (<code>enable_vgpr_workitem_id > 1</code>)	1	32 bit work item id in Z dimension of work-group for wavefront lane.

The setting of registers is done by GPU CP/ADC/SPI hardware as follows:

1. SGPRs before the Work-Group Ids are set by CP using the 16 User Data registers.
2. Work-group Id registers X, Y, Z are set by ADC which supports any combination including none.
3. Scratch Wave Offset is set by SPI in a per wave basis which is why its value cannot be included with the flat scratch init value which is per queue.
4. The VGPRs are set by SPI which only supports specifying either (X), (X, Y) or (X, Y, Z).

Flat Scratch register pair are adjacent SGRRs so they can be moved as a 64 bit value to the hardware required SGPRn-3 and SGPRn-4 respectively.

The global segment can be accessed either using buffer instructions (GFX6 which has V# 64 bit address support), flat instructions (GFX7-9), or global instructions (GFX9).

If buffer operations are used then the compiler can generate a V# with the following properties:

- base address of 0
- no swizzle
- ATC: 1 if IOMMU present (such as APU)
- ptr64: 1
- MTYPE set to support memory coherence that matches the runtime (such as CC for APU and NC for dGPU).

Kernel Prolog

M0

GFX6-GFX8 The M0 register must be initialized with a value at least the total LDS size if the kernel may access LDS via DS or flat operations. Total LDS size is available in dispatch packet. For M0, it is also possible to use maximum possible value of LDS for given target (0x7FFF for GFX6 and 0xFFFF for GFX7-GFX8).

GFX9 The M0 register is not used for range checking LDS accesses and so does not need to be initialized in the prolog.

Flat Scratch

If the kernel may use flat operations to access scratch memory, the prolog code must set up FLAT_SCRATCH register pair (FLAT_SCRATCH_LO/FLAT_SCRATCH_HI which are in SGPRn-4/SGPRn-3). Initialization uses Flat Scratch Init and Scratch Wave Offset SGPR registers (see *Initial Kernel Execution State*):

GFX6 Flat scratch is not supported.

GFX7-8

1. The low word of Flat Scratch Init is 32 bit byte offset from SH_HIDDEN_PRIVATE_BASE_VIMID to the base of scratch backing memory being managed by SPI for the queue executing the kernel dispatch. This is the same value used in the Scratch Segment Buffer V# base address. The prolog must add the value of Scratch Wave Offset to get the wave's byte scratch backing memory offset from SH_HIDDEN_PRIVATE_BASE_VIMID. Since FLAT_SCRATCH_LO is in units of 256 bytes, the offset must be right shifted by 8 before moving into FLAT_SCRATCH_LO.
2. The second word of Flat Scratch Init is 32 bit byte size of a single work-items scratch memory usage. This is directly loaded from the kernel dispatch packet Private Segment Byte Size and rounded up to a multiple of DWORD. Having CP load it once avoids loading it at the beginning of every wavefront. The prolog must move it to FLAT_SCRATCH_LO for use as FLAT_SCRATCH_SIZE.

GFX9 The Flat Scratch Init is the 64 bit address of the base of scratch backing memory being managed by SPI for the queue executing the kernel dispatch. The prolog must add the value of Scratch Wave Offset and moved to the FLAT_SCRATCH pair for use as the flat scratch base in flat memory instructions.

Memory Model

This section describes the mapping of LLVM memory model onto AMDGPU machine code (see *Memory Model for Concurrent Operations*). *The implementation is WIP.*

The AMDGPU backend supports the memory synchronization scopes specified in *Memory Scopes*.

The code sequences used to implement the memory model are defined in table *AMDHSA Memory Model Code Sequences GFX6-GFX9*.

The sequences specify the order of instructions that a single thread must execute. The `s_waitcnt` and `buffer_wbinvl_vol` are defined with respect to other memory instructions executed by the same thread. This allows them to be moved earlier or later which can allow them to be combined with other instances of the same instruction, or hoisted/sunk out of loops to improve performance. Only the instructions related to the memory model are given; additional `s_waitcnt` instructions are required to ensure registers are defined before being used. These may be able to be combined with the memory model `s_waitcnt` instructions as described above.

The AMDGPU memory model supports both the HSA *[HSA]* memory model, and the OpenCL *[OpenCL]* memory model. The HSA memory model uses a single happens-before relation for all address spaces (see *Address Spaces*). The OpenCL memory model which has separate happens-before relations for the global and local address spaces, and only a fence specifying both global and local address space joins the relationships. Since the LLVM `memfence` instruction does not allow an address space to be specified the OpenCL fence has to conservatively assume both local and global address space was specified. However, optimizations can often be done to eliminate the additional `s_waitcnt` instructions when there are no intervening corresponding `ds/flat_load/store/atomic` memory instructions. The code sequences in the table indicate what can be omitted for the OpenCL memory. The target triple environment is used to determine if the source language is OpenCL (see *OpenCL*).

`ds/flat_load/store/atomic` instructions to local memory are termed LDS operations.

`buffer/global/flat_load/store/atomic` instructions to global memory are termed vector memory operations.

For GFX6-GFX9:

- Each agent has multiple compute units (CU).
- Each CU has multiple SIMDs that execute wavefronts.
- The wavefronts for a single work-group are executed in the same CU but may be executed by different SIMDs.
- Each CU has a single LDS memory shared by the wavefronts of the work-groups executing on it.
- All LDS operations of a CU are performed as wavefront wide operations in a global order and involve no caching. Completion is reported to a wavefront in execution order.
- The LDS memory has multiple request queues shared by the SIMDs of a CU. Therefore, the LDS operations performed by different waves of a work-group can be reordered relative to each other, which can result in re-ordering the visibility of vector memory operations with respect to LDS operations of other wavefronts in the same work-group. A `s_waitcnt lgkmcnt(0)` is required to ensure synchronization between LDS operations and vector memory operations between waves of a work-group, but not between operations performed by the same wavefront.
- The vector memory operations are performed as wavefront wide operations and completion is reported to a wavefront in execution order. The exception is that for GFX7-9 `flat_load/store/atomic` instructions can report out of vector memory order if they access LDS memory, and out of LDS operation order if they access global memory.
- The vector memory operations access a vector L1 cache shared by all wavefronts on a CU. Therefore, no special action is required for coherence between wavefronts in the same work-group. A `buffer_wbinvll_vol` is required for coherence between waves executing in different work-groups as they may be executing on different CUs.
- The scalar memory operations access a scalar L1 cache shared by all wavefronts on a group of CUs. The scalar and vector L1 caches are not coherent. However, scalar operations are used in a restricted way so do not impact the memory model. See *Memory Spaces*.
- The vector and scalar memory operations use an L2 cache shared by all CUs on the same agent.
- The L2 cache has independent channels to service disjoint ranges of virtual addresses.
- Each CU has a separate request queue per channel. Therefore, the vector and scalar memory operations performed by waves executing in different work-groups (which may be executing on different CUs) of an agent can be reordered relative to each other. A `s_waitcnt vmcnt(0)` is required to ensure synchronization between vector memory operations of different CUs. It ensures a previous vector memory operation has completed before executing a subsequent vector memory or LDS operation and so can be used to meet the requirements of acquire and release.
- The L2 cache can be kept coherent with other agents on some targets, or ranges of virtual addresses can be set up to bypass it to ensure system coherence.

Private address space uses `buffer_load/store` using the scratch V# (GFX6-8), or `scratch_load/store` (GFX9). Since only a single thread is accessing the memory, atomic memory orderings are not meaningful and all accesses are treated as non-atomic.

Constant address space uses `buffer/global_load` instructions (or equivalent scalar memory instructions). Since the constant address space contents do not change during the execution of a kernel dispatch it is not legal to perform stores, and atomic memory orderings are not meaningful and all access are treated as non-atomic.

A memory synchronization scope wider than work-group is not meaningful for the group (LDS) address space and is treated as work-group.

The memory model does not support the region address space which is treated as non-atomic.

Acquire memory ordering is not meaningful on store atomic instructions and is treated as non-atomic.

Release memory ordering is not meaningful on load atomic instructions and is treated as a non-atomic.

Acquire-release memory ordering is not meaningful on load or store atomic instructions and is treated as acquire and release respectively.

AMDGPU backend only uses scalar memory operations to access memory that is proven to not change during the execution of the kernel dispatch. This includes constant address space and global address space for program scope const variables. Therefore the kernel machine code does not have to maintain the scalar L1 cache to ensure it is coherent with the vector L1 cache. The scalar and vector L1 caches are invalidated between kernel dispatches by CP since constant address space data may change between kernel dispatch executions. See [Memory Spaces](#).

The one exception is if scalar writes are used to spill SGPR registers. In this case the AMDGPU backend ensures the memory location used to spill is never accessed by vector memory operations at the same time. If scalar writes are used then a `s_dcache_wb` is inserted before the `s_endpgm` and before a function return since the locations may be used for vector memory instructions by a future wave that uses the same scratch area, or a function call that creates a frame at the same address, respectively. There is no need for a `s_dcache_inv` as all scalar writes are write-before-read in the same thread.

Scratch backing memory (which is used for the private address space) is accessed with MTYPE NC_NV (non-coherent non-volatile). Since the private address space is only accessed by a single thread, and is always write-before-read, there is never a need to invalidate these entries from the L1 cache. Hence all cache invalidates are done as `*_vol` to only invalidate the volatile cache lines.

On dGPU the kernarg backing memory is accessed as UC (uncached) to avoid needing to invalidate the L2 cache. This also causes it to be treated as non-volatile and so is not invalidated by `*_vol`. On APU it is accessed as CC (cache coherent) and so the L2 cache will be coherent with the CPU and other agents.

Table 4.56: AMDHSA Memory Model Code Sequences GFX6-GFX9

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
Non-Atomic				
load	<i>none</i>	<i>none</i>	<ul style="list-style-type: none"> • global • generic 	non-volatile 1. buffer/global/flat_load volatile 1. buffer/global/flat_load glc=1
load	<i>none</i>	<i>none</i>	<ul style="list-style-type: none"> • local 	1. ds_load
store	<i>none</i>	<i>none</i>	<ul style="list-style-type: none"> • global • generic 	1. buffer/global/flat_store
store	<i>none</i>	<i>none</i>	<ul style="list-style-type: none"> • local 	1. ds_store
Unordered Atomic				
load atomic	unordered	<i>any</i>	<i>any</i>	<i>Same as non-atomic.</i>
store atomic	unordered	<i>any</i>	<i>any</i>	<i>Same as non-atomic.</i>

Continued on next page

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
atomicrmw	unordered	<i>any</i>	<i>any</i>	<i>Same as monotonic atomic.</i>
Monotonic Atomic				
load atomic	monotonic	<ul style="list-style-type: none"> • singlethread • wavefront • workgroup 	<ul style="list-style-type: none"> • global • generic 	1. buffer/global/flat_load
load atomic	monotonic	<ul style="list-style-type: none"> • singlethread • wavefront • workgroup 	<ul style="list-style-type: none"> • local 	1. ds_load
load atomic	monotonic	<ul style="list-style-type: none"> • agent • system 	<ul style="list-style-type: none"> • global • generic 	1. buffer/global/flat_load glc=1
store atomic	monotonic	<ul style="list-style-type: none"> • singlethread • wavefront • workgroup • agent • system 	<ul style="list-style-type: none"> • global • generic 	1. buffer/global/flat_store
store atomic	monotonic	<ul style="list-style-type: none"> • singlethread • wavefront • workgroup 	<ul style="list-style-type: none"> • local 	1. ds_store
atomicrmw	monotonic	<ul style="list-style-type: none"> • singlethread • wavefront • workgroup • agent • system 	<ul style="list-style-type: none"> • global • generic 	1. buffer/global/flat_atomic
atomicrmw	monotonic	<ul style="list-style-type: none"> • singlethread • wavefront • workgroup 	<ul style="list-style-type: none"> • local 	1. ds_atomic
Acquire Atomic				
load atomic	acquire	<ul style="list-style-type: none"> • singlethread • wavefront 	<ul style="list-style-type: none"> • global • local • generic 	1. buffer/global/ds/flat_load

Continued on next page

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
load atomic	acquire	<ul style="list-style-type: none"> • workgroup 	<ul style="list-style-type: none"> • global 	1. buffer/global_load
load atomic	acquire	<ul style="list-style-type: none"> • workgroup 	<ul style="list-style-type: none"> • local • generic 	1. ds/flat_load 2. s_waitcnt lgkmcnt(0) <ul style="list-style-type: none"> • If OpenCL, omit waitcnt. • Must happen before any following global/generic load/load atomic/store/store atomic/atomicrmw. • Ensures any following global data read is no older than the load atomic value being acquired.

Continued on next page

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
load atomic	acquire	<ul style="list-style-type: none"> • agent • system 	<ul style="list-style-type: none"> • global 	<ol style="list-style-type: none"> 1. buffer/global_load glc=1 2. s_waitcnt vncnt(0) <ul style="list-style-type: none"> • Must happen before following buffer_wbinvl1_vol. • Ensures the load has completed before invalidating the cache. 3. buffer_wbinvl1_vol <ul style="list-style-type: none"> • Must happen before any following global/generic load/load atomic/atomicrmw. • Ensures that following loads will not see stale global data.

Continued on next page

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
load atomic	acquire	<ul style="list-style-type: none"> • agent • system 	<ul style="list-style-type: none"> • generic 	<ol style="list-style-type: none"> 1. flat_load glc=1 2. s_waitcnt vmcnt(0) & lgkmcnt(0) <ul style="list-style-type: none"> • If OpenCL omit lgkmcnt(0). • Must happen before following buffer_wbinvl1_vol. • Ensures the flat_load has completed before invalidating the cache. 3. buffer_wbinvl1_vol <ul style="list-style-type: none"> • Must happen before any following global/generic load/load atomic/atomicmw. • Ensures that following loads will not see stale global data.
atomicmw	acquire	<ul style="list-style-type: none"> • singlethread • wavefront 	<ul style="list-style-type: none"> • global • local • generic 	<ol style="list-style-type: none"> 1. buffer/global/ds/flat_atomic
atomicmw	acquire	<ul style="list-style-type: none"> • workgroup 	<ul style="list-style-type: none"> • global 	<ol style="list-style-type: none"> 1. buffer/global_atomic

Continued on next page

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
atomicrmw	acquire	<ul style="list-style-type: none"> workgroup 	<ul style="list-style-type: none"> local generic 	<ol style="list-style-type: none"> ds/flat_atomic waitcnt lgkmcnt(0) <ul style="list-style-type: none"> If OpenCL, omit waitcnt. Must happen before any following global/generic load/load atomic/store/store atomic/atomicrmw. Ensures any following global data read is no older than the atomicrmw value being acquired.

Continued on next page

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
atomicrmw	acquire	<ul style="list-style-type: none">• agent• system	<ul style="list-style-type: none">• global	<ol style="list-style-type: none">1. buffer/global_atomic2. s_waitcnt vncnt(0)<ul style="list-style-type: none">• Must happen before following buffer_wbinvl1_vol.• Ensures the atomicrmw has completed before invalidating the cache.3. buffer_wbinvl1_vol<ul style="list-style-type: none">• Must happen before any following global/generic load/load atomic/atomicrmw.• Ensures that following loads will not see stale global data.

Continued on next page

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
atomicrmw	acquire	<ul style="list-style-type: none"> • agent • system 	<ul style="list-style-type: none"> • generic 	<ol style="list-style-type: none"> 1. flat_atomic 2. s_waitcnt vmcnt(0) & lgkmcnt(0) <ul style="list-style-type: none"> • If OpenCL, omit lgkmcnt(0). • Must happen before following buffer_wbinvl1_vol. • Ensures the atomicrmw has completed before invalidating the cache. 3. buffer_wbinvl1_vol <ul style="list-style-type: none"> • Must happen before any following global/generic load/load atomic/atomicrmw. • Ensures that following loads will not see stale global data.
fence	acquire	<ul style="list-style-type: none"> • singlethread • wavefront 	<i>none</i>	<i>none</i>

Continued on next page

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
fence	acquire	<ul style="list-style-type: none">workgroup	<i>none</i>	<div>1. s_waitcnt lgkmcnt(0)</div> <ul style="list-style-type: none">If OpenCL and address space is not generic, omit waitcnt. However, since LLVM currently has no address space on the fence need to conservatively always generate. If fence had an address space then set to address space of OpenCL fence flag, or to generic if both local and global flags are specified.Must happen after any preceding local/generic load atomic/atomicrmw with an equal or wider sync scope and memory ordering stronger than unordered (this is termed the fence-paired-atomic).Must happen before any following global/generic
4.24.	User Guide for AMDGPU Backend			1263

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
fence	acquire	<ul style="list-style-type: none"> • agent • system 	<i>none</i>	<ol style="list-style-type: none"> 1. s_waitcnt vmcnt(0) & lgkmcnt(0) <ul style="list-style-type: none"> • If OpenCL and address space is not generic, omit lgkmcnt(0). However, since LLVM currently has no address space on the fence need to conservatively always generate (see comment for previous fence). • Could be split into separate s_waitcnt vmcnt(0) and s_waitcnt lgkmcnt(0) to allow them to be independently moved according to the following rules. • s_waitcnt vmcnt(0) must happen after any preceding global/generic load atomic/atomicrmw with an equal or wider sync scope and memory ordering stronger

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
Release Atomic				
store atomic	release	<ul style="list-style-type: none"> • singlethread • wavefront 	<ul style="list-style-type: none"> • global • local • generic 	1. buffer/global/ds/flat_store
store atomic	release	<ul style="list-style-type: none"> • workgroup 	<ul style="list-style-type: none"> • global • generic 	1. s_waitcnt lgkmcnt(0) <ul style="list-style-type: none"> • If OpenCL, omit waitcnt. • Must happen after any preceding local/generic load/store/load atomic/store atomic/atomicrmw. • Must happen before the following store. • Ensures that all memory operations to local have completed before performing the store that is being released. 2. buffer/global/flat_store
store atomic	release	<ul style="list-style-type: none"> • workgroup 	<ul style="list-style-type: none"> • local 	1. ds_store

Continued on next page

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
store atomic	release	<ul style="list-style-type: none"> • agent • system 	<ul style="list-style-type: none"> • global • generic 	<ol style="list-style-type: none"> 1. s_waitcnt vmcnt(0) & lgkmcnt(0) <ul style="list-style-type: none"> • If OpenCL, omit lgkmcnt(0). • Could be split into separate s_waitcnt vmcnt(0) and s_waitcnt lgkmcnt(0) to allow them to be independently moved according to the following rules. • s_waitcnt vmcnt(0) must happen after any preceding global/generic load/store/load atomic/store atomic/atomicrmw. • s_waitcnt lgkmcnt(0) must happen after any preceding local/generic load/store/load atomic/store atomic/atomicrmw. • Must happen before the following store. • Ensures that all memory operations to global memory completed before performing the

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
atomicrmw	release	<ul style="list-style-type: none"> • singlethread • wavefront 	<ul style="list-style-type: none"> • global • local • generic 	1. buffer/global/ds/flat_atomic
atomicrmw	release	<ul style="list-style-type: none"> • workgroup 	<ul style="list-style-type: none"> • global • generic 	<p>1. s_waitcnt lgkmcnt(0)</p> <ul style="list-style-type: none"> • If OpenCL, omit waitcnt. • Must happen after any preceding local/generic load/store/load atomic/store atomic/atomicrmw. • Must happen before the following atomicrmw. • Ensures that all memory operations to local have completed before performing the atomicrmw that is being released. <p>2. buffer/global/flat_atomic</p>
atomicrmw	release	<ul style="list-style-type: none"> • workgroup 	<ul style="list-style-type: none"> • local 	1. ds_atomic

Continued on next page

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
atomicrmw	release	<ul style="list-style-type: none"> • agent • system 	<ul style="list-style-type: none"> • global • generic 	<ol style="list-style-type: none"> 1. s_waitcnt vmcnt(0) & lgkmcnt(0) <ul style="list-style-type: none"> • If OpenCL, omit lgkmcnt(0). • Could be split into separate s_waitcnt vmcnt(0) and s_waitcnt lgkmcnt(0) to allow them to be independently moved according to the following rules. • s_waitcnt vmcnt(0) must happen after any preceding global/generic load/store/load atomic/store atomic/atomicrmw. • s_waitcnt lgkmcnt(0) must happen after any preceding local/generic load/store/load atomic/store atomic/atomicrmw. • Must happen before the following atomicrmw. • Ensures that all memory operations to global and local have completed before per-

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
fence	release	<ul style="list-style-type: none"> • singlethread • wavefront 	<i>none</i>	<i>none</i>

Continued on next page

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
fence	release	<ul style="list-style-type: none"> workgroup 	<i>none</i>	<ol style="list-style-type: none"> s_waitcnt lgkmcnt(0) <ul style="list-style-type: none"> If OpenCL and address space is not generic, omit waitcnt. However, since LLVM currently has no address space on the fence need to conservatively always generate (see comment for previous fence). Must happen after any preceding local/generic load/load atomic/store/store atomic/atomicrmw. Must happen before any following store atomic/atomicrmw with an equal or wider sync scope and memory ordering stronger than unordered (this is termed the fence-paired-atomic). Ensures that all memory operations to be completed before performing the
1270			Chapter 4. Subsystem Documentation	

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
fence	release	<ul style="list-style-type: none"> • agent • system 	<i>none</i>	<ol style="list-style-type: none"> 1. s_waitcnt vmcnt(0) & lgkmcnt(0) <ul style="list-style-type: none"> • If OpenCL and address space is not generic, omit lgkmcnt(0). However, since LLVM currently has no address space on the fence need to conservatively always generate (see comment for previous fence). • Could be split into separate s_waitcnt vmcnt(0) and s_waitcnt lgkmcnt(0) to allow them to be independently moved according to the following rules. • s_waitcnt vmcnt(0) must happen after any preceding global/generic load/store/load atomic/store atomic/atomicrmw. • s_waitcnt lgkmcnt(0) must happen after any preceding lo-
4.24.	User Guide for AMDGPU Backend			1271

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
Acquire-Release Atomic				
atomicrmw	acq_rel	<ul style="list-style-type: none"> • singlethread • wavefront 	<ul style="list-style-type: none"> • global • local • generic 	1. buffer/global/ds/flat_atomic
atomicrmw	acq_rel	<ul style="list-style-type: none"> • workgroup 	<ul style="list-style-type: none"> • global 	<p>1. s_waitcnt lgkmcnt(0)</p> <ul style="list-style-type: none"> • If OpenCL, omit waitcnt. • Must happen after any preceding local/generic load/store/load atomic/store atomic/atomicrmw. • Must happen before the following atomicrmw. • Ensures that all memory operations to local have completed before performing the atomicrmw that is being released. <p>2. buffer/global_atomic</p>

Continued on next page

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
atomicrmw	acq_rel	<ul style="list-style-type: none"> workgroup 	<ul style="list-style-type: none"> local 	<ol style="list-style-type: none"> ds_atomic s_waitcnt lgkmcnt(0) <ul style="list-style-type: none"> If OpenCL, omit waitcnt. Must happen before any following global/generic load/load atomic/store/store atomic/atomicrmw. Ensures any following global data read is no older than the load atomic value being acquired.

Continued on next page

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
atomicrmw	acq_rel	<ul style="list-style-type: none"> workgroup 	<ul style="list-style-type: none"> generic 	<ol style="list-style-type: none"> s_waitcnt lgkmcnt(0) <ul style="list-style-type: none"> If OpenCL, omit waitcnt. Must happen after any preceding local/generic load/store/load atomic/store atomic/atomicrmw. flat_atomic s_waitcnt lgkmcnt(0) <ul style="list-style-type: none"> If OpenCL, omit waitcnt. Must happen before any following global/generic load/load atomic/store/store atomic/atomicrmw. Ensures any following global data read is no older than the load atomic value being acquired.

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
atomicrmw	acq_rel	<ul style="list-style-type: none"> • agent • system 	<ul style="list-style-type: none"> • global 	<ol style="list-style-type: none"> 1. s_waitcnt vmcnt(0) & lgkmcnt(0) <ul style="list-style-type: none"> • If OpenCL, omit lgkmcnt(0). • Could be split into separate s_waitcnt vmcnt(0) and s_waitcnt lgkmcnt(0) to allow them to be independently moved according to the following rules. • s_waitcnt vmcnt(0) must happen after any preceding global/generic load/store/load atomic/store atomic/atomicrmw. • s_waitcnt lgkmcnt(0) must happen after any preceding local/generic load/store/load atomic/store atomic/atomicrmw. • Must happen before the following atomicrmw. • Ensures that all memory operations to global have completed before performing the
4.24.	User Guide for AMDGPU Backend			global have completed before performing the

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
atomicrmw	acq_rel	<ul style="list-style-type: none"> • agent • system 	<ul style="list-style-type: none"> • generic 	<ol style="list-style-type: none"> 1. s_waitcnt vmcnt(0) & lgkmcnt(0) <ul style="list-style-type: none"> • If OpenCL, omit lgkmcnt(0). • Could be split into separate s_waitcnt vmcnt(0) and s_waitcnt lgkmcnt(0) to allow them to be independently moved according to the following rules. • s_waitcnt vmcnt(0) must happen after any preceding global/generic load/store/load atomic/store atomic/atomicrmw. • s_waitcnt lgkmcnt(0) must happen after any preceding local/generic load/store/load atomic/store atomic/atomicrmw. • Must happen before the following atomicrmw. • Ensures that all memory operations to global memory are completed before performing the

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
fence	acq_rel	<ul style="list-style-type: none">• singlethread• wavefront	<i>none</i>	<i>none</i>

Continued on next page

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
fence	acq_rel	<ul style="list-style-type: none"> workgroup 	<i>none</i>	<ol style="list-style-type: none"> s_waitcnt lgkmcnt(0) <ul style="list-style-type: none"> If OpenCL and address space is not generic, omit waitcnt. However, since LLVM currently has no address space on the fence need to conservatively always generate (see comment for previous fence). Must happen after any preceding local/generic load/load atomic/store/store atomic/atomicmw. Must happen before any following global/generic load/load atomic/store/store atomic/atomicmw. Ensures that all memory operations to local have completed before performing any following global memory operations. Ensures that the preceding local/generic load atomic/atomicmw with an
1278			Chapter 4. Subsystem Documentation	

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
fence	acq_rel	<ul style="list-style-type: none"> agent system 	<i>none</i>	<ol style="list-style-type: none"> s_waitcnt vmcnt(0) & lgkmcnt(0) <ul style="list-style-type: none"> If OpenCL and address space is not generic, omit lgkmcnt(0). However, since LLVM currently has no address space on the fence need to conservatively always generate (see comment for previous fence). Could be split into separate s_waitcnt vmcnt(0) and s_waitcnt lgkmcnt(0) to allow them to be independently moved according to the following rules. s_waitcnt vmcnt(0) must happen after any preceding global/generic load/store/load atomic/store atomic/atomicrmw. s_waitcnt lgkmcnt(0) must happen after any preceding lo-
4.24.	User Guide for AMDGPU Backend			1279

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
Sequential Consistent Atomic				
load atomic	seq_cst	<ul style="list-style-type: none">• singlethread• wavefront• workgroup	<ul style="list-style-type: none">• global• local• generic	<i>Same as corresponding load atomic acquire.</i>

Continued on next page

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
load atomic	seq_cst	<ul style="list-style-type: none">• agent• system	<ul style="list-style-type: none">• global• local• generic	<ol style="list-style-type: none">1. s_waitcnt vmcnt(0) <ul style="list-style-type: none">• Must happen after preceding global/generic load atomic/store atomic/atomicrmw with memory ordering of seq_cst and with equal or wider sync scope. (Note that seq_cst fences have their own s_waitcnt vmcnt(0) and so do not need to be considered.)• Ensures any preceding sequential consistent global memory instructions have completed before executing this sequentially consistent instruction. This prevents reordering a seq_cst store followed by a seq_cst load (Note that seq_cst is stronger than acquire/release as the re-ordering of load acquire followed
4.24.	User Guide for AMDGPU Backend			1281

Table 4.56 – continued from previous page

LLVM Instr	LLVM Memory Ordering	LLVM Memory Sync Scope	AMDGPU Address Space	AMDGPU Machine Code
store atomic	seq_cst	<ul style="list-style-type: none">• singlethread• wavefront• workgroup	<ul style="list-style-type: none">• global• local• generic	<i>Same as corresponding store atomic release.</i>
store atomic	seq_cst	<ul style="list-style-type: none">• agent• system	<ul style="list-style-type: none">• global• generic	<i>Same as corresponding store atomic release.</i>
atomicrmw	seq_cst	<ul style="list-style-type: none">• singlethread• wavefront• workgroup	<ul style="list-style-type: none">• global• local• generic	<i>Same as corresponding atomicrmw acq_rel.</i>
atomicrmw	seq_cst	<ul style="list-style-type: none">• agent• system	<ul style="list-style-type: none">• global• generic	<i>Same as corresponding atomicrmw acq_rel.</i>
fence	seq_cst	<ul style="list-style-type: none">• singlethread• wavefront• workgroup• agent• system	<i>none</i>	<i>Same as corresponding fence acq_rel.</i>

The memory order also adds the single thread optimization constraints defined in table [AMDHSA Memory Model Single Thread Optimization Constraints GFX6-GFX9](#).

Table 4.57: AMDHSA Memory Model Single Thread Optimization Constraints GFX6-GFX9

LLVM Memory	Optimization Constraints
Ordering	
unordered	<i>none</i>
monotonic	<i>none</i>
acquire	<ul style="list-style-type: none"> • If a load atomic/atomicrmw then no following load/load atomic/store/ store atomic/atomicrmw/fence instruction can be moved before the acquire. • If a fence then same as load atomic, plus no preceding associated fence-paired-atomic can be moved after the fence.
release	<ul style="list-style-type: none"> • If a store atomic/atomicrmw then no preceding load/load atomic/store/ store atomic/atomicrmw/fence instruction can be moved after the release. • If a fence then same as store atomic, plus no following associated fence-paired-atomic can be moved before the fence.
acq_rel	Same constraints as both acquire and release.
seq_cst	<ul style="list-style-type: none"> • If a load atomic then same constraints as acquire, plus no preceding sequentially consistent load atomic/store atomic/atomicrmw/fence instruction can be moved after the seq_cst. • If a store atomic then the same constraints as release, plus no following sequentially consistent load atomic/store atomic/atomicrmw/fence instruction can be moved before the seq_cst. • If an atomicrmw/fence then same constraints as acq_rel.

Trap Handler ABI

For code objects generated by AMDGPU backend for HSA [\[HSA\]](#) compatible runtimes (such as ROCm [\[AMD-ROCm\]](#)), the runtime installs a trap handler that supports the `s_trap` instruction with the following usage:

Table 4.58: AMDGPU Trap Handler for AMDHSA OS

Usage	Code Sequence	Trap Handler Inputs	Description
reserved	<code>s_trap 0x00</code>		Reserved by hardware.
<code>debugtrap(arg)</code>	<code>s_trap 0x01</code>	SGPR0-1: <code>queue_ptr</code> VGPR0: <code>arg</code>	Reserved for HSA <code>debugtrap</code> intrinsic (not implemented).
<code>llvm.trap</code>	<code>s_trap 0x02</code>	SGPR0-1: <code>queue_ptr</code>	Causes dispatch to be terminated and its associated queue put into the error state.
<code>llvm.debugtrap</code>	<code>s_trap 0x03</code>	SGPR0-1: <code>queue_ptr</code>	If debugger not installed handled same as <code>llvm.trap</code> .
debugger breakpoint	<code>s_trap 0x07</code>		Reserved for debugger breakpoints.
debugger	<code>s_trap 0x08</code>		Reserved for debugger.
debugger	<code>s_trap 0xfe</code>		Reserved for debugger.
debugger	<code>s_trap 0xff</code>		Reserved for debugger.

Non-AMDHSA

Trap Handler ABI

For code objects generated by AMDGPU backend for non-amdhsa OS, the runtime does not install a trap handler. The `llvm.trap` and `llvm.debugtrap` instructions are handled as follows:

Table 4.59: AMDGPU Trap Handler for Non-AMDHSA OS

Usage	Code Sequence	Description
<code>llvm.trap</code>	<code>s_endpgm</code>	Causes wavefront to be terminated.
<code>llvm.debugtrap</code>	<i>none</i>	Compiler warning given that there is no trap handler installed.

4.24.5 Source Languages

OpenCL

When generating code for the OpenCL language the target triple environment should be `opencl` or `amdglzcl` (see *Target Triples*).

When the language is OpenCL the following differences occur:

1. The OpenCL memory model is used (see *Memory Model*).
2. The AMDGPU backend adds additional arguments to the kernel.
3. Additional metadata is generated (*Code Object Metadata*).

HCC

When generating code for the OpenCL language the target triple environment should be `hcc` (see *Target Triples*).

When the language is OpenCL the following differences occur:

1. The HSA memory model is used (see [Memory Model](#)).

Assembler

AMDGPU backend has LLVM-MC based assembler which is currently in development. It supports AMDGCN GFX6-GFX8.

This section describes general syntax for instructions and operands. For more information about instructions, their semantics and supported combinations of operands, refer to one of instruction set architecture manuals [[AMD-Southern-Islands](#)] [[AMD-Sea-Islands](#)] [[AMD-Volcanic-Islands](#)].

An instruction has the following syntax (register operands are normally comma-separated while extra operands are space-separated):

```
<opcode> <register_operand0>, ... <extra_operand0> ...
```

Operands

The following syntax for register operands is supported:

- SGPR registers: s0, ... or s[0], ...
- VGPR registers: v0, ... or v[0], ...
- TTMP registers: tmp0, ... or tmp[0], ...
- Special registers: exec (exec_lo, exec_hi), vcc (vcc_lo, vcc_hi), flat_scratch (flat_scratch_lo, flat_scratch_hi)
- Special trap registers: tba (tba_lo, tba_hi), tma (tma_lo, tma_hi)
- Register pairs, quads, etc: s[2:3], v[10:11], tmp[5:6], s[4:7], v[12:15], tmp[4:7], s[8:15], ...
- Register lists: [s0, s1], [tmp0, tmp1, tmp2, tmp3]
- Register index expressions: v[2*2], s[1-1:2-1]
- 'off' indicates that an operand is not enabled

The following extra operands are supported:

- offset, offset0, offset1
- idxen, offen bits
- glc, slc, tfe bits
- waitcnt: integer or combination of counter values
- VOP3 modifiers:
 - abs (| |), neg (-)
- DPP modifiers:
 - row_shl, row_shr, row_ror, row_rol
 - row_mirror, row_half_mirror, row_bcast
 - wave_shl, wave_shr, wave_ror, wave_rol, quad_perm
 - row_mask, bank_mask, bound_ctrl
- SDWA modifiers:

- dst_sel, src0_sel, src1_sel (BYTE_N, WORD_M, DWORD)
- dst_unused (UNUSED_PAD, UNUSED_SEXT, UNUSED_PRESERVE)
- abs, neg, sext

Instruction Examples

DS

```
ds_add_u32 v2, v4 offset:16
ds_write_src2_b64 v2 offset0:4 offset1:8
ds_cmpst_f32 v2, v4, v6
ds_min_rtn_f64 v[8:9], v2, v[4:5]
```

For full list of supported instructions, refer to “LDS/GDS instructions” in ISA Manual.

FLAT

```
flat_load_dword v1, v[3:4]
flat_store_dwordx3 v[3:4], v[5:7]
flat_atomic_swap v1, v[3:4], v5 glc
flat_atomic_cmpswap v1, v[3:4], v[5:6] glc slc
flat_atomic_fmax_x2 v[1:2], v[3:4], v[5:6] glc
```

For full list of supported instructions, refer to “FLAT instructions” in ISA Manual.

MUBUF

```
buffer_load_dword v1, off, s[4:7], s1
buffer_store_dwordx4 v[1:4], v2, tmp[4:7], s1 offen offset:4 glc tfe
buffer_store_format_xy v[1:2], off, s[4:7], s1
buffer_wbinvl1
buffer_atomic_inc v1, v2, s[8:11], s4 idxen offset:4 slc
```

For full list of supported instructions, refer to “MUBUF Instructions” in ISA Manual.

SMRD/SMEM

```
s_load_dword s1, s[2:3], 0xfc
s_load_dwordx8 s[8:15], s[2:3], s4
s_load_dwordx16 s[88:103], s[2:3], s4
s_dcache_inv_vol
s_memtime s[4:5]
```

For full list of supported instructions, refer to “Scalar Memory Operations” in ISA Manual.

SOP1

```
s_mov_b32 s1, s2
s_mov_b64 s[0:1], 0x80000000
s_cmov_b32 s1, 200
s_wqm_b64 s[2:3], s[4:5]
s_bcmt0_i32_b64 s1, s[2:3]
s_swappc_b64 s[2:3], s[4:5]
s_cbranch_join s[4:5]
```

For full list of supported instructions, refer to “SOP1 Instructions” in ISA Manual.

SOP2

```
s_add_u32 s1, s2, s3
s_and_b64 s[2:3], s[4:5], s[6:7]
s_cselect_b32 s1, s2, s3
s_andn2_b32 s2, s4, s6
s_lshr_b64 s[2:3], s[4:5], s6
s_ashr_i32 s2, s4, s6
s_bfm_b64 s[2:3], s4, s6
s_bfe_i64 s[2:3], s[4:5], s6
s_cbranch_g_fork s[4:5], s[6:7]
```

For full list of supported instructions, refer to “SOP2 Instructions” in ISA Manual.

SOPC

```
s_cmp_eq_i32 s1, s2
s_bitcmp1_b32 s1, s2
s_bitcmp0_b64 s[2:3], s4
s_setvskip s3, s5
```

For full list of supported instructions, refer to “SOPC Instructions” in ISA Manual.

SOPP

```
s_barrier
s_nop 2
s_endpgm
s_waitcnt 0 ; Wait for all counters to be 0
s_waitcnt vmcnt(0) & expcnt(0) & lgkmcnt(0) ; Equivalent to above
s_waitcnt vmcnt(1) ; Wait for vmcnt counter to be 1.
s_sethalt 9
s_sleep 10
s_sendmsg 0x1
s_sendmsg sendmsg(MSG_INTERRUPT)
s_trap 1
```

For full list of supported instructions, refer to “SOPP Instructions” in ISA Manual.

Unless otherwise mentioned, little verification is performed on the operands of SOPP Instructions, so it is up to the programmer to be familiar with the range or acceptable values.

VALU

For vector ALU instruction opcodes (VOP1, VOP2, VOP3, VOPC, VOP_DPP, VOP_SDWA), the assembler will automatically use optimal encoding based on its operands. To force specific encoding, one can add a suffix to the opcode of the instruction:

- `_e32` for 32-bit VOP1/VOP2/VOPC
- `_e64` for 64-bit VOP3
- `_dpp` for VOP_DPP
- `_sdwa` for VOP_SDWA

VOP1/VOP2/VOP3/VOPC examples:

```
v_mov_b32 v1, v2
v_mov_b32_e32 v1, v2
v_nop
v_cvt_f64_i32_e32 v[1:2], v2
v_floor_f32_e32 v1, v2
v_bfrev_b32_e32 v1, v2
v_add_f32_e32 v1, v2, v3
v_mul_i32_i24_e64 v1, v2, 3
v_mul_i32_i24_e32 v1, -3, v3
v_mul_i32_i24_e32 v1, -100, v3
v_addc_u32 v1, s[0:1], v2, v3, s[2:3]
v_max_f16_e32 v1, v2, v3
```

VOP_DPP examples:

```
v_mov_b32 v0, v0 quad_perm:[0,2,1,1]
v_sin_f32 v0, v0 row_shl:1 row_mask:0xa bank_mask:0x1 bound_ctrl:0
v_mov_b32 v0, v0 wave_shl:1
v_mov_b32 v0, v0 row_mirror
v_mov_b32 v0, v0 row_bcast:31
v_mov_b32 v0, v0 quad_perm:[1,3,0,1] row_mask:0xa bank_mask:0x1 bound_ctrl:0
v_add_f32 v0, v0, |v0| row_shl:1 row_mask:0xa bank_mask:0x1 bound_ctrl:0
v_max_f16 v1, v2, v3 row_shl:1 row_mask:0xa bank_mask:0x1 bound_ctrl:0
```

VOP_SDWA examples:

```
v_mov_b32 v1, v2 dst_sel:BYTE_0 dst_unused:UNUSED_PRESERVE src0_sel:DWORD
v_min_u32 v200, v200, v1 dst_sel:WORD_1 dst_unused:UNUSED_PAD src0_sel:BYTE_1 src1_
↪sel:DWORD
v_sin_f32 v0, v0 dst_unused:UNUSED_PAD src0_sel:WORD_1
v_fract_f32 v0, |v0| dst_sel:DWORD dst_unused:UNUSED_PAD src0_sel:WORD_1
v_cmpx_le_u32 vcc, v1, v2 src0_sel:BYTE_2 src1_sel:WORD_0
```

For full list of supported instructions, refer to “Vector ALU instructions”.

HSA Code Object Directives

AMDGPU ABI defines auxiliary data in output code object. In assembly source, one can specify them with assembler directives.

.hsa_code_object_version major, minor

major and *minor* are integers that specify the version of the HSA code object that will be generated by the assembler.

.hsa_code_object_isa [major, minor, stepping, vendor, arch]

major, *minor*, and *stepping* are all integers that describe the instruction set architecture (ISA) version of the assembly program.

vendor and *arch* are quoted strings. *vendor* should always be equal to “AMD” and *arch* should always be equal to “AMDGPU”.

By default, the assembler will derive the ISA version, *vendor*, and *arch* from the value of the `-mcpu` option that is passed to the assembler.

.amdgpu_hsa_kernel (name)

This directive specifies that the symbol with given name is a kernel entry point (label) and the object should contain corresponding symbol of type `STT_AMDGPU_HSA_KERNEL`.

.amd_kernel_code_t

This directive marks the beginning of a list of key / value pairs that are used to specify the `amd_kernel_code_t` object that will be emitted by the assembler. The list must be terminated by the `.end_amd_kernel_code_t` directive. For any `amd_kernel_code_t` values that are unspecified a default value will be used. The default value for all keys is 0, with the following exceptions:

- *kernel_code_version_major* defaults to 1.
- *machine_kind* defaults to 1.
- *machine_version_major*, *machine_version_minor*, and *machine_version_stepping* are derived from the value of the `-mcpu` option that is passed to the assembler.
- *kernel_code_entry_byte_offset* defaults to 256.
- *wavefront_size* defaults to 6.
- *kernarg_segment_alignment*, *group_segment_alignment*, and *private_segment_alignment* default to 4. Note that alignments are specified as a power of two, so a value of *n* means an alignment of 2^n .

The `.amd_kernel_code_t` directive must be placed immediately after the function label and before any instructions.

For a full list of `amd_kernel_code_t` keys, refer to AMDGPU ABI document, comments in `lib/Target/AMDGPU/AmdKernelCodeT.h` and `test/CodeGen/AMDGPU/hsa.s`.

Here is an example of a minimal `amd_kernel_code_t` specification:

```
.hsa_code_object_version 1,0
.hsa_code_object_isa

.hsatext
.globl hello_world
.p2align 8
.amdgpu_hsa_kernel hello_world

hello_world:
```

```
.amd_kernel_code_t
  enable_sgpr_kernarg_segment_ptr = 1
  is_ptr64 = 1
  compute_pgm_rsrc1_vgprs = 0
  compute_pgm_rsrc1_sgprs = 0
  compute_pgm_rsrc2_user_sgpr = 2
  kernarg_segment_byte_size = 8
  wavefront_sgpr_count = 2
  workitem_vgpr_count = 3
.end_amd_kernel_code_t

s_load_dwordx2 s[0:1], s[0:1] 0x0
v_mov_b32 v0, 3.14159
s_waitcnt lgkmcnt(0)
v_mov_b32 v1, s0
v_mov_b32 v2, s1
flat_store_dword v[1:2], v0
s_endpgm
.Lfunc_end0:
  .size  hello_world, .Lfunc_end0-hello_world
```

4.24.6 Additional Documentation

4.25 Stack maps and patch points in LLVM

- *Definitions*
- *Motivation*
- *Intrinsics*
 - ‘*llvm.experimental.stackmap*’ Intrinsic
 - ‘*llvm.experimental.patchpoint.**’ Intrinsic
- *Stack Map Format*
 - *Stack Map Section*
- *Stack Map Usage*
 - *Direct Stack Map Entries*
- *Supported Architectures*

4.25.1 Definitions

In this document we refer to the “runtime” collectively as all components that serve as the LLVM client, including the LLVM IR generator, object code consumer, and code patcher.

A stack map records the location of `live` values at a particular instruction address. These `live` values do not refer to all the LLVM values live across the stack map. Instead, they are only the values that the runtime requires to be live at this point. For example, they may be the values the runtime will need to resume program execution at that point independent of the compiled function containing the stack map.

LLVM emits stack map data into the object code within a designated *Stack Map Section*. This stack map data contains a record for each stack map. The record stores the stack map's instruction address and contains an entry for each mapped value. Each entry encodes a value's location as a register, stack offset, or constant.

A patch point is an instruction address at which space is reserved for patching a new instruction sequence at run time. Patch points look much like calls to LLVM. They take arguments that follow a calling convention and may return a value. They also imply stack map generation, which allows the runtime to locate the patchpoint and find the location of live values at that point.

4.25.2 Motivation

This functionality is currently experimental but is potentially useful in a variety of settings, the most obvious being a runtime (JIT) compiler. Example applications of the patchpoint intrinsics are implementing an inline call cache for polymorphic method dispatch or optimizing the retrieval of properties in dynamically typed languages such as JavaScript.

The intrinsics documented here are currently used by the JavaScript compiler within the open source WebKit project, see the [FTL JIT](#), but they are designed to be used whenever stack maps or code patching are needed. Because the intrinsics have experimental status, compatibility across LLVM releases is not guaranteed.

The stack map functionality described in this document is separate from the functionality described in *Computing stack maps*. *GCFUNCTIONMetadata* provides the location of pointers into a collected heap captured by the *GCRoot* intrinsic, which can also be considered a “stack map”. Unlike the stack maps defined above, the *GCFUNCTIONMetadata* stack map interface does not provide a way to associate live register values of arbitrary type with an instruction address, nor does it specify a format for the resulting stack map. The stack maps described here could potentially provide richer information to a garbage collecting runtime, but that usage will not be discussed in this document.

4.25.3 Intrinsics

The following two kinds of intrinsics can be used to implement stack maps and patch points: `llvm.experimental.stackmap` and `llvm.experimental.patchpoint`. Both kinds of intrinsics generate a stack map record, and they both allow some form of code patching. They can be used independently (i.e. `llvm.experimental.patchpoint` implicitly generates a stack map without the need for an additional call to `llvm.experimental.stackmap`). The choice of which to use depends on whether it is necessary to reserve space for code patching and whether any of the intrinsic arguments should be lowered according to calling conventions. `llvm.experimental.stackmap` does not reserve any space, nor does it expect any call arguments. If the runtime patches code at the stack map's address, it will destructively overwrite the program text. This is unlike `llvm.experimental.patchpoint`, which reserves space for in-place patching without overwriting surrounding code. The `llvm.experimental.patchpoint` intrinsic also lowers a specified number of arguments according to its calling convention. This allows patched code to make in-place function calls without marshaling.

Each instance of one of these intrinsics generates a stack map record in the *Stack Map Section*. The record includes an ID, allowing the runtime to uniquely identify the stack map, and the offset within the code from the beginning of the enclosing function.

'`llvm.experimental.stackmap`' Intrinsic

Syntax:

```
declare void
  @llvm.experimental.stackmap(i64 <id>, i32 <numShadowBytes>, ...)
```

Overview:

The `llvm.experimental.stackmap` intrinsic records the location of specified values in the stack map without generating any code.

Operands:

The first operand is an ID to be encoded within the stack map. The second operand is the number of shadow bytes following the intrinsic. The variable number of operands that follow are the `live` values for which locations will be recorded in the stack map.

To use this intrinsic as a bare-bones stack map, with no code patching support, the number of shadow bytes can be set to zero.

Semantics:

The stack map intrinsic generates no code in place, unless nops are needed to cover its shadow (see below). However, its offset from function entry is stored in the stack map. This is the relative instruction address immediately following the instructions that precede the stack map.

The stack map ID allows a runtime to locate the desired stack map record. LLVM passes this ID through directly to the stack map record without checking uniqueness.

LLVM guarantees a shadow of instructions following the stack map's instruction offset during which neither the end of the basic block nor another call to `llvm.experimental.stackmap` or `llvm.experimental.patchpoint` may occur. This allows the runtime to patch the code at this point in response to an event triggered from outside the code. The code for instructions following the stack map may be emitted in the stack map's shadow, and these instructions may be overwritten by destructive patching. Without shadow bytes, this destructive patching could overwrite program text or data outside the current function. We disallow overlapping stack map shadows so that the runtime does not need to consider this corner case.

For example, a stack map with 8 byte shadow:

```
call void @runtime()
call void (i64, i32, ...) * @llvm.experimental.stackmap(i64 77, i32 8,
                                                         i64* %ptr)

%val = load i64* %ptr
%add = add i64 %val, 3
ret i64 %add
```

May require one byte of nop-padding:

```
0x00 callq _runtime
0x05 nop          <--- stack map address
0x06 movq (%rdi), %rax
0x07 addq $3, %rax
0x0a popq %rdx
0x0b ret          <---- end of 8-byte shadow
```

Now, if the runtime needs to invalidate the compiled code, it may patch 8 bytes of code at the stack map's address at follows:

```
0x00 callq _runtime
0x05 movl $0xffff, %rax <--- patched code at stack map address
0x0a callq *%rax        <---- end of 8-byte shadow
```

This way, after the normal call to the runtime returns, the code will execute a patched call to a special entry point that can rebuild a stack frame from the values located by the stack map.

`'llvm.experimental.patchpoint.*'` Intrinsic

Syntax:

```
declare void
  @llvm.experimental.patchpoint.void(i64 <id>, i32 <numBytes>,
                                     i8* <target>, i32 <numArgs>, ...)
declare i64
  @llvm.experimental.patchpoint.i64(i64 <id>, i32 <numBytes>,
                                    i8* <target>, i32 <numArgs>, ...)
```

Overview:

The `'llvm.experimental.patchpoint.*'` intrinsics creates a function call to the specified `<target>` and records the location of specified values in the stack map.

Operands:

The first operand is an ID, the second operand is the number of bytes reserved for the patchable region, the third operand is the target address of a function (optionally null), and the fourth operand specifies how many of the following variable operands are considered function call arguments. The remaining variable number of operands are the `live values` for which locations will be recorded in the stack map.

Semantics:

The patch point intrinsic generates a stack map. It also emits a function call to the address specified by `<target>` if the address is not a constant null. The function call and its arguments are lowered according to the calling convention specified at the intrinsic's callsite. Variants of the intrinsic with non-void return type also return a value according to calling convention.

On PowerPC, note that `<target>` must be the ABI function pointer for the intended target of the indirect call. Specifically, when compiling for the ELF V1 ABI, `<target>` is the function-descriptor address normally used as the C/C++ function-pointer representation.

Requesting zero patch point arguments is valid. In this case, all variable operands are handled just like `llvm.experimental.stackmap.*`. The difference is that space will still be reserved for patching, a call will be emitted, and a return value is allowed.

The location of the arguments are not normally recorded in the stack map because they are already fixed by the calling convention. The remaining `live values` will have their location recorded, which could be a register, stack location, or constant. A special calling convention has been introduced for use with stack maps, `anyregcc`, which forces the arguments to be loaded into registers but allows those register to be dynamically allocated. These argument registers will have their register locations recorded in the stack map in addition to the remaining `live values`.

The patch point also emits nops to cover at least `<numBytes>` of instruction encoding space. Hence, the client must ensure that `<numBytes>` is enough to encode a call to the target address on the supported targets. If the call target is constant null, then there is no minimum requirement. A zero-byte null target patchpoint is valid.

The runtime may patch the code emitted for the patch point, including the call sequence and nops. However, the runtime may not assume anything about the code LLVM emits within the reserved space. Partial patching is not allowed. The runtime must patch all reserved bytes, padding with nops if necessary.

This example shows a patch point reserving 15 bytes, with one argument in `$rdi`, and a return value in `$rax` per native calling convention:

```
%target = inttoptr i64 -281474976710654 to i8*
%val = call i64 @i64, i32, ...)*
      @llvm.experimental.patchpoint.i64(i64 78, i32 15,
                                         i8* %target, i32 1, i64* %ptr)
%add = add i64 %val, 3
ret i64 %add
```

May generate:

```
0x00 movabsq $0xffff000000000002, %r11 <--- patch point address
0x0a callq   *%r11
0x0d nop
0x0e nop
0x0f addq    $0x3, %rax
0x10 movl    %rax, 8(%rsp)
```

Note that no stack map locations will be recorded. If the patched code sequence does not need arguments fixed to specific calling convention registers, then the `anyregcc` convention may be used:

```
%val = call anyregcc @llvm.experimental.patchpoint(i64 78, i32 15,
                                                    i8* %target, i32 1,
                                                    i64* %ptr)
```

The stack map now indicates the location of the `%ptr` argument and return value:

```
Stack Map: ID=78, Loc0=%r9 Loc1=%r8
```

The patch code sequence may now use the argument that happened to be allocated in `%r8` and return a value allocated in `%r9`:

```
0x00 movslq 4(%r8) %r9 <--- patched code at patch point address
0x03 nop
...
0x0e nop <--- end of reserved 15-bytes
0x0f addq    $0x3, %r9
0x10 movl    %r9, 8(%rsp)
```

4.25.4 Stack Map Format

The existence of a stack map or patch point intrinsic within an LLVM Module forces code emission to create a *Stack Map Section*. The format of this section follows:

```
Header {
  uint8 : Stack Map Version (current version is 3)
  uint8 : Reserved (expected to be 0)
  uint16 : Reserved (expected to be 0)
}
uint32 : NumFunctions
uint32 : NumConstants
uint32 : NumRecords
```

```

StkSizeRecord[NumFunctions] {
    uint64 : Function Address
    uint64 : Stack Size
    uint64 : Record Count
}
Constants[NumConstants] {
    uint64 : LargeConstant
}
StkMapRecord[NumRecords] {
    uint64 : PatchPoint ID
    uint32 : Instruction Offset
    uint16 : Reserved (record flags)
    uint16 : NumLocations
    Location[NumLocations] {
        uint8 : Register | Direct | Indirect | Constant | ConstantIndex
        uint8 : Reserved (expected to be 0)
        uint16 : Location Size
        uint16 : Dwarf RegNum
        uint16 : Reserved (expected to be 0)
        int32 : Offset or SmallConstant
    }
    uint32 : Padding (only if required to align to 8 byte)
    uint16 : Padding
    uint16 : NumLiveOuts
    LiveOuts[NumLiveOuts]
        uint16 : Dwarf RegNum
        uint8 : Reserved
        uint8 : Size in Bytes
    }
    uint32 : Padding (only if required to align to 8 byte)
}

```

The first byte of each location encodes a type that indicates how to interpret the `RegNum` and `Offset` fields as follows:

Encoding	Type	Value	Description
0x1	Register	Reg	Value in a register
0x2	Direct	Reg + Offset	Frame index value
0x3	Indirect	[Reg + Offset]	Spilled value
0x4	Constant	Offset	Small constant
0x5	ConstIndex	Constants[Offset]	Large constant

In the common case, a value is available in a register, and the `Offset` field will be zero. Values spilled to the stack are encoded as `Indirect` locations. The runtime must load those values from a stack address, typically in the form `[BP + Offset]`. If an `alloca` value is passed directly to a stack map intrinsic, then LLVM may fold the frame index into the stack map as an optimization to avoid allocating a register or stack slot. These frame indices will be encoded as `Direct` locations in the form `BP + Offset`. LLVM may also optimize constants by emitting them directly in the stack map, either in the `Offset` of a `Constant` location or in the constant pool, referred to by `ConstantIndex` locations.

At each callsite, a “liveout” register list is also recorded. These are the registers that are live across the stackmap and therefore must be saved by the runtime. This is an important optimization when the patchpoint intrinsic is used with a calling convention that by default preserves most registers as callee-save.

Each entry in the liveout register list contains a DWARF register number and size in bytes. The stackmap format deliberately omits specific subregister information. Instead the runtime must interpret this information conservatively.

For example, if the stackmap reports one byte at `%rax`, then the value may be in either `%al` or `%ah`. It doesn't matter in practice, because the runtime will simply save `%rax`. However, if the stackmap reports 16 bytes at `%ymm0`, then the runtime can safely optimize by saving only `%xmm0`.

The stack map format is a contract between an LLVM SVN revision and the runtime. It is currently experimental and may change in the short term, but minimizing the need to update the runtime is important. Consequently, the stack map design is motivated by simplicity and extensibility. Compactness of the representation is secondary because the runtime is expected to parse the data immediately after compiling a module and encode the information in its own format. Since the runtime controls the allocation of sections, it can reuse the same stack map space for multiple modules.

Stackmap support is currently only implemented for 64-bit platforms. However, a 32-bit implementation should be able to use the same format with an insignificant amount of wasted space.

Stack Map Section

A JIT compiler can easily access this section by providing its own memory manager via the LLVM C API `LLVMCreateSimpleMCJITMemoryManager()`. When creating the memory manager, the JIT provides a callback: `LLVMMemoryManagerAllocateDataSectionCallback()`. When LLVM creates this section, it invokes the callback and passes the section name. The JIT can record the in-memory address of the section at this time and later parse it to recover the stack map data.

On Darwin, the stack map section name is “`__llvm_stackmaps`”. The segment name is “`__LLVM_STACKMAPS`”.

4.25.5 Stack Map Usage

The stack map support described in this document can be used to precisely determine the location of values at a specific position in the code. LLVM does not maintain any mapping between those values and any higher-level entity. The runtime must be able to interpret the stack map record given only the ID, offset, and the order of the locations, records, and functions, which LLVM preserves.

Note that this is quite different from the goal of debug information, which is a best-effort attempt to track the location of named variables at every instruction.

An important motivation for this design is to allow a runtime to commandeer a stack frame when execution reaches an instruction address associated with a stack map. The runtime must be able to rebuild a stack frame and resume program execution using the information provided by the stack map. For example, execution may resume in an interpreter or a recompiled version of the same function.

This usage restricts LLVM optimization. Clearly, LLVM must not move stores across a stack map. However, loads must also be handled conservatively. If the load may trigger an exception, hoisting it above a stack map could be invalid. For example, the runtime may determine that a load is safe to execute without a type check given the current state of the type system. If the type system changes while some activation of the load's function exists on the stack, the load becomes unsafe. The runtime can prevent subsequent execution of that load by immediately patching any stack map location that lies between the current call site and the load (typically, the runtime would simply patch all stack map locations to invalidate the function). If the compiler had hoisted the load above the stack map, then the program could crash before the runtime could take back control.

To enforce these semantics, stackmap and patchpoint intrinsics are considered to potentially read and write all memory. This may limit optimization more than some clients desire. This limitation may be avoided by marking the call site as “readonly”. In the future we may also allow meta-data to be added to the intrinsic call to express aliasing, thereby allowing optimizations to hoist certain loads above stack maps.

Direct Stack Map Entries

As shown in *Stack Map Section*, a Direct stack map location records the address of frame index. This address is itself the value that the runtime requested. This differs from Indirect locations, which refer to a stack locations from which the requested values must be loaded. Direct locations can communicate the address if an alloca, while Indirect locations handle register spills.

For example:

```
entry:
  %a = alloca i64...
  llvm.experimental.stackmap(i64 <ID>, i32 <shadowBytes>, i64* %a)
```

The runtime can determine this alloca's relative location on the stack immediately after compilation, or at any time thereafter. This differs from Register and Indirect locations, because the runtime can only read the values in those locations when execution reaches the instruction address of the stack map.

This functionality requires LLVM to treat entry-block allocas specially when they are directly consumed by an intrinsic. (This is the same requirement imposed by the `llvm.gcroot` intrinsic.) LLVM transformations must not substitute the alloca with any intervening value. This can be verified by the runtime simply by checking that the stack map's location is a Direct location type.

4.25.6 Supported Architectures

Support for StackMap generation and the related intrinsics requires some code for each backend. Today, only a subset of LLVM's backends are supported. The currently supported architectures are X86_64, PowerPC, and Aarch64.

4.26 Design and Usage of the InAlloca Attribute

4.26.1 Introduction

The *inalloca* attribute is designed to allow taking the address of an aggregate argument that is being passed by value through memory. Primarily, this feature is required for compatibility with the Microsoft C++ ABI. Under that ABI, class instances that are passed by value are constructed directly into argument stack memory. Prior to the addition of *inalloca*, calls in LLVM were indivisible instructions. There was no way to perform intermediate work, such as object construction, between the first stack adjustment and the final control transfer. With *inalloca*, all arguments passed in memory are modelled as a single alloca, which can be stored to prior to the call. Unfortunately, this complicated feature comes with a large set of restrictions designed to bound the lifetime of the argument memory around the call.

For now, it is recommended that frontends and optimizers avoid producing this construct, primarily because it forces the use of a base pointer. This feature may grow in the future to allow general mid-level optimization, but for now, it should be regarded as less efficient than passing by value with a copy.

4.26.2 Intended Usage

The example below is the intended LLVM IR lowering for some C++ code that passes two default-constructed `Foo` objects to `g` in the 32-bit Microsoft C++ ABI.

```
// Foo is non-trivial.
struct Foo { int a, b; Foo(); ~Foo(); Foo(const Foo &); };
void g(Foo a, Foo b);
void f() {
```

```
g(Foo(), Foo());  
}
```

```
%struct.Foo = type { i32, i32 }  
declare void @Foo_ctor(%struct.Foo* %this)  
declare void @Foo_dtor(%struct.Foo* %this)  
declare void @g(<{ %struct.Foo, %struct.Foo }>* inalloca %memargs)  
  
define void @f() {  
entry:  
  %base = call i8* @llvm.stacksave()  
  %memargs = alloca <{ %struct.Foo, %struct.Foo }>  
  %b = getelementptr <{ %struct.Foo, %struct.Foo }>* %memargs, i32 1  
  call void @Foo_ctor(%struct.Foo* %b)  
  
  ; If a's ctor throws, we must destruct b.  
  %a = getelementptr <{ %struct.Foo, %struct.Foo }>* %memargs, i32 0  
  invoke void @Foo_ctor(%struct.Foo* %a)  
    to label %invoke.cont unwind %invoke.unwind  
  
invoke.cont:  
  call void @g(<{ %struct.Foo, %struct.Foo }>* inalloca %memargs)  
  call void @llvm.stackrestore(i8* %base)  
  ...  
  
invoke.unwind:  
  call void @Foo_dtor(%struct.Foo* %b)  
  call void @llvm.stackrestore(i8* %base)  
  ...  
}
```

To avoid stack leaks, the frontend saves the current stack pointer with a call to *llvm.stacksave*. Then, it allocates the argument stack space with *alloca* and calls the default constructor. The default constructor could throw an exception, so the frontend has to create a landing pad. The frontend has to destroy the already constructed argument *b* before restoring the stack pointer. If the constructor does not unwind, *g* is called. In the Microsoft C++ ABI, *g* will destroy its arguments, and then the stack is restored in *f*.

4.26.3 Design Considerations

Lifetime

The biggest design consideration for this feature is object lifetime. We cannot model the arguments as static allocas in the entry block, because all calls need to use the memory at the top of the stack to pass arguments. We cannot vend pointers to that memory at function entry because after code generation they will alias.

The rule against allocas between argument allocations and the call site avoids this problem, but it creates a cleanup problem. Cleanup and lifetime is handled explicitly with stack save and restore calls. In the future, we may want to introduce a new construct such as *freea* or *afree* to make it clear that this stack adjusting cleanup is less powerful than a full stack save and restore.

Nested Calls and Copy Elision

We also want to be able to support copy elision into these argument slots. This means we have to support multiple live argument allocations.

Consider the evaluation of:

```
// Foo is non-trivial.
struct Foo { int a; Foo(); Foo(const &Foo); ~Foo(); };
Foo bar(Foo b);
int main() {
    bar(bar(Foo()));
}
```

In this case, we want to be able to elide copies into `bar`'s argument slots. That means we need to have more than one set of argument frames active at the same time. First, we need to allocate the frame for the outer call so we can pass it in as the hidden struct return pointer to the middle call. Then we do the same for the middle call, allocating a frame and passing its address to `Foo`'s default constructor. By wrapping the evaluation of the inner `bar` with stack save and restore, we can have multiple overlapping active call frames.

Callee-cleanup Calling Conventions

Another wrinkle is the existence of callee-cleanup conventions. On Windows, all methods and many other functions adjust the stack to clear the memory used to pass their arguments. In some sense, this means that the allocas are automatically cleared by the call. However, LLVM instead models this as a write of undef to all of the `inalloca` values passed to the call instead of a stack adjustment. Frontends should still restore the stack pointer to avoid a stack leak.

Exceptions

There is also the possibility of an exception. If argument evaluation or copy construction throws an exception, the landing pad must do cleanup, which includes adjusting the stack pointer to avoid a stack leak. This means the cleanup of the stack memory cannot be tied to the call itself. There needs to be a separate IR-level instruction that can perform independent cleanup of arguments.

Efficiency

Eventually, it should be possible to generate efficient code for this construct. In particular, using `inalloca` should not require a base pointer. If the backend can prove that all points in the CFG only have one possible stack level, then it can address the stack directly from the stack pointer. While this is not yet implemented, the plan is that the `inalloca` attribute should not change much, but the frontend IR generation recommendations may change.

4.27 Using ARM NEON instructions in big endian mode

- *Introduction*
 - *Example: C-level intrinsics -> assembly*
- *Problem*
- *LDR and LD1*
- *Considerations*
 - *LLVM IR Lane ordering*
 - *AAPCS*

- *Alignment*
- *Summary*
- *Implementation*
- *Bitconverts*

4.27.1 Introduction

Generating code for big endian ARM processors is for the most part straightforward. NEON loads and stores however have some interesting properties that make code generation decisions less obvious in big endian mode.

The aim of this document is to explain the problem with NEON loads and stores, and the solution that has been implemented in LLVM.

In this document the term “vector” refers to what the ARM ABI calls a “short vector”, which is a sequence of items that can fit in a NEON register. This sequence can be 64 or 128 bits in length, and can constitute 8, 16, 32 or 64 bit items. This document refers to A64 instructions throughout, but is almost applicable to the A32/ARMv7 instruction sets also. The ABI format for passing vectors in A32 is slightly different to A64. Apart from that, the same concepts apply.

Example: C-level intrinsics -> assembly

It may be helpful first to illustrate how C-level ARM NEON intrinsics are lowered to instructions.

This trivial C function takes a vector of four ints and sets the zero'th lane to the value “42”:

```
#include <arm_neon.h>
int32x4_t f(int32x4_t p) {
    return vsetq_lane_s32(42, p, 0);
}
```

arm_neon.h intrinsics generate “generic” IR where possible (that is, normal IR instructions not `llvm.arm.neon.*` intrinsic calls). The above generates:

```
define <4 x i32> @f(<4 x i32> %p) {
    %vset_lane = insertelement <4 x i32> %p, i32 42, i32 0
    ret <4 x i32> %vset_lane
}
```

Which then becomes the following trivial assembly:

```
f:                                // @f
    movz        w8, #0x2a
    ins         v0.s[0], w8
    ret
```

4.27.2 Problem

The main problem is how vectors are represented in memory and in registers.

First, a recap. The “endianness” of an item affects its representation in memory only. In a register, a number is just a sequence of bits - 64 bits in the case of AArch64 general purpose registers. Memory, however, is a sequence of

addressable units of 8 bits in size. Any number greater than 8 bits must therefore be split up into 8-bit chunks, and endianness describes the order in which these chunks are laid out in memory.

A “little endian” layout has the least significant byte first (lowest in memory address). A “big endian” layout has the *most* significant byte first. This means that when loading an item from big endian memory, the lowest 8-bits in memory must go in the most significant 8-bits, and so forth.

4.27.3 LDR and LD1

A vector is a consecutive sequence of items that are operated on simultaneously.

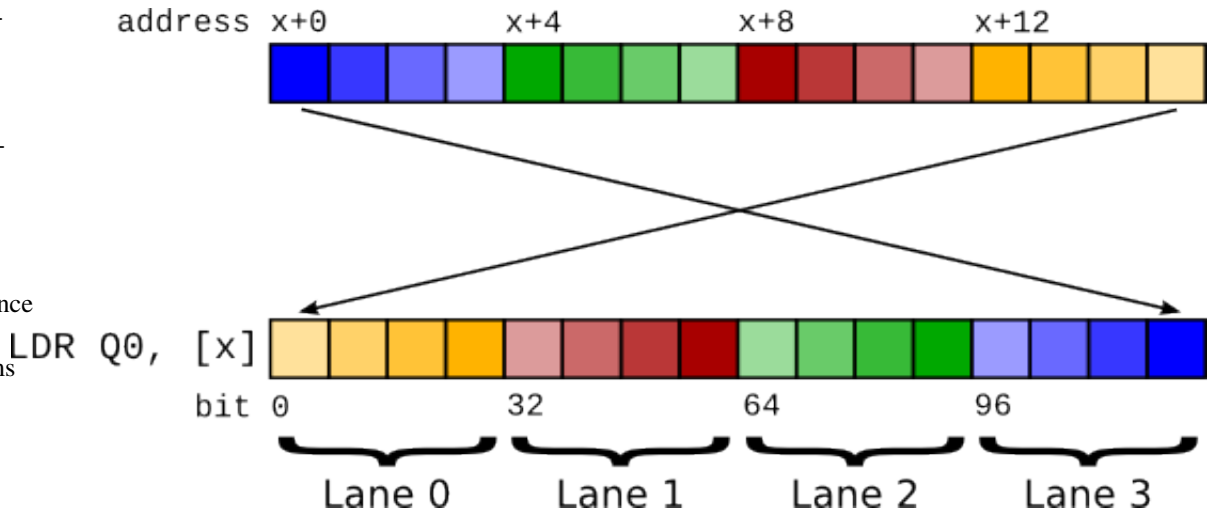


Fig. 4.1: Big endian vector load using LDR.

To load a 64-bit vector, 64 bits need to be read from memory. In little endian mode, we can do this by just performing a 64-bit load - `LDR q0, [foo]`. However if we try this in big endian mode, because of the byte swapping the lane indices end up being swapped! The zero'th item as laid out in memory becomes the n'th lane in the vector.

Because of this, the instruction `LD1` performs a vector load but performs byte swapping not on the entire 64 bits, but on the individual items within the vector. This means that the register content is the same as it would have been on a little endian system.

It may seem that `LD1` should suffice to perform vector loads on a big endian machine. However there are pros and cons to the two approaches that make it less than simple which register format to pick.

There are two options:

1. The content of a vector register is the same *as if* it had been loaded with an `LDR` instruction.
2. The content of a vector register is the same *as if* it had been loaded with an `LD1` instruction.

Because `LD1 == LDR + REV` and similarly `LDR == LD1 + REV` (on a big endian system), we can simulate either type of load with the other type of load plus a `REV` instruction. So we're not deciding which instructions to use, but which format to use (which will then influence which instruction is best to use).

Note that throughout this section we only mention loads. Stores have exactly the same problems as their associated loads, so have been skipped for brevity.

Fig. 4.2: Big endian vector load using LD1. Note that the lanes retain the correct ordering.

4.27.4 Considerations

LLVM IR Lane ordering

LLVM IR has first class vector types. In LLVM IR, the zero'th element of a vector resides at the lowest memory address. The optimizer relies on this property in certain areas, for example when concatenating vectors together. The intention is for arrays and vectors to have identical memory layouts - `[4 x i8]` and `<4 x i8>` should be represented the same in memory. Without this property there would be many special cases that the optimizer would have to cleverly handle.

Use of `LDR` would break this lane ordering property. This doesn't preclude the use of `LDR`, but we would have to do one of two things:

1. Insert a `REV` instruction to reverse the lane order after every `LDR`.
2. Disable all optimizations that rely on lane layout, and for every access to an individual lane (`insertelement`/`extractelement`/`shufflevector`) reverse the lane index.

AAPCS

The ARM procedure call standard (AAPCS) defines the ABI for passing vectors between functions in registers. It states:

When a short vector is transferred between registers and memory it is treated as an opaque object. That is a short vector is stored in memory as if it were stored with a single `STR` of the entire register; a short vector is loaded from memory using the corresponding `LDR` instruction. On a little-endian system this means that element 0 will always contain the lowest addressed element of a short vector; on a big-endian system element 0 will contain the highest-addressed element of a short vector.

—Procedure Call Standard for the ARM 64-bit Architecture (AArch64), 4.1.2 Short Vectors

The use of `LDR` and `STR` as the ABI defines has at least one advantage over `LD1` and `ST1`. `LDR` and `STR` are oblivious to the size of the individual lanes of a vector. `LD1` and `ST1` are not - the lane size is encoded within them. This is important across an ABI boundary, because it would become necessary to know the lane width the callee expects. Consider the following code:

```
<callee.c>
void callee(uint32x2_t v) {
    ...
}

<caller.c>
extern void callee(uint32x2_t);
void caller() {
    callee(...);
}
```

If `callee` changed its signature to `uint16x4_t`, which is equivalent in register content, if we passed as `LD1` we'd break this code until `caller` was updated and recompiled.

There is an argument that if the signatures of the two functions are different then the behaviour should be undefined. But there may be functions that are agnostic to the lane layout of the vector, and treating the vector as an opaque value (just loading it and storing it) would be impossible without a common format across ABI boundaries.

So to preserve ABI compatibility, we need to use the `LDR` lane layout across function calls.

Alignment

In strict alignment mode, `LDR qX` requires its address to be 128-bit aligned, whereas `LD1` only requires it to be as aligned as the lane size. If we canonicalised on using `LDR`, we'd still need to use `LD1` in some places to avoid alignment faults (the result of the `LD1` would then need to be reversed with `REV`).

Most operating systems however do not run with alignment faults enabled, so this is often not an issue.

Summary

The following table summarises the instructions that are required to be emitted for each property mentioned above for each of the two solutions.

	LDR layout	LD1 layout
Lane ordering	LDR + REV	LD1
AAPCS	LDR	LD1 + REV
Alignment for strict mode	LDR / LD1 + REV	LD1

Neither approach is perfect, and choosing one boils down to choosing the lesser of two evils. The issue with lane ordering, it was decided, would have to change target-agnostic compiler passes and would result in a strange IR in which lane indices were reversed. It was decided that this was worse than the changes that would have to be made to support `LD1`, so `LD1` was chosen as the canonical vector load instruction (and by inference, `ST1` for vector stores).

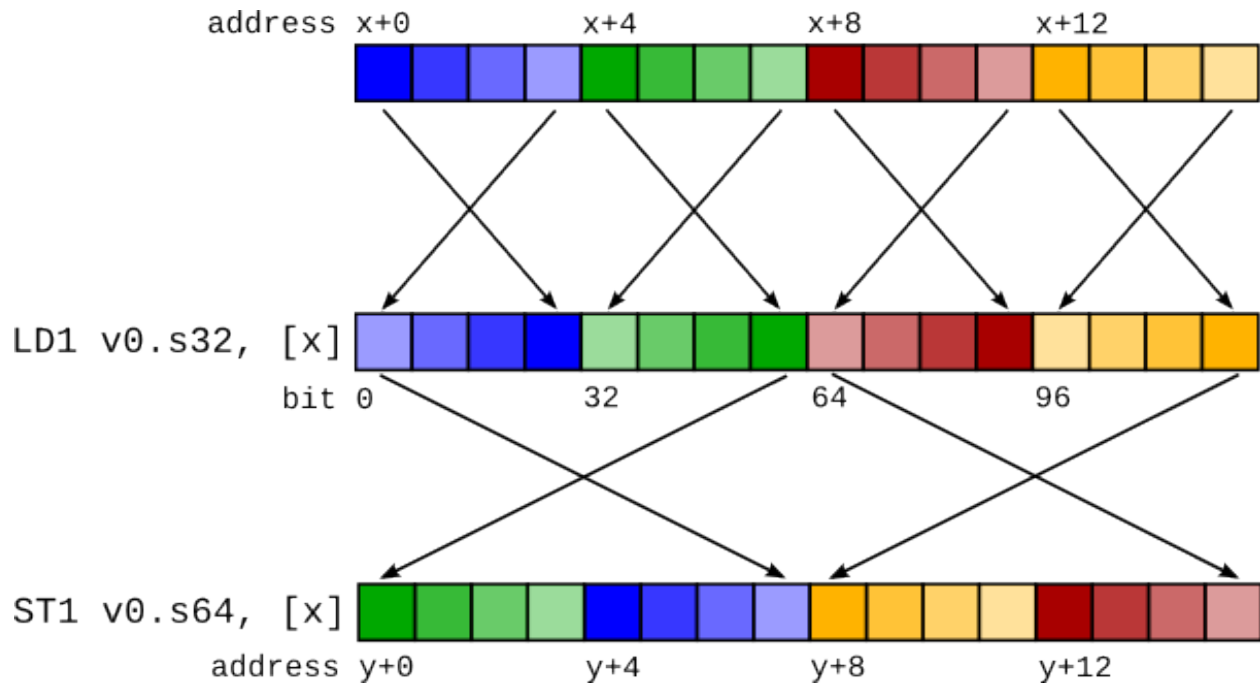
4.27.5 Implementation

There are 3 parts to the implementation:

1. Predicate `LDR` and `STR` instructions so that they are never allowed to be selected to generate vector loads and stores. The exception is one-lane vectors¹ - these by definition cannot have lane ordering problems so are fine to use `LDR/STR`.
2. Create code generation patterns for bitconverts that create `REV` instructions.
3. Make sure appropriate bitconverts are created so that vector values get passed over call boundaries as 1-element vectors (which is the same as if they were loaded with `LDR`).

¹ One lane vectors may seem useless as a concept but they serve to distinguish between values held in general purpose registers and values held in NEON/VFP registers. For example, an `i64` would live in an `x` register, but `<1 x i64>` would live in a `d` register.

Bitconverts



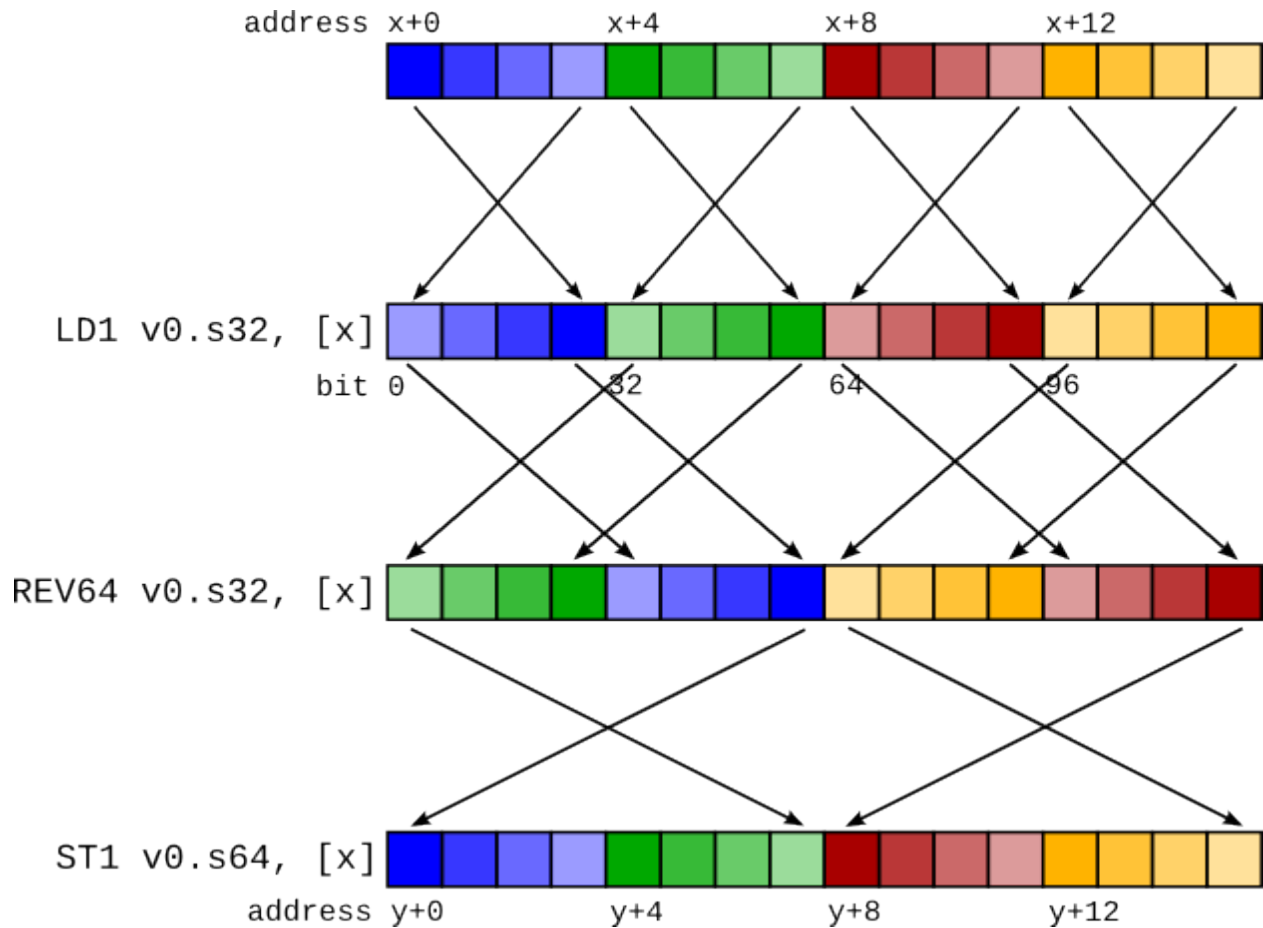
The main problem with the LD1 solution is dealing with bitconverts (or bitcasts, or reinterpret casts). These are pseudo instructions that only change the compiler's interpretation of data, not the underlying data itself. A requirement is that if data is loaded and then saved again (called a "round trip"), the memory contents should be the same after the store as before the load. If a vector is loaded and is then bitconverted to a different vector type before storing, the round trip will currently be broken.

Take for example this code sequence:

```
%0 = load <4 x i32> %x
%1 = bitcast <4 x i32> %0 to <2 x i64>
store <2 x i64> %1, <2 x i64>* %y
```

This would produce a code sequence such as that in the figure on the right. The mismatched LD1 and ST1 cause the stored data to differ from the loaded data.

When we see a bitcast from type X to type Y, what we need to do is to change the in-register representation of the data to be *as if* it had just been loaded by a LD1 of type Y.



Conceptually this is simple - we can insert a REV undoing the LD1 of type X (converting the in-register representation to the same as if it had been loaded by LDR) and then insert another REV to change the representation to be as if it had been loaded by an LD1 of type Y.

For the previous example, this would be:

```
LD1    v0.4s, [x]

REV64 v0.4s, v0.4s           // There is no REV128 instruction, so it must be
↪synthesized
EXT    v0.16b, v0.16b, v0.16b, #8 // with a REV64 then an EXT to swap the two 64-
↪bit elements.

REV64 v0.2d, v0.2d
EXT    v0.16b, v0.16b, v0.16b, #8

ST1    v0.2d, [y]
```

It turns out that these REV pairs can, in almost all cases, be squashed together into a single REV. For the example above, a REV128 4s + REV128 2d is actually a REV64 4s, as shown in the figure on the right.

4.28 LLVM Code Coverage Mapping Format

- *Introduction*
- *Quick Start*
- *High Level Overview*
- *Advanced Concepts*
 - *Mapping Region*
 - * *Source Range:*
 - * *File ID:*
 - * *Counter:*
- *LLVM IR Representation*
 - *Coverage Mapping Header:*
 - *Function record:*
 - *Encoded data:*
 - * *Dissecting the sample:*
- *Encoding*
 - *Types*
 - * *LEB128*
 - * *Strings*
 - *File ID Mapping*
 - *Counter*
 - * *Tag:*
 - * *Data:*
 - *Counter Expressions*
 - *Mapping Regions*
 - * *Sub-Array of Regions*
 - * *Mapping Region*
 - * *Header*
 - *Counter:*
 - *Pseudo-Counter:*
 - * *Source Range*

4.28.1 Introduction

LLVM's code coverage mapping format is used to provide code coverage analysis using LLVM's and Clang's instrumentation based profiling (Clang's `-fprofile-instr-generate` option).

This document is aimed at those who use LLVM's code coverage mapping to provide code coverage analysis for their own programs, and for those who would like to know how it works under the hood. A prior knowledge of how Clang's profile guided optimization works is useful, but not required.

We start by showing how to use LLVM and Clang for code coverage analysis, then we briefly describe LLVM's code coverage mapping format and the way that Clang and LLVM's code coverage tool work with this format. After the basics are down, more advanced features of the coverage mapping format are discussed - such as the data structures, LLVM IR representation and the binary encoding.

4.28.2 Quick Start

Here's a short story that describes how to generate code coverage overview for a sample source file called *test.c*.

- First, compile an instrumented version of your program using Clang's `-fprofile-instr-generate` option with the additional `-fcoverage-mapping` option:

```
clang -o test -fprofile-instr-generate -fcoverage-mapping test.c
```

- Then, run the instrumented binary. The runtime will produce a file called *default.profraw* containing the raw profile instrumentation data:

```
./test
```

- After that, merge the profile data using the *llvm-profdata* tool:

```
llvm-profdata merge -o test.profdata default.profraw
```

- Finally, run LLVM's code coverage tool (*llvm-cov*) to produce the code coverage overview for the sample source file:

```
llvm-cov show ./test -instr-profile=test.profdata test.c
```

4.28.3 High Level Overview

LLVM's code coverage mapping format is designed to be a self contained data format, that can be embedded into the LLVM IR and object files. It's described in this document as a **mapping** format because its goal is to store the data that is required for a code coverage tool to map between the specific source ranges in a file and the execution counts obtained after running the instrumented version of the program.

The mapping data is used in two places in the code coverage process:

1. When clang compiles a source file with `-fcoverage-mapping`, it generates the mapping information that describes the mapping between the source ranges and the profiling instrumentation counters. This information gets embedded into the LLVM IR and conveniently ends up in the final executable file when the program is linked.
2. It is also used by *llvm-cov* - the mapping information is extracted from an object file and is used to associate the execution counts (the values of the profile instrumentation counters), and the source ranges in a file. After that, the tool is able to generate various code coverage reports for the program.

The coverage mapping format aims to be a "universal format" that would be suitable for usage by any frontend, and not just by Clang. It also aims to provide the frontend the possibility of generating the minimal coverage mapping data in order to reduce the size of the IR and object files - for example, instead of emitting mapping information for each statement in a function, the frontend is allowed to group the statements with the same execution count into regions of code, and emit the mapping information only for those regions.

4.28.4 Advanced Concepts

The remainder of this guide is meant to give you insight into the way the coverage mapping format works.

The coverage mapping format operates on a per-function level as the profile instrumentation counters are associated with a specific function. For each function that requires code coverage, the frontend has to create coverage mapping data that can map between the source code ranges and the profile instrumentation counters for that function.

Mapping Region

The function's coverage mapping data contains an array of mapping regions. A mapping region stores the *source code range* that is covered by this region, the *file id*, the *coverage mapping counter* and the region's kind. There are several kinds of mapping regions:

- Code regions associate portions of source code and *coverage mapping counters*. They make up the majority of the mapping regions. They are used by the code coverage tool to compute the execution counts for lines, highlight the regions of code that were never executed, and to obtain the various code coverage statistics for a function. For example:
- Skipped regions are used to represent source ranges that were skipped by Clang's preprocessor. They don't associate with *coverage mapping counters*, as the frontend knows that they are never executed. They are used by the code coverage tool to mark the skipped lines inside a function as non-code lines that don't have execution counts. For example:
- Expansion regions are used to represent Clang's macro expansions. They have an additional property - *expanded file id*. This property can be used by the code coverage tool to find the mapping regions that are created as a result of this macro expansion, by checking if their file id matches the expanded file id. They don't associate with *coverage mapping counters*, as the code coverage tool can determine the execution count for this region by looking up the execution count of the first region with a corresponding file id. For example:

Source Range:

The source range record contains the starting and ending location of a certain mapping region. Both locations include the line and the column numbers.

File ID:

The file id an integer value that tells us in which source file or macro expansion is this region located. It enables Clang to produce mapping information for the code defined inside macros, like this example demonstrates:

Counter:

A coverage mapping counter can represents a reference to the profile instrumentation counter. The execution count for a region with such counter is determined by looking up the value of the corresponding profile instrumentation counter.

It can also represent a binary arithmetical expression that operates on coverage mapping counters or other expressions. The execution count for a region with an expression counter is determined by evaluating the expression's arguments and then adding them together or subtracting them from one another. In the example below, a subtraction expression is used to compute the execution count for the compound statement that follows the *else* keyword:

Finally, a coverage mapping counter can also represent an execution count of of zero. The zero counter is used to provide coverage mapping for unreachable statements and expressions, like in the example below:

The zero counters allow the code coverage tool to display proper line execution counts for the unreachable lines and highlight the unreachable code. Without them, the tool would think that those lines and regions were still executed, as it doesn't possess the frontend's knowledge.

4.28.5 LLVM IR Representation

The coverage mapping data is stored in the LLVM IR using a single global constant structure variable called `__llvm_coverage_mapping` with the `__llvm_covmap` section specifier.

For example, let's consider a C file and how it gets compiled to LLVM:

```
int foo() {
    return 42;
}
int bar() {
    return 13;
}
```

The coverage mapping variable generated by Clang has 3 fields:

- Coverage mapping header.
- An array of function records.
- Coverage mapping data which is an array of bytes. Zero paddings are added at the end to force 8 byte alignment.

```
@__llvm_coverage_mapping = internal constant { { i32, i32, i32, i32 }, [2 x { i64, i32, i64 }], [40 x i8] }
{
  { i32, i32, i32, i32 } ; Coverage map header
  {
    i32 2, ; The number of function records
    i32 20, ; The length of the string that contains the encoded translation unit_
    ↪filenames
    i32 20, ; The length of the string that contains the encoded coverage mapping data
    i32 1, ; Coverage mapping format version
  },
  [2 x { i64, i32, i64 }] [ ; Function records
    { i64, i32, i64 } {
      i64 0x5cf8c24cdb18bdac, ; Function's name MD5
      i32 9, ; Function's encoded coverage mapping data string length
      i64 0 ; Function's structural hash
    },
    { i64, i32, i64 } {
      i64 0xe413754a191db537, ; Function's name MD5
      i32 9, ; Function's encoded coverage mapping data string length
      i64 0 ; Function's structural hash
    },
  ],
  [40 x i8] c"..."; Encoded data (dissected later)
}, section "__llvm_covmap", align 8
```

The function record layout has evolved since version 1. In version 1, the function record for `foo` is defined as follows:

```
{ i8*, i32, i32, i64 } { i8* getelementptr inbounds ([3 x i8]* @__profn_foo, i32 0, i32 0), ; Function's name
  i32 3, ; Function's name length
  i32 9, ; Function's encoded coverage mapping data string length
  i64 0 ; Function's structural hash
}
```

Coverage Mapping Header:

The coverage mapping header has the following fields:

- The number of function records.
- The length of the string in the third field of `__llvm_coverage_mapping` that contains the encoded translation unit filenames.
- The length of the string in the third field of `__llvm_coverage_mapping` that contains the encoded coverage mapping data.
- The format version. The current version is 2 (encoded as a 1).

Function record:

A function record is a structure of the following type:

```
{ i64, i32, i64 }
```

It contains function name's MD5, the length of the encoded mapping data for that function, and function's structural hash value.

Encoded data:

The encoded data is stored in a single string that contains the encoded filenames used by this translation unit and the encoded coverage mapping data for each function in this translation unit.

The encoded data has the following structure:

```
[filenames, coverageMappingDataForFunctionRecord0, coverageMappingDataForFunctionRecord1, ..., padding]
```

If necessary, the encoded data is padded with zeroes so that the size of the data string is rounded up to the nearest multiple of 8 bytes.

Dissecting the sample:

Here's an overview of the encoded data that was stored in the IR for the *coverage mapping sample* that was shown earlier:

- The IR contains the following string constant that represents the encoded coverage mapping data for the sample translation unit:

```
c"\01\12\Users\alex\test.  
↪c\01\00\00\01\01\01\0C\02\02\01\00\00\01\01\04\0C\02\02\00\00"
```

- The string contains values that are encoded in the LEB128 format, which is used throughout for storing integers. It also contains a string value.
- The length of the substring that contains the encoded translation unit filenames is the value of the second field in the `__llvm_coverage_mapping` structure, which is 20, thus the filenames are encoded in this string:

```
c"\01\12\Users\alex\test.c"
```

This string contains the following data:

- Its first byte has a value of 0x01. It stores the number of filenames contained in this string.
- Its second byte stores the length of the first filename in this string.
- The remaining 18 bytes are used to store the first filename.
- The length of the substring that contains the encoded coverage mapping data for the first function is the value of the third field in the first structure in an array of *function records* stored in the third field of the `__llvm_coverage_mapping` structure, which is the 9. Therefore, the coverage mapping for the first function record is encoded in this string:

```
c"\01\00\00\01\01\01\0C\02\02"
```

This string consists of the following bytes:

0x01	The number of file ids used by this function. There is only one file id used by the mapping data in this function.
0x00	An index into the filenames array which corresponds to the file “/Users/alex/test.c”.
0x00	The number of counter expressions used by this function. This function doesn’t use any expressions.
0x01	The number of mapping regions that are stored in an array for the function’s file id #0.
0x01	The coverage mapping counter for the first region in this function. The value of 1 tells us that it’s a coverage mapping counter that is a reference to the profile instrumentation counter with an index of 0.
0x01	The starting line of the first mapping region in this function.
0x0C	The starting column of the first mapping region in this function.
0x02	The ending line of the first mapping region in this function.
0x02	The ending column of the first mapping region in this function.

- The length of the substring that contains the encoded coverage mapping data for the second function record is also 9. It’s structured like the mapping data for the first function record.
- The two trailing bytes are zeroes and are used to pad the coverage mapping data to give it the 8 byte alignment.

4.28.6 Encoding

The per-function coverage mapping data is encoded as a stream of bytes, with a simple structure. The structure consists of the encoding *types* like variable-length unsigned integers, that are used to encode *File ID Mapping*, *Counter Expressions* and the *Mapping Regions*.

The format of the structure follows:

```
[file id mapping, counterexpressions, mapping regions]
```

The translation unit filenames are encoded using the same encoding *types* as the per-function coverage mapping data, with the following structure:

```
[numFilenames : LEB128, filename0 : string, filename1 : string,
...]
```

Types

This section describes the basic types that are used by the encoding format and can appear after `:` in the `[foo : type]` description.

LEB128

LEB128 is an unsigned integer value that is encoded using DWARF's LEB128 encoding, optimizing for the case where values are small (1 byte for values less than 128).

Strings

```
[length : LEB128, characters...]
```

String values are encoded with a *LEB value* for the length of the string and a sequence of bytes for its characters.

File ID Mapping

```
[numIndices : LEB128, filenameIndex0 : LEB128, filenameIndex1 : LEB128, ...]
```

File id mapping in a function's coverage mapping stream contains the indices into the translation unit's filenames array.

Counter

```
[value : LEB128]
```

A *coverage mapping counter* is stored in a single *LEB value*. It is composed of two things — the *tag* which is stored in the lowest 2 bits, and the *counter data* which is stored in the remaining bits.

Tag:

The counter's tag encodes the counter's kind and, if the counter is an expression, the expression's kind. The possible tag values are:

- 0 - The counter is zero.
- 1 - The counter is a reference to the profile instrumentation counter.
- 2 - The counter is a subtraction expression.
- 3 - The counter is an addition expression.

Data:

The counter's data is interpreted in the following manner:

- When the counter is a reference to the profile instrumentation counter, then the counter's data is the id of the profile counter.
- When the counter is an expression, then the counter's data is the index into the array of counter expressions.

Counter Expressions

```
[numExpressions : LEB128, expr0LHS : LEB128, expr0RHS : LEB128, expr1LHS : LEB128, expr1RHS : LEB128, ...]
```

Counter expressions consist of two counters as they represent binary arithmetic operations. The expression's kind is determined from the *tag* of the counter that references this expression.

Mapping Regions

```
[numRegionArrays : LEB128, regionsForFile0, regionsForFile1, ...]
```

The mapping regions are stored in an array of sub-arrays where every region in a particular sub-array has the same file id.

The file id for a sub-array of regions is the index of that sub-array in the main array e.g. The first sub-array will have the file id of 0.

Sub-Array of Regions

```
[numRegions : LEB128, region0, region1, ...]
```

The mapping regions for a specific file id are stored in an array that is sorted in an ascending order by the region's starting location.

Mapping Region

```
[header, source range]
```

The mapping region record contains two sub-records — the *header*, which stores the counter and/or the region's kind, and the *source range* that contains the starting and ending location of this region.

Header

```
[counter]
```

or

```
[pseudo-counter]
```

The header encodes the region's counter and the region's kind.

The value of the counter's tag distinguishes between the counters and pseudo-counters — if the tag is zero, than this header contains a pseudo-counter, otherwise this header contains an ordinary counter.

Counter:

A mapping region whose header has a counter with a non-zero tag is a code region.

Pseudo-Counter:

```
[value : LEB128]
```

A pseudo-counter is stored in a single *LEB value*, just like the ordinary counter. It has the following interpretation:

- bits 0-1: tag, which is always 0.
- bit 2: expansionRegionTag. If this bit is set, then this mapping region is an expansion region.
- remaining bits: data. If this region is an expansion region, then the data contains the expanded file id of that region.

Otherwise, the data contains the region's kind. The possible region kind values are:

- 0 - This mapping region is a code region with a counter of zero.
- 2 - This mapping region is a skipped region.

Source Range

```
[deltaLineStart : LEB128, columnStart : LEB128, numLines : LEB128,  
columnEnd : LEB128]
```

The source range record contains the following fields:

- *deltaLineStart*: The difference between the starting line of the current mapping region and the starting line of the previous mapping region.

If the current mapping region is the first region in the current sub-array, then it stores the starting line of that region.

- *columnStart*: The starting column of the mapping region.
- *numLines*: The difference between the ending line and the starting line of the current mapping region.
- *columnEnd*: The ending column of the mapping region.

4.29 Garbage Collection Safepoints in LLVM

- *Status*
- *Overview*
 - *Base & Derived Pointers*
 - *GC Transitions*
- *Intrinsics*
 - *'llvm.experimental.gc.statepoint' Intrinsic*
 - *'llvm.experimental.gc.result' Intrinsic*
 - *'llvm.experimental.gc.relocate' Intrinsic*
- *Stack Map Format*
- *Safepoint Semantics & Verification*

- *Utility Passes for Safepoint Insertion*
 - *RewriteStatepointsForGC*
 - *PlaceSafepoints*
- *Supported Architectures*
- *Problem Areas and Active Work*
- *Bugs and Enhancements*

4.29.1 Status

This document describes a set of extensions to LLVM to support garbage collection. By now, these mechanisms are well proven with commercial java implementation with a fully relocating collector having shipped using them. There are a couple places where bugs might still linger; these are called out below.

They are still listed as “experimental” to indicate that no forward or backward compatibility guarantees are offered across versions. If your use case is such that you need some form of forward compatibility guarantee, please raise the issue on the `llvm-dev` mailing list.

LLVM still supports an alternate mechanism for conservative garbage collection support using the `gcroot` intrinsic. The `gcroot` mechanism is mostly of historical interest at this point with one exception - its implementation of shadow stacks has been used successfully by a number of language frontends and is still supported.

4.29.2 Overview

To collect dead objects, garbage collectors must be able to identify any references to objects contained within executing code, and, depending on the collector, potentially update them. The collector does not need this information at all points in code - that would make the problem much harder - but only at well-defined points in the execution known as ‘safepoints’. For most collectors, it is sufficient to track at least one copy of each unique pointer value. However, for a collector which wishes to relocate objects directly reachable from running code, a higher standard is required.

One additional challenge is that the compiler may compute intermediate results (“derived pointers”) which point outside of the allocation or even into the middle of another allocation. The eventual use of this intermediate value must yield an address within the bounds of the allocation, but such “exterior derived pointers” may be visible to the collector. Given this, a garbage collector can not safely rely on the runtime value of an address to indicate the object it is associated with. If the garbage collector wishes to move any object, the compiler must provide a mapping, for each pointer, to an indication of its allocation.

To simplify the interaction between a collector and the compiled code, most garbage collectors are organized in terms of three abstractions: load barriers, store barriers, and safepoints.

1. A load barrier is a bit of code executed immediately after the machine load instruction, but before any use of the value loaded. Depending on the collector, such a barrier may be needed for all loads, merely loads of a particular type (in the original source language), or none at all.
2. Analogously, a store barrier is a code fragment that runs immediately before the machine store instruction, but after the computation of the value stored. The most common use of a store barrier is to update a ‘card table’ in a generational garbage collector.
3. A safepoint is a location at which pointers visible to the compiled code (i.e. currently in registers or on the stack) are allowed to change. After the safepoint completes, the actual pointer value may differ, but the ‘object’ (as seen by the source language) pointed to will not.

Note that the term ‘safepoint’ is somewhat overloaded. It refers to both the location at which the machine state is parsable and the coordination protocol involved in bring application threads to a point at which the

collector can safely use that information. The term “statepoint” as used in this document refers exclusively to the former.

This document focuses on the last item - compiler support for safepoints in generated code. We will assume that an outside mechanism has decided where to place safepoints. From our perspective, all safepoints will be function calls. To support relocation of objects directly reachable from values in compiled code, the collector must be able to:

1. identify every copy of a pointer (including copies introduced by the compiler itself) at the safepoint,
2. identify which object each pointer relates to, and
3. potentially update each of those copies.

This document describes the mechanism by which an LLVM based compiler can provide this information to a language runtime/collector, and ensure that all pointers can be read and updated if desired.

At a high level, LLVM has been extended to support compiling to an abstract machine which extends the actual target with a non-integral pointer type suitable for representing a garbage collected reference to an object. In particular, such non-integral pointer type have no defined mapping to an integer representation. This semantic quirk allows the runtime to pick a integer mapping for each point in the program allowing relocations of objects without visible effects.

Warning: Non-Integral Pointer Types are a newly added concept in LLVM IR. It’s possible that we’ve missed disabling some of the optimizations which assume an integral value for pointers. If you find such a case, please file a bug or share a patch.

Warning: There is one currently known semantic hole in the definition of non-integral pointers which has not been addressed upstream. To work around this, you need to disable speculation of loads unless the memory type (non-integral pointer vs anything else) is known to unchanged. That is, it is not safe to speculate a load if doing causes a non-integral pointer value to be loaded as any other type or vice versa. In practice, this restriction is well isolated to `isSafeToSpeculate` in `ValueTracking.cpp`.

This high level abstract machine model is used for most of the LLVM optimizer. Before starting code generation, we switch representations to an explicit form. In theory, a frontend could directly generate this low level explicit form, but doing so is likely to inhibit optimization.

The heart of the explicit approach is to construct (or rewrite) the IR in a manner where the possible updates performed by the garbage collector are explicitly visible in the IR. Doing so requires that we:

1. create a new SSA value for each potentially relocated pointer, and ensure that no uses of the original (non-relocated) value is reachable after the safepoint,
2. specify the relocation in a way which is opaque to the compiler to ensure that the optimizer can not introduce new uses of an unrelocated value after a statepoint. This prevents the optimizer from performing unsound optimizations.
3. recording a mapping of live pointers (and the allocation they’re associated with) for each statepoint.

At the most abstract level, inserting a safepoint can be thought of as replacing a call instruction with a call to a multiple return value function which both calls the original target of the call, returns its result, and returns updated values for any live pointers to garbage collected objects.

Note that the task of identifying all live pointers to garbage collected values, transforming the IR to expose a pointer giving the base object for every such live pointer, and inserting all the intrinsics correctly is explicitly out of scope for this document. The recommended approach is to use the *utility passes* described below.

This abstract function call is concretely represented by a sequence of intrinsic calls known collectively as a “statepoint relocation sequence”.

Let’s consider a simple call in LLVM IR:

```
define i8 @test1(i8 @test1(i8 addrspace(1)* %obj)
    gc "statepoint-example" {
    call void ()* @foo()
    ret i8 addrspace(1)* %obj
}
```

Depending on our language we may need to allow a safepoint during the execution of `foo`. If so, we need to let the collector update local values in the current frame. If we don't, we'll be accessing a potential invalid reference once we eventually return from the call.

In this example, we need to relocate the SSA value `%obj`. Since we can't actually change the value in the SSA value `%obj`, we need to introduce a new SSA value `%obj.relocated` which represents the potentially changed value of `%obj` after the safepoint and update any following uses appropriately. The resulting relocation sequence is:

```
define i8 @test1(i8 @test1(i8 addrspace(1)* %obj)
    gc "statepoint-example" {
    %0 = call token (i64, i32, void ()*, i32, i32, ...) @llvm.experimental.gc.
    →statepoint.p0f_isVoidf(i64 0, i32 0, void ()* @foo, i32 0, i32 0, i32 0, i32 0, i8_
    →addrspace(1)* %obj)
    %obj.relocated = call coldcc i8 @llvm.experimental.gc.relocate.
    →pli8(token %0, i32 7, i32 7)
    ret i8 addrspace(1)* %obj.relocated
}
```

Ideally, this sequence would have been represented as a `M` argument, `N` return value function (where `M` is the number of values being relocated + the original call arguments and `N` is the original return value + each relocated value), but LLVM does not easily support such a representation.

Instead, the statepoint intrinsic marks the actual site of the safepoint or statepoint. The statepoint returns a token value (which exists only at compile time). To get back the original return value of the call, we use the `gc.result` intrinsic. To get the relocation of each pointer in turn, we use the `gc.relocate` intrinsic with the appropriate index. Note that both the `gc.relocate` and `gc.result` are tied to the statepoint. The combination forms a “statepoint relocation sequence” and represents the entirety of a parseable call or ‘statepoint’.

When lowered, this example would generate the following x86 assembly:

```
.globl    test1
.align    16, 0x90
pushq    %rax
callq    foo
.Ltmp1:
movq     (%rsp), %rax # This load is redundant (oops!)
popq     %rdx
retq
```

Each of the potentially relocated values has been spilled to the stack, and a record of that location has been recorded to the *Stack Map section*. If the garbage collector needs to update any of these pointers during the call, it knows exactly what to change.

The relevant parts of the StackMap section for our example are:

```
# This describes the call site
# Stack Maps: callsite 2882400000
    .quad 2882400000
    .long .Ltmp1-test1
    .short 0
# .. 8 entries skipped ..
# This entry describes the spill slot which is directly addressable
```

```
# off RSP with offset 0.  Given the value was spilled with a pushq,
# that makes sense.
# Stack Maps:  Loc 8: Direct RSP      [encoding: .byte 2, .byte 8, .short 7, .int 0]
               .byte 2
               .byte 8
               .short      7
               .long 0
```

This example was taken from the tests for the *RewriteStatepointsForGC* utility pass. As such, its full StackMap can be easily examined with the following command.

```
opt -rewrite-statepoints-for-gc test/Transforms/RewriteStatepointsForGC/basics.ll -S_
↳ | llc -debug-only=stackmaps
```

Base & Derived Pointers

A “base pointer” is one which points to the starting address of an allocation (object). A “derived pointer” is one which is offset from a base pointer by some amount. When relocating objects, a garbage collector needs to be able to relocate each derived pointer associated with an allocation to the same offset from the new address.

“Interior derived pointers” remain within the bounds of the allocation they’re associated with. As a result, the base object can be found at runtime provided the bounds of allocations are known to the runtime system.

“Exterior derived pointers” are outside the bounds of the associated object; they may even fall within *another* allocations address range. As a result, there is no way for a garbage collector to determine which allocation they are associated with at runtime and compiler support is needed.

The `gc.relocate` intrinsic supports an explicit operand for describing the allocation associated with a derived pointer. This operand is frequently referred to as the base operand, but does not strictly speaking have to be a base pointer, but it does need to lie within the bounds of the associated allocation. Some collectors may require that the operand be an actual base pointer rather than merely an internal derived pointer. Note that during lowering both the base and derived pointer operands are required to be live over the associated call safepoint even if the base is otherwise unused afterwards.

If we extend our previous example to include a pointless derived pointer, we get:

```
define i8 @test1(i8 @test1(i8 addrspc(1)* %obj)
  gc "statepoint-example" {
    %gep = getelementptr i8, i8 addrspc(1)* %obj, i64 20000
    %token = call token (i64, i32, void ()*, i32, i32, ...) @llvm.experimental.gc.
↳statepoint.pOf_isVoidf(i64 0, i32 0, void ()* @foo, i32 0, i32 0, i32 0, i32 0, i8_
↳addrspc(1)* %obj, i8 addrspc(1)* %gep)
    %obj.relocated = call i8 addrspc(1)* @llvm.experimental.gc.relocate.pli8(token
↳%token, i32 7, i32 7)
    %gep.relocated = call i8 addrspc(1)* @llvm.experimental.gc.relocate.pli8(token
↳%token, i32 7, i32 8)
    %p = getelementptr i8, i8 addrspc(1)* %gep, i64 -20000
    ret i8 addrspc(1)* %p
  }
```

Note that in this example `%p` and `%obj.relocate` are the same address and we could replace one with the other, potentially removing the derived pointer from the live set at the safepoint entirely.

GC Transitions

As a practical consideration, many garbage-collected systems allow code that is collector-aware (“managed code”) to call code that is not collector-aware (“unmanaged code”). It is common that such calls must also be safepoints, since it is desirable to allow the collector to run during the execution of unmanaged code. Furthermore, it is common that coordinating the transition from managed to unmanaged code requires extra code generation at the call site to inform the collector of the transition. In order to support these needs, a statepoint may be marked as a GC transition, and data that is necessary to perform the transition (if any) may be provided as additional arguments to the statepoint.

Note that although in many cases statepoints may be inferred to be GC transitions based on the function symbols involved (e.g. a call from a function with GC strategy “foo” to a function with GC strategy “bar”), indirect calls that are also GC transitions must also be supported. This requirement is the driving force behind the decision to require that GC transitions are explicitly marked.

Let’s revisit the sample given above, this time treating the call to `@foo` as a GC transition. Depending on our target, the transition code may need to access some extra state in order to inform the collector of the transition. Let’s assume a hypothetical GC—somewhat unimaginatively named “hypothetical-gc”—that requires that a TLS variable must be written to before and after a call to unmanaged code. The resulting relocation sequence is:

```
@flag = thread_local global i32 0, align 4

define i8 @test1(i8 @rspace(1) * %obj)
    gc "hypothetical-gc" {

        %0 = call token (i64, i32, void ()*, i32, i32, ...) * @llvm.experimental.gc.
        ↪ statepoint.p0f_isVoidf(i64 0, i32 0, void ()* @foo, i32 0, i32 1, i32* @Flag, i32 0,
        ↪ i8 @rspace(1) * %obj)
        %obj.relocated = call coldcc i8 @rspace(1) * @llvm.experimental.gc.relocate.
        ↪ pli8(token %0, i32 7, i32 7)
        ret i8 @rspace(1) * %obj.relocated
    }
```

During lowering, this will result in a instruction selection DAG that looks something like:

```
CALLSEQ_START
...
GC_TRANSITION_START (lowered i32 *@Flag), SRCVALUE i32* Flag
STATEPOINT
GC_TRANSITION_END (lowered i32 *@Flag), SRCVALUE i32 *Flag
...
CALLSEQ_END
```

In order to generate the necessary transition code, the backend for each target supported by “hypothetical-gc” must be modified to lower `GC_TRANSITION_START` and `GC_TRANSITION_END` nodes appropriately when the “hypothetical-gc” strategy is in use for a particular function. Assuming that such lowering has been added for X86, the generated assembly would be:

```
.globl      test1
.align     16, 0x90
pushq     %rax
movl     $1, %fs:Flag@TPOFF
callq    foo
movl     $0, %fs:Flag@TPOFF
.Ltmp1:
movq     (%rsp), %rax # This load is redundant (oops!)
popq     %rdx
retq
```

Note that the design as presented above is not fully implemented: in particular, strategy-specific lowering is not present, and all GC transitions are emitted as a single no-op before and after the call instruction. These no-ops are often removed by the backend during dead machine instruction elimination.

4.29.3 Intrinsics

‘`llvm.experimental.gc.statepoint`’ Intrinsic

Syntax:

```
declare token
@llvm.experimental.gc.statepoint(i64 <id>, i32 <num patch bytes>,
    func_type <target>,
    i64 <#call args>, i64 <flags>,
    ... (call parameters),
    i64 <# transition args>, ... (transition parameters),
    i64 <# deopt args>, ... (deopt parameters),
    ... (gc parameters))
```

Overview:

The statepoint intrinsic represents a call which is parse-able by the runtime.

Operands:

The ‘id’ operand is a constant integer that is reported as the ID field in the generated stackmap. LLVM does not interpret this parameter in any way and its meaning is up to the statepoint user to decide. Note that LLVM is free to duplicate code containing statepoint calls, and this may transform IR that had a unique ‘id’ per lexical call to statepoint to IR that does not.

If ‘num patch bytes’ is non-zero then the call instruction corresponding to the statepoint is not emitted and LLVM emits ‘num patch bytes’ bytes of nops in its place. LLVM will emit code to prepare the function arguments and retrieve the function return value in accordance to the calling convention; the former before the nop sequence and the latter after the nop sequence. It is expected that the user will patch over the ‘num patch bytes’ bytes of nops with a calling sequence specific to their runtime before executing the generated machine code. There are no guarantees with respect to the alignment of the nop sequence. Unlike *Stack maps and patch points in LLVM* statepoints do not have a concept of shadow bytes. Note that semantically the statepoint still represents a call or invoke to ‘target’, and the nop sequence after patching is expected to represent an operation equivalent to a call or invoke to ‘target’.

The ‘target’ operand is the function actually being called. The target can be specified as either a symbolic LLVM function, or as an arbitrary Value of appropriate function type. Note that the function type must match the signature of the callee and the types of the ‘call parameters’ arguments.

The ‘#call args’ operand is the number of arguments to the actual call. It must exactly match the number of arguments passed in the ‘call parameters’ variable length section.

The ‘flags’ operand is used to specify extra information about the statepoint. This is currently only used to mark certain statepoints as GC transitions. This operand is a 64-bit integer with the following layout, where bit 0 is the least significant bit:

Bit #	Usage
0	Set if the statepoint is a GC transition, cleared otherwise.
1-63	Reserved for future use; must be cleared.

The ‘call parameters’ arguments are simply the arguments which need to be passed to the call target. They will be lowered according to the specified calling convention and otherwise handled like a normal call instruction. The number of arguments must exactly match what is specified in ‘# call args’. The types must match the signature of ‘target’.

The ‘transition parameters’ arguments contain an arbitrary list of Values which need to be passed to GC transition code. They will be lowered and passed as operands to the appropriate GC_TRANSITION nodes in the selection DAG. It is assumed that these arguments must be available before and after (but not necessarily during) the execution of the callee. The ‘# transition args’ field indicates how many operands are to be interpreted as ‘transition parameters’.

The ‘deopt parameters’ arguments contain an arbitrary list of Values which is meaningful to the runtime. The runtime may read any of these values, but is assumed not to modify them. If the garbage collector might need to modify one of these values, it must also be listed in the ‘gc pointer’ argument list. The ‘# deopt args’ field indicates how many operands are to be interpreted as ‘deopt parameters’.

The ‘gc parameters’ arguments contain every pointer to a garbage collector object which potentially needs to be updated by the garbage collector. Note that the argument list must explicitly contain a base pointer for every derived pointer listed. The order of arguments is unimportant. Unlike the other variable length parameter sets, this list is not length prefixed.

Semantics:

A statepoint is assumed to read and write all memory. As a result, memory operations can not be reordered past a statepoint. It is illegal to mark a statepoint as being either ‘readonly’ or ‘readnone’.

Note that legal IR can not perform any memory operation on a ‘gc pointer’ argument of the statepoint in a location statically reachable from the statepoint. Instead, the explicitly relocated value (from a `gc.relocate`) must be used.

‘llvm.experimental.gc.result’ Intrinsic

Syntax:

```
declare type*
  @llvm.experimental.gc.result(token %statepoint_token)
```

Overview:

`gc.result` extracts the result of the original call instruction which was replaced by the `gc.statepoint`. The `gc.result` intrinsic is actually a family of three intrinsics due to an implementation limitation. Other than the type of the return value, the semantics are the same.

Operands:

The first and only argument is the `gc.statepoint` which starts the safepoint sequence of which this `gc.result` is a part. Despite the typing of this as a generic token, *only* the value defined by a `gc.statepoint` is legal here.

Semantics:

The `gc.result` represents the return value of the call target of the statepoint. The type of the `gc.result` must exactly match the type of the target. If the call target returns void, there will be no `gc.result`.

A `gc.result` is modeled as a ‘readnone’ pure function. It has no side effects since it is just a projection of the return value of the previous call represented by the `gc.statepoint`.

‘`llvm.experimental.gc.relocate`’ Intrinsic

Syntax:

```
declare <pointer type>
  @llvm.experimental.gc.relocate(token %statepoint_token,
                                i32 %base_offset,
                                i32 %pointer_offset)
```

Overview:

A `gc.relocate` returns the potentially relocated value of a pointer at the safepoint.

Operands:

The first argument is the `gc.statepoint` which starts the safepoint sequence of which this `gc.relocation` is a part. Despite the typing of this as a generic token, *only* the value defined by a `gc.statepoint` is legal here.

The second argument is an index into the statepoints list of arguments which specifies the allocation for the pointer being relocated. This index must land within the ‘gc parameter’ section of the statepoint’s argument list. The associated value must be within the object with which the pointer being relocated is associated. The optimizer is free to change *which* interior derived pointer is reported, provided that it does not replace an actual base pointer with another interior derived pointer. Collectors are allowed to rely on the base pointer operand remaining an actual base pointer if so constructed.

The third argument is an index into the statepoint’s list of arguments which specify the (potentially) derived pointer being relocated. It is legal for this index to be the same as the second argument if-and-only-if a base pointer is being relocated. This index must land within the ‘gc parameter’ section of the statepoint’s argument list.

Semantics:

The return value of `gc.relocate` is the potentially relocated value of the pointer specified by its arguments. It is unspecified how the value of the returned pointer relates to the argument to the `gc.statepoint` other than that a) it points to the same source language object with the same offset, and b) the ‘based-on’ relationship of the newly relocated pointers is a projection of the unrelocated pointers. In particular, the integer value of the pointer returned is unspecified.

A `gc.relocate` is modeled as a `readnone` pure function. It has no side effects since it is just a way to extract information about work done during the actual call modeled by the `gc.statepoint`.

4.29.4 Stack Map Format

Locations for each pointer value which may need read and/or updated by the runtime or collector are provided via the *Stack Map format* specified in the PatchPoint documentation.

Each statepoint generates the following Locations:

- Constant which describes the calling convention of the call target. This constant is a valid *calling convention identifier* for the version of LLVM used to generate the stackmap. No additional compatibility guarantees are made for this constant over what LLVM provides elsewhere w.r.t. these identifiers.
- Constant which describes the flags passed to the statepoint intrinsic
- Constant which describes number of following deopt *Locations* (not operands)
- Variable number of Locations, one for each deopt parameter listed in the IR statepoint (same number as described by previous Constant). At the moment, only deopt parameters with a bitwidth of 64 bits or less are supported. Values of a type larger than 64 bits can be specified and reported only if a) the value is constant at the call site, and b) the constant can be represented with less than 64 bits (assuming zero extension to the original bitwidth).
- Variable number of relocation records, each of which consists of exactly two Locations. Relocation records are described in detail below.

Each relocation record provides sufficient information for a collector to relocate one or more derived pointers. Each record consists of a pair of Locations. The second element in the record represents the pointer (or pointers) which need updated. The first element in the record provides a pointer to the base of the object with which the pointer(s) being relocated is associated. This information is required for handling generalized derived pointers since a pointer may be outside the bounds of the original allocation, but still needs to be relocated with the allocation. Additionally:

- It is guaranteed that the base pointer must also appear explicitly as a relocation pair if used after the statepoint.
- There may be fewer relocation records than gc parameters in the IR statepoint. Each *unique* pair will occur at least once; duplicates are possible.
- The Locations within each record may either be of pointer size or a multiple of pointer size. In the later case, the record must be interpreted as describing a sequence of pointers and their corresponding base pointers. If the Location is of size $N \times \text{sizeof}(\text{pointer})$, then there will be N records of one pointer each contained within the Location. Both Locations in a pair can be assumed to be of the same size.

Note that the Locations used in each section may describe the same physical location. e.g. A stack slot may appear as a deopt location, a gc base pointer, and a gc derived pointer.

The LiveOut section of the StkMapRecord will be empty for a statepoint record.

4.29.5 Safepoint Semantics & Verification

The fundamental correctness property for the compiled code's correctness w.r.t. the garbage collector is a dynamic one. It must be the case that there is no dynamic trace such that a operation involving a potentially relocated pointer is observably-after a safepoint which could relocate it. 'observably-after' in this usage means that an outside observer could observe this sequence of events in a way which precludes the operation being performed before the safepoint.

To understand why this 'observable-after' property is required, consider a null comparison performed on the original copy of a relocated pointer. Assuming that control flow follows the safepoint, there is no way to observe externally whether the null comparison is performed before or after the safepoint. (Remember, the original Value is unmodified by the safepoint.) The compiler is free to make either scheduling choice.

The actual correctness property implemented is slightly stronger than this. We require that there be no *static path* on which a potentially relocated pointer is 'observably-after' it may have been relocated. This is slightly stronger than

is strictly necessary (and thus may disallow some otherwise valid programs), but greatly simplifies reasoning about correctness of the compiled code.

By construction, this property will be upheld by the optimizer if correctly established in the source IR. This is a key invariant of the design.

The existing IR Verifier pass has been extended to check most of the local restrictions on the intrinsics mentioned in their respective documentation. The current implementation in LLVM does not check the key relocation invariant, but this is ongoing work on developing such a verifier. Please ask on `llvm-dev` if you're interested in experimenting with the current version.

4.29.6 Utility Passes for Safepoint Insertion

RewriteStatepointsForGC

The pass `RewriteStatepointsForGC` transforms a function's IR to lower from the abstract machine model described above to the explicit statepoint model of relocations. To do this, it replaces all calls or invokes of functions which might contain a safepoint poll with a `gc.statepoint` and associated full relocation sequence, including all required `gc.relocates`.

Note that by default, this pass only runs for the “statepoint-example” or “core-clr” gc strategies. You will need to add your custom strategy to this whitelist or use one of the predefined ones.

As an example, given this code:

```
define i8 @test1(i8 @addrspace(1)* %obj)
{
  gc "statepoint-example" {
    call void @foo()
    ret i8 @addrspace(1)* %obj
  }
}
```

The pass would produce this IR:

```
define i8 @test1(i8 @addrspace(1)* %obj)
{
  gc "statepoint-example" {
    %0 = call token (i64, i32, void ()*, i32, i32, ...) @llvm.experimental.gc.
    ↪statepoint.pOf_isVoidf(i64 2882400000, i32 0, void ()* @foo, i32 0, i32 0, i32 0, ↪
    ↪i32 5, i32 0, i32 -1, i32 0, i32 0, i32 0, i8 @addrspace(1)* %obj)
    %obj.relocated = call coldcc i8 @addrspace(1)* @llvm.experimental.gc.relocate.
    ↪pli8(token %0, i32 12, i32 12)
    ret i8 @addrspace(1)* %obj.relocated
  }
}
```

In the above examples, the `addrspace(1)` marker on the pointers is the mechanism that the `statepoint-example` GC strategy uses to distinguish references from non references. The pass assumes that all `addrspace(1)` pointers are non-integral pointer types. Address space 1 is not globally reserved for this purpose.

This pass can be used as a utility function by a language frontend that doesn't want to manually reason about liveness, base pointers, or relocation when constructing IR. As currently implemented, `RewriteStatepointsForGC` must be run after SSA construction (i.e. `mem2ref`).

`RewriteStatepointsForGC` will ensure that appropriate base pointers are listed for every relocation created. It will do so by duplicating code as needed to propagate the base pointer associated with each pointer being relocated to the appropriate safepoints. The implementation assumes that the following IR constructs produce base pointers: loads from the heap, addresses of global variables, function arguments, function return values. Constant pointers (such as null) are also assumed to be base pointers. In practice, this constraint can be relaxed to producing interior derived pointers provided the target collector can find the associated allocation from an arbitrary interior derived pointer.

By default `RewriteStatepointsForGC` passes in `0xABCDEF00` as the statepoint ID and 0 as the number of patchable bytes to the newly constructed `gc.statepoint`. These values can be configured on a per-callsite basis using the attributes `"statepoint-id"` and `"statepoint-num-patch-bytes"`. If a call site is marked with a `"statepoint-id"` function attribute and its value is a positive integer (represented as a string), then that value is used as the ID of the newly constructed `gc.statepoint`. If a call site is marked with a `"statepoint-num-patch-bytes"` function attribute and its value is a positive integer, then that value is used as the 'num patch bytes' parameter of the newly constructed `gc.statepoint`. The `"statepoint-id"` and `"statepoint-num-patch-bytes"` attributes are not propagated to the `gc.statepoint` call or invoke if they could be successfully parsed.

In practice, `RewriteStatepointsForGC` should be run much later in the pass pipeline, after most optimization is already done. This helps to improve the quality of the generated code when compiled with garbage collection support.

PlaceSafepoints

The pass `PlaceSafepoints` inserts safepoint polls sufficient to ensure running code checks for a safepoint request on a timely manner. This pass is expected to be run before `RewriteStatepointsForGC` and thus does not produce full relocation sequences.

As an example, given input IR of the following:

```
define void @test() gc "statepoint-example" {
  call void @foo()
  ret void
}

declare void @do_safepoint()
define void @gc.safepoint_poll() {
  call void @do_safepoint()
  ret void
}
```

This pass would produce the following IR:

```
define void @test() gc "statepoint-example" {
  call void @do_safepoint()
  call void @foo()
  ret void
}
```

In this case, we've added an (unconditional) entry safepoint poll. Note that despite appearances, the entry poll is not necessarily redundant. We'd have to know that `foo` and `test` were not mutually recursive for the poll to be redundant. In practice, you'd probably want your poll definition to contain a conditional branch of some form.

At the moment, `PlaceSafepoints` can insert safepoint polls at method entry and loop backedges locations. Extending this to work with return polls would be straight forward if desired.

`PlaceSafepoints` includes a number of optimizations to avoid placing safepoint polls at particular sites unless needed to ensure timely execution of a poll under normal conditions. `PlaceSafepoints` does not attempt to ensure timely execution of a poll under worst case conditions such as heavy system paging.

The implementation of a safepoint poll action is specified by looking up a function of the name `gc.safepoint_poll` in the containing Module. The body of this function is inserted at each poll site desired. While calls or invokes inside this method are transformed to a `gc.statepoints`, recursive poll insertion is not performed.

This pass is useful for any language frontend which only has to support garbage collection semantics at safepoints. If you need other abstract frame information at safepoints (e.g. for deoptimization or introspection), you can insert

safepoint polls in the frontend. If you have the later case, please ask on [llvm-dev](#) for suggestions. There's been a good amount of work done on making such a scheme work well in practice which is not yet documented here.

4.29.7 Supported Architectures

Support for statepoint generation requires some code for each backend. Today, only X86_64 is supported.

4.29.8 Problem Areas and Active Work

1. Support for languages which allow unmanaged pointers to garbage collected objects (i.e. pass a pointer to an object to a C routine) via pinning.
2. Support for garbage collected objects allocated on the stack. Specifically, allocas are always assumed to be in address space 0 and we need a cast/promotion operator to let rewriting identify them.
3. The current statepoint lowering is known to be somewhat poor. In the very long term, we'd like to integrate statepoints with the register allocator; in the near term this is unlikely to happen. We've found the quality of lowering to be relatively unimportant as hot-statepoints are almost always inliner bugs.
4. Concerns have been raised that the statepoint representation results in a large amount of IR being produced for some examples and that this contributes to higher than expected memory usage and compile times. There's no immediate plans to make changes due to this, but alternate models may be explored in the future.
5. Relocations along exceptional paths are currently broken in ToT. In particular, there is current no way to represent a rethrow on a path which also has relocations. See [this llvm-dev discussion](#) for more detail.

4.29.9 Bugs and Enhancements

Currently known bugs and enhancements under consideration can be tracked by performing a [bugzilla search](#) for [Statepoint] in the summary field. When filing new bugs, please use this tag so that interested parties see the newly filed bug. As with most LLVM features, design discussions take place on [llvm-dev](#), and patches should be sent to [llvm-commits](#) for review.

4.30 MergeFunctions pass, how it works

- *Introduction*
 - *Why would I want to read this document?*
 - *What should I know to be able to follow along with this document?*
 - *What I gain by reading this document?*
 - *Narrative structure*
- *Basics*
 - *How to do it?*
 - * *Possible solutions*
 - *Random-access*
 - *Logarithmical search*

- *Present state*
 - * *MergeFunctions, main fields and runOnModule*
 - *runOnModule*
 - *Comparison and logarithmical search*
- *Functions comparison*
 - *FunctionComparator::compare(void)*
 - *FunctionComparator::cmpType*
 - *cmpValues(const Value*, const Value*)*
 - * *What we associate in cmpValues?*
 - * *How to implement cmpValues?*
 - *cmpConstants*
 - *compare(const BasicBlock*, const BasicBlock*)*
 - *cmpGEP*
 - *cmpOperation*
 - *O(log(N))*
- *Merging process, mergeTwoFunctions*
 - *If “F” may be overridden*
 - * *HasGlobalAliases, removeUsers*
 - * *No global aliases, replaceDirectCallers*
 - *If “F” could not be overridden, fix it!*
- *That’s it.*

4.30.1 Introduction

Sometimes code contains equal functions, or functions that does exactly the same thing even though they are non-equal on the IR level (e.g.: multiplication on 2 and ‘shl 1’). It could happen due to several reasons: mainly, the usage of templates and automatic code generators. Though, sometimes user itself could write the same thing twice :-)

The main purpose of this pass is to recognize such functions and merge them.

Why would I want to read this document?

Document is the extension to pass comments and describes the pass logic. It describes algorithm that is used in order to compare functions, it also explains how we could combine equal functions correctly, keeping module valid.

Material is brought in top-down form, so reader could start learn pass from ideas and end up with low-level algorithm details, thus preparing him for reading the sources.

So main goal is do describe algorithm and logic here; the concept. This document is good for you, if you *don’t want* to read the source code, but want to understand pass algorithms. Author tried not to repeat the source-code and cover only common cases, and thus avoid cases when after minor code changes we need to update this document.

What should I know to be able to follow along with this document?

Reader should be familiar with common compile-engineering principles and LLVM code fundamentals. In this article we suppose reader is familiar with [Single Static Assignment](#) concepts. Understanding of [IR structure](#) is also important.

We will use such terms as “[module](#)”, “[function](#)”, “[basic block](#)”, “[user](#)”, “[value](#)”, “[instruction](#)”.

As a good start point, Kaleidoscope tutorial could be used:

[LLVM Tutorial: Table of Contents](#)

Especially it’s important to understand chapter 3 of tutorial:

[Kaleidoscope: Code generation to LLVM IR](#)

Reader also should know how passes work in LLVM, they could use next article as a reference and start point here:

[Writing an LLVM Pass](#)

What else? Well perhaps reader also should have some experience in LLVM pass debugging and bug-fixing.

What I gain by reading this document?

Main purpose is to provide reader with comfortable form of algorithms description, namely the human reading text. Since it could be hard to understand algorithm straight from the source code: pass uses some principles that have to be explained first.

Author wishes to everybody to avoid case, when you read code from top to bottom again and again, and yet you don’t understand why we implemented it that way.

We hope that after this article reader could easily debug and improve MergeFunctions pass and thus help LLVM project.

Narrative structure

Article consists of three parts. First part explains pass functionality on the top-level. Second part describes the comparison procedure itself. The third part describes the merging process.

In every part author also tried to put the contents into the top-down form. First, the top-level methods will be described, while the terminal ones will be at the end, in the tail of each part. If reader will see the reference to the method that wasn’t described yet, they will find its description a bit below.

4.30.2 Basics

How to do it?

Do we need to merge functions? Obvious thing is: yes that’s a quite possible case, since usually we *do* have duplicates. And it would be good to get rid of them. But how to detect such a duplicates? The idea is next: we split functions onto small bricks (parts), then we compare “bricks” amount, and if it equal, compare “bricks” themselves, and then do our conclusions about functions themselves.

What the difference it could be? For example, on machine with 64-bit pointers (let’s assume we have only one address space), one function stores 64-bit integer, while another one stores a pointer. So if the target is a machine mentioned above, and if functions are identical, except the parameter type (we could consider it as a part of function type), then we can treat `uint64_t` and “`void*`” as equal.

It was just an example; possible details are described a bit below.

As another example reader may imagine two more functions. First function performs multiplication on 2, while the second one performs arithmetic right shift on 1.

Possible solutions

Let's briefly consider possible options about how and what we have to implement in order to create full-featured functions merging, and also what it would mean for us.

Equal functions detection, obviously supposes "detector" method to be implemented, latter should answer the question "whether functions are equal". This "detector" method consists of tiny "sub-detectors", each of them answers exactly the same question, but for function parts.

As the second step, we should merge equal functions. So it should be a "merger" method. "Merger" accepts two functions *F1* and *F2*, and produces *F1F2* function, the result of merging.

Having such a routines in our hands, we can process whole module, and merge all equal functions.

In this case, we have to compare every function with every another function. As reader could notice, this way seems to be quite expensive. Of course we could introduce hashing and other helpers, but it is still just an optimization, and thus the level of $O(N*N)$ complexity.

Can we reach another level? Could we introduce logarithmical search, or random access lookup? The answer is: "yes".

Random-access

How it could be done? Just convert each function to number, and gather all of them in special hash-table. Functions with equal hash are equal. Good hashing means, that every function part must be taken into account. That means we have to convert every function part into some number, and then add it into hash. Lookup-up time would be small, but such approach adds some delay due to hashing routine.

Logarithmical search

We could introduce total ordering among the functions set, once we had it we could then implement a logarithmical search. Lookup time still depends on N , but adds a little of delay ($\log(N)$).

Present state

Both of approaches (random-access and logarithmical) has been implemented and tested. And both of them gave a very good improvement. And what was most surprising, logarithmical search was faster; sometimes up to 15%. Hashing needs some extra CPU time, and it is the main reason why it works slower; in most of cases total "hashing" time was greater than total "logarithmical-search" time.

So, preference has been granted to the "logarithmical search".

Though in the case of need, *logarithmical-search* (read "total-ordering") could be used as a milestone on our way to the *random-access* implementation.

Every comparison is based either on the numbers or on flags comparison. In *random-access* approach we could use the same comparison algorithm. During comparison we exit once we find the difference, but here we might have to scan whole function body every time (note, it could be slower). Like in "total-ordering", we will track every numbers and flags, but instead of comparison, we should get numbers sequence and then create the hash number. So, once again, *total-ordering* could be considered as a milestone for even faster (in theory) random-access approach.

MergeFunctions, main fields and runOnModule

There are two most important fields in class:

`FnTree` – the set of all unique functions. It keeps items that couldn't be merged with each other. It is defined as:

```
std::set<FunctionNode> FnTree;
```

Here `FunctionNode` is a wrapper for `llvm::Function` class, with implemented “<” operator among the functions set (below we explain how it works exactly; this is a key point in fast functions comparison).

`Deferred` – merging process can affect bodies of functions that are in `FnTree` already. Obviously such functions should be rechecked again. In this case we remove them from `FnTree`, and mark them as to be rescanned, namely put them into `Deferred` list.

runOnModule

The algorithm is pretty simple:

1. Put all module's functions into the *worklist*.
2. Scan *worklist*'s functions twice: first enumerate only strong functions and then only weak ones:
 - 2.1. Loop body: take function from *worklist* (call it *FCur*) and try to insert it into *FnTree*: check whether *FCur* is equal to one of functions in *FnTree*. If there is equal function in *FnTree* (call it *FExists*): merge function *FCur* with *FExists*. Otherwise add function from *worklist* to *FnTree*.
3. Once *worklist* scanning and merging operations is complete, check *Deferred* list. If it is not empty: refill *worklist* contents with *Deferred* list and do step 2 again, if *Deferred* is empty, then exit from method.

Comparison and logarithmical search

Let's recall our task: for every function *F* from module *M*, we have to find equal functions *F'* in shortest time, and merge them into the single function.

Defining total ordering among the functions set allows to organize functions into the binary tree. The lookup procedure complexity would be estimated as $O(\log(N))$ in this case. But how to define *total-ordering*?

We have to introduce a single rule applicable to every pair of functions, and following this rule then evaluate which of them is greater. What kind of rule it could be? Let's declare it as “compare” method, that returns one of 3 possible values:

- 1, left is *less* than right,
- 0, left and right are *equal*,
- 1, left is *greater* than right.

Of course it means, that we have to maintain *strict and non-strict order relation properties*:

- reflexivity ($a \leq a$, $a == a$, $a \geq a$),
- antisymmetry (if $a \leq b$ and $b \leq a$ then $a == b$),
- transitivity ($a \leq b$ and $b \leq c$, then $a \leq c$)
- asymmetry (if $a < b$, then $a > b$ or $a == b$).

As it was mentioned before, comparison routine consists of “sub-comparison-routines”, each of them also consists “sub-comparison-routines”, and so on, finally it ends up with a primitives comparison.

Below, we will use the next operations:

1. `cmpNumbers(number1, number2)` is method that returns -1 if left is less than right; 0, if left and right are equal; and 1 otherwise.
2. `cmpFlags(flag1, flag2)` is hypothetical method that compares two flags. The logic is the same as in `cmpNumbers`, where `true` is 1, and `false` is 0.

The rest of article is based on *MergeFunctions.cpp* source code (`<llvm_dir>/lib/Transforms/IPO/MergeFunctions.cpp`). We would like to ask reader to keep this file open nearby, so we could use it as a reference for further explanations.

Now we're ready to proceed to the next chapter and see how it works.

4.30.3 Functions comparison

At first, let's define how exactly we compare complex objects.

Complex objects comparison (function, basic-block, etc) is mostly based on its sub-objects comparison results. So it is similar to the next "tree" objects comparison:

1. For two trees *T1* and *T2* we perform *depth-first-traversal* and have two sequences as a product: "*T1Items*" and "*T2Items*".
2. Then compare chains "*T1Items*" and "*T2Items*" in most-significant-item-first order. Result of items comparison would be the result of *T1* and *T2* comparison itself.

FunctionComparator::compare(void)

Brief look at the source code tells us, that comparison starts in "`int FunctionComparator::compare(void)`" method.

1. First parts to be compared are function's attributes and some properties that outsidess "attributes" term, but still could make function different without changing its body. This part of comparison is usually done within simple *cmpNumbers* or *cmpFlags* operations (e.g. `cmpFlags(F1->hasGC(), F2->hasGC())`). Below is full list of function's properties to be compared on this stage:

- *Attributes* (those are returned by `Function::getAttributes()` method).
- *GC*, for equivalence, *RHS* and *LHS* should be both either without *GC* or with the same one.
- *Section*, just like a *GC*: *RHS* and *LHS* should be defined in the same section.
- *Variable arguments*. *LHS* and *RHS* should be both either with or without *var-args*.
- *Calling convention* should be the same.

2. Function type. Checked by `FunctionComparator::cmpType(Type*, Type*)` method. It checks return type and parameters type; the method itself will be described later.

3. Associate function formal parameters with each other. Then comparing function bodies, if we see the usage of *LHS*'s *i*-th argument in *LHS*'s body, then, we want to see usage of *RHS*'s *i*-th argument at the same place in *RHS*'s body, otherwise functions are different. On this stage we grant the preference to those we met later in function body (value we met first would be *less*). This is done by "`FunctionComparator::cmpValues(const Value*, const Value*)`" method (will be described a bit later).

4. Function body comparison. As it written in method comments:

"We do a CFG-ordered walk since the actual ordering of the blocks in the linked list is immaterial. Our walk starts at the entry block for both functions, then takes each block from each terminator in order. As an artifact, this also means that unreachable blocks are ignored."

So, using this walk we get BBs from *left* and *right* in the same order, and compare them by "`FunctionComparator::compare(const BasicBlock*, const BasicBlock*)`" method.

We also associate BBs with each other, like we did it with function formal arguments (see `cmpValues` method below).

FunctionComparator::cmpType

Consider how types comparison works.

1. Coerce pointer to integer. If left type is a pointer, try to coerce it to the integer type. It could be done if its address space is 0, or if address spaces are ignored at all. Do the same thing for the right type.
2. If left and right types are equal, return 0. Otherwise we need to give preference to one of them. So proceed to the next step.
3. If types are of different kind (different type IDs). Return result of type IDs comparison, treating them as a numbers (use `cmpNumbers` operation).
4. If types are vectors or integers, return result of their pointers comparison, comparing them as numbers.
5. Check whether type ID belongs to the next group (call it equivalent-group):
 - Void
 - Float
 - Double
 - X86_FP80
 - FP128
 - PPC_FP128
 - Label
 - Metadata.

If ID belongs to group above, return 0. Since it's enough to see that types has the same `TypeID`. No additional information is required.

6. Left and right are pointers. Return result of address space comparison (numbers comparison).
7. Complex types (structures, arrays, etc.). Follow complex objects comparison technique (see the very first paragraph of this chapter). Both *left* and *right* are to be expanded and their element types will be checked the same way. If we get -1 or 1 on some stage, return it. Otherwise return 0.
8. Steps 1-6 describe all the possible cases, if we passed steps 1-6 and didn't get any conclusions, then invoke `llvm_unreachable`, since it's quite unexpected case.

cmpValues(const Value*, const Value*)

Method that compares local values.

This method gives us an answer on a very curious question: whether we could treat local values as equal, and which value is greater otherwise. It's better to start from example:

Consider situation when we're looking at the same place in left function "*FL*" and in right function "*FR*". And every part of *left* place is equal to the corresponding part of *right* place, and (!) both parts use *Value* instances, for example:

```
instr0 i32 %LV    ; left side, function FL
instr0 i32 %RV    ; right side, function FR
```

So, now our conclusion depends on *Value* instances comparison.

Main purpose of this method is to determine relation between such values.

What we expect from equal functions? At the same place, in functions “*FL*” and “*FR*” we expect to see *equal* values, or values *defined* at the same place in “*FL*” and “*FR*”.

Consider small example here:

```
define void %f(i32 %pf0, i32 %pf1) {
    instr0 i32 %pf0 instr1 i32 %pf1 instr2 i32 123
}
```

```
define void %g(i32 %pg0, i32 %pg1) {
    instr0 i32 %pg0 instr1 i32 %pg0 instr2 i32 123
}
```

In this example, *pf0* is associated with *pg0*, *pf1* is associated with *pg1*, and we also declare that *pf0* < *pf1*, and thus *pg0* < *pf1*.

Instructions with opcode “*instr0*” would be *equal*, since their types and opcodes are equal, and values are *associated*.

Instruction with opcode “*instr1*” from *f* is *greater* than instruction with opcode “*instr1*” from *g*; here we have equal types and opcodes, but “*pf1* is greater than “*pg0*”.

And instructions with opcode “*instr2*” are equal, because their opcodes and types are equal, and the same constant is used as a value.

What we associate in cmpValues?

- Function arguments. *i*-th argument from left function associated with *i*-th argument from right function.
- BasicBlock instances. In basic-block enumeration loop we associate *i*-th BasicBlock from the left function with *i*-th BasicBlock from the right function.
- Instructions.
- Instruction operands. Note, we can meet *Value* here we have never seen before. In this case it is not a function argument, nor *BasicBlock*, nor *Instruction*. It is global value. It is constant, since its the only supposed global here. Method also compares:
- Constants that are of the same type.
- If right constant could be losslessly bit-casted to the left one, then we also compare them.

How to implement cmpValues?

Association is a case of equality for us. We just treat such values as equal. But, in general, we need to implement antisymmetric relation. As it was mentioned above, to understand what is *less*, we can use order in which we meet values. If both of values has the same order in function (met at the same time), then treat values as *associated*. Otherwise – it depends on who was first.

Every time we run top-level compare method, we initialize two identical maps (one for the left side, another one for the right side):

```
map<Value, int> sn_mapL, sn_mapR;
```

The key of the map is the *Value* itself, the *value* – is its order (call it *serial number*).

To add value *V* we need to perform the next procedure:

```
sn_map.insert(std::make_pair(V, sn_map.size()));
```

For the first *Value*, map will return 0, for second *Value* map will return 1, and so on.

Then we can check whether left and right values met at the same time with simple comparison:

```
cmpNumbers(sn_mapL[Left], sn_mapR[Right]);
```

Of course, we can combine insertion and comparison:

```
std::pair<iterator, bool>
LeftRes = sn_mapL.insert(std::make_pair(Left, sn_mapL.size())), RightRes
= sn_mapR.insert(std::make_pair(Right, sn_mapR.size()));
return cmpNumbers(LeftRes.first->second, RightRes.first->second);
```

Let's look, how whole method could be implemented.

1. we have to start from the bad news. Consider function self and cross-referencing cases:

```
// self-reference unsigned fact0(unsigned n) { return n > 1 ? n
* fact0(n-1) : 1; } unsigned fact1(unsigned n) { return n > 1 ? n *
fact1(n-1) : 1; }

// cross-reference unsigned ping(unsigned n) { return n!= 0 ? pong(n-1) : 0;
} unsigned pong(unsigned n) { return n!= 0 ? ping(n-1) : 0; }
```

This comparison has been implemented in initial *MergeFunctions* pass version. But, unfortunately, it is not transitive. And this is the only case we can't convert to less-equal-greater comparison. It is a seldom case, 4-5 functions of 10000 (checked on test-suite), and, we hope, reader would forgive us for such a sacrifice in order to get the $O(\log(N))$ pass time.

2. If left/right *Value* is a constant, we have to compare them. Return 0 if it is the same constant, or use `cmpConstants` method otherwise.
3. If left/right is *InlineAsm* instance. Return result of *Value* pointers comparison.
4. Explicit association of *L* (left value) and *R* (right value). We need to find out whether values met at the same time, and thus are *associated*. Or we need to put the rule: when we treat $L < R$. Now it is easy: just return result of numbers comparison:

```
std::pair<iterator, bool>
LeftRes = sn_mapL.insert(std::make_pair(Left, sn_mapL.size())),
RightRes = sn_mapR.insert(std::make_pair(Right, sn_mapR.size()));
if (LeftRes.first->second == RightRes.first->second) return 0;
if (LeftRes.first->second < RightRes.first->second) return -1;
return 1;
```

Now when *cmpValues* returns 0, we can proceed comparison procedure. Otherwise, if we get (-1 or 1), we need to pass this result to the top level, and finish comparison procedure.

cmpConstants

Performs constants comparison as follows:

1. Compare constant types using `cmpType` method. If result is -1 or 1, goto step 2, otherwise proceed to step 3.
2. If types are different, we still can check whether constants could be losslessly bitcasted to each other. The further explanation is modification of `canLosslesslyBitCastTo` method.

- 2.1 Check whether constants are of the first class types (`isFirstClassType` check):
 - 2.1.1. If both constants are *not* of the first class type: return result of `cmpType`.
 - 2.1.2. Otherwise, if left type is not of the first class, return -1. If right type is not of the first class, return 1.
 - 2.1.3. If both types are of the first class type, proceed to the next step (2.1.3.1).
 - 2.1.3.1. If types are vectors, compare their bitwidth using the `cmpNumbers`. If result is not 0, return it.
 - 2.1.3.2. Different types, but not a vectors:
 - if both of them are pointers, good for us, we can proceed to step 3.
 - if one of types is pointer, return result of `isPointer` flags comparison (`cmpFlags` operation).
 - otherwise we have no methods to prove bitcastability, and thus return result of types comparison (-1 or 1).

Steps below are for the case when types are equal, or case when constants are bitcastable:

3. One of constants is a “*null*” value. Return the result of `cmpFlags(L->isNullValue, R->isNullValue)` comparison.

4. Compare value IDs, and return result if it is not 0:

```
if (int Res = cmpNumbers(L->getValueID(), R->getValueID()))
    return Res;
```

5. Compare the contents of constants. The comparison depends on kind of constants, but on this stage it is just a lexicographical comparison. Just see how it was described in the beginning of “*Functions comparison*” paragraph. Mathematically it is equal to the next case: we encode left constant and right constant (with similar way *bitcode-writer* does). Then compare left code sequence and right code sequence.

compare(const BasicBlock*, const BasicBlock*)

Compares two *BasicBlock* instances.

It enumerates instructions from left *BB* and right *BB*.

1. It assigns serial numbers to the left and right instructions, using `cmpValues` method.
2. If one of left or right is *GEP* (`GetElementPtr`), then treat *GEP* as greater than other instructions, if both instructions are *GEPs* use `cmpGEP` method for comparison. If result is -1 or 1, pass it to the top-level comparison (return it).
 - 3.1. Compare operations. Call `cmpOperation` method. If result is -1 or 1, return it.
 - 3.2. Compare number of operands, if result is -1 or 1, return it.
 - 3.3. Compare operands themselves, use `cmpValues` method. Return result if it is -1 or 1.
 - 3.4. Compare type of operands, using `cmpType` method. Return result if it is -1 or 1.
 - 3.5. Proceed to the next instruction.
4. We can finish instruction enumeration in 3 cases:
 - 4.1. We reached the end of both left and right basic-blocks. We didn’t exit on steps 1-3, so contents is equal, return 0.
 - 4.2. We have reached the end of the left basic-block. Return -1.
 - 4.3. Return 1 (the end of the right basic block).

cmpGEP

Compares two GEPs (`getelementptr` instructions).

It differs from regular operations comparison with the only thing: possibility to use `accumulateConstantOffset` method.

So, if we get constant offset for both left and right *GEPs*, then compare it as numbers, and return comparison result.

Otherwise treat it like a regular operation (see previous paragraph).

cmpOperation

Compares instruction opcodes and some important operation properties.

1. Compare opcodes, if it differs return the result.
2. Compare number of operands. If it differs – return the result.
3. Compare operation types, use *cmpType*. All the same – if types are different, return result.
4. Compare *subclassOptionalData*, get it with `getRawSubclassOptionalData` method, and compare it like a numbers.
5. Compare operand types.
6. For some particular instructions check equivalence (relation in our case) of some significant attributes. For example we have to compare alignment for `load` instructions.

O(log(N))

Methods described above implement order relationship. And latter, could be used for nodes comparison in a binary tree. So we can organize functions set into the binary tree and reduce the cost of lookup procedure from $O(N*N)$ to $O(\log(N))$.

4.30.4 Merging process, mergeTwoFunctions

Once *MergeFunctions* detected that current function (*G*) is equal to one that were analyzed before (function *F*) it calls `mergeTwoFunctions(Function*, Function*)`.

Operation affects `FnTree` contents with next way: *F* will stay in `FnTree`. *G* being equal to *F* will not be added to `FnTree`. Calls of *G* would be replaced with something else. It changes bodies of callers. So, functions that calls *G* would be put into `Deferred` set and removed from `FnTree`, and analyzed again.

The approach is next:

1. Most wished case: when we can use alias and both of *F* and *G* are weak. We make both of them with aliases to the third strong function *H*. Actually *H* is *F*. See below how it's made (but it's better to look straight into the source code). Well, this is a case when we can just replace *G* with *F* everywhere, we use `replaceAllUsesWith` operation here (*RAUW*).
2. *F* could not be overridden, while *G* could. It would be good to do the next: after merging the places where overridable function were used, still use overridable stub. So try to make *G* alias to *F*, or create overridable tail call wrapper around *F* and replace *G* with that call.
3. Neither *F* nor *G* could be overridden. We can't use *RAUW*. We can just change the callers: call *F* instead of *G*. That's what `replaceDirectCallers` does.

Below is detailed body description.

If “F” may be overridden

As follows from `maybeOverridden` comments: “whether the definition of this global may be replaced by something non-equivalent at link time”. If so, that’s ok: we can use alias to *F* instead of *G* or change call instructions itself.

HasGlobalAliases, removeUsers

First consider the case when we have global aliases of one function name to another. Our purpose is make both of them with aliases to the third strong function. Though if we keep *F* alive and without major changes we can leave it in `FnTree`. Try to combine these two goals.

Do stub replacement of *F* itself with an alias to *F*.

1. Create stub function *H*, with the same name and attributes like function *F*. It takes maximum alignment of *F* and *G*.
2. Replace all uses of function *F* with uses of function *H*. It is the two steps procedure instead. First of all, we must take into account, all functions from whom *F* is called would be changed: since we change the call argument (from *F* to *H*). If so we must to review these caller functions again after this procedure. We remove callers from `FnTree`, method with name `removeUsers(F)` does that (don’t confuse with `replaceAllUsesWith`):
 - 2.1. Inside `removeUsers(Value* V)` we go through the all values that use value *V* (or *F* in our context). If value is instruction, we go to function that holds this instruction and mark it as to-be-analyzed-again (put to `Deferred` set), we also remove caller from `FnTree`.
 - 2.2. Now we can do the replacement: `call F->replaceAllUsesWith(H)`.
3. *H* (that now “officially” plays *F*’s role) is replaced with alias to *F*. Do the same with *G*: replace it with alias to *F*. So finally everywhere *F* was used, we use *H* and it is alias to *F*, and everywhere *G* was used we also have alias to *F*.
4. Set *F* linkage to private. Make it strong :-)

No global aliases, replaceDirectCallers

If global aliases are not supported. We call `replaceDirectCallers` then. Just go through all calls of *G* and replace it with calls of *F*. If you look into method you will see that it scans all uses of *G* too, and if use is callee (if user is call instruction and *G* is used as what to be called), we replace it with use of *F*.

If “F” could not be overridden, fix it!

We call `writeThunkOrAlias(Function *F, Function *G)`. Here we try to replace *G* with alias to *F* first. Next conditions are essential:

- target should support global aliases,
- the address itself of *G* should be not significant, not named and not referenced anywhere,
- function should come with external, local or weak linkage.

Otherwise we write thunk: some wrapper that has *G*’s interface and calls *F*, so *G* could be replaced with this wrapper.

writeAlias

As follows from *llvm* reference:

“Aliases act as *second name* for the aliasee value”. So we just want to create second name for *F* and use it instead of *G*:

1. create global alias itself (*GA*),
2. adjust alignment of *F* so it must be maximum of current and *G*'s alignment;
3. replace uses of *G*:
 - 3.1. first mark all callers of *G* as to-be-analyzed-again, using `removeUsers` method (see chapter above),
 - 3.2. call `G->replaceAllUsesWith(GA)`.
4. Get rid of *G*.

writeThunk

As it written in method comments:

“Replace *G* with a simple tail call to `bitcast(F)`. Also replace direct uses of *G* with `bitcast(F)`. Deletes *G*.”

In general it does the same as usual when we want to replace callee, except the first point:

1. We generate tail call wrapper around *F*, but with interface that allows use it instead of *G*.
2. “As-usual”: `removeUsers` and `replaceAllUsesWith` then.
3. Get rid of *G*.

4.30.5 That's it.

We have described how to detect equal functions, and how to merge them, and in first chapter we have described how it works all-together. Author hopes, reader have some picture from now, and it helps him improve and debug this pass.

Reader is welcomed to send us any questions and proposals ;-)

4.31 Type Metadata

Type metadata is a mechanism that allows IR modules to co-operatively build pointer sets corresponding to addresses within a given set of globals. LLVM's [control flow integrity](#) implementation uses this metadata to efficiently check (at each call site) that a given address corresponds to either a valid vtable or function pointer for a given class or function type, and its whole-program devirtualization pass uses the metadata to identify potential callees for a given virtual call.

To use the mechanism, a client creates metadata nodes with two elements:

1. a byte offset into the global (generally zero for functions)
2. a metadata object representing an identifier for the type

These metadata nodes are associated with globals by using global object metadata attachments with the `!type` metadata kind.

Each type identifier must exclusively identify either global variables or functions.

Limitation

The current implementation only supports attaching metadata to functions on the x86-32 and x86-64 architectures.

An intrinsic, `llvm.type.test`, is used to test whether a given pointer is associated with a type identifier.

4.31.1 Representing Type Information using Type Metadata

This section describes how Clang represents C++ type information associated with virtual tables using type metadata. Consider the following inheritance hierarchy:

```
struct A {
    virtual void f();
};

struct B : A {
    virtual void f();
    virtual void g();
};

struct C {
    virtual void h();
};

struct D : A, C {
    virtual void f();
    virtual void h();
};
```

The virtual table objects for A, B, C and D look like this (under the Itanium ABI):

Table 4.60: Virtual Table Layout for A, B, C, D

Class	0	1	2	3	4	5	6
A	A::offset-to-top	&A::rtti	&A::f				
B	B::offset-to-top	&B::rtti	&B::f	&B::g			
C	C::offset-to-top	&C::rtti	&C::h				
D	D::offset-to-top	&D::rtti	&D::f	&D::h	D::offset-to-top	&D::rtti	thunk for &D::h

When an object of type A is constructed, the address of `&A::f` in A's virtual table object is stored in the object's vtable pointer. In ABI parlance this address is known as an [address point](#). Similarly, when an object of type B is constructed, the address of `&B::f` is stored in the vtable pointer. In this way, the vtable in B's virtual table object is compatible with A's vtable.

D is a little more complicated, due to the use of multiple inheritance. Its virtual table object contains two vttables, one compatible with A's vtable and the other compatible with C's vtable. Objects of type D contain two virtual pointers, one belonging to the A subobject and containing the address of the vtable compatible with A's vtable, and the other belonging to the C subobject and containing the address of the vtable compatible with C's vtable.

The full set of compatibility information for the above class hierarchy is shown below. The following table shows the name of a class, the offset of an address point within that class's vtable and the name of one of the classes with which that address point is compatible.

Table 4.61: Type Offsets for A, B, C, D

VTable for	Offset	Compatible Class
A	16	A
B	16	A
		B
C	16	C
D	16	A
		D
	48	C

The next step is to encode this compatibility information into the IR. The way this is done is to create type metadata named after each of the compatible classes, with which we associate each of the compatible address points in each vtable. For example, these type metadata entries encode the compatibility information for the above hierarchy:

```
@_ZTV1A = constant [...], !type !0
@_ZTV1B = constant [...], !type !0, !type !1
@_ZTV1C = constant [...], !type !2
@_ZTV1D = constant [...], !type !0, !type !3, !type !4

!0 = !{i64 16, !"_ZTS1A"}
!1 = !{i64 16, !"_ZTS1B"}
!2 = !{i64 16, !"_ZTS1C"}
!3 = !{i64 16, !"_ZTS1D"}
!4 = !{i64 48, !"_ZTS1C"}
```

With this type metadata, we can now use the `llvm.type.test` intrinsic to test whether a given pointer is compatible with a type identifier. Working backwards, if `llvm.type.test` returns true for a particular pointer, we can also statically determine the identities of the virtual functions that a particular virtual call may call. For example, if a program assumes a pointer to be a member of `! "_ZTS1A"`, we know that the address can be only be one of `_ZTV1A+16`, `_ZTV1B+16` or `_ZTV1D+16` (i.e. the address points of the vtables of A, B and D respectively). If we then load an address from that pointer, we know that the address can only be one of `&A::f`, `&B::f` or `&D::f`.

4.31.2 Testing Addresses For Type Membership

If a program tests an address using `llvm.type.test`, this will cause a link-time optimization pass, `LowerTypeTests`, to replace calls to this intrinsic with efficient code to perform type member tests. At a high level, the pass will lay out referenced globals in a consecutive memory region in the object file, construct bit vectors that map onto that memory region, and generate code at each of the `llvm.type.test` call sites to test pointers against those bit vectors. Because of the layout manipulation, the globals' definitions must be available at LTO time. For more information, see the [control flow integrity design document](#).

A type identifier that identifies functions is transformed into a jump table, which is a block of code consisting of one branch instruction for each of the functions associated with the type identifier that branches to the target function. The pass will redirect any taken function addresses to the corresponding jump table entry. In the object file's symbol table, the jump table entries take the identities of the original functions, so that addresses taken outside the module will pass any verification done inside the module.

Jump tables may call external functions, so their definitions need not be available at LTO time. Note that if an externally defined function is associated with a type identifier, there is no guarantee that its identity within the module will be the same as its identity outside of the module, as the former will be the jump table entry if a jump table is necessary.

The `GlobalLayoutBuilder` class is responsible for laying out the globals efficiently to minimize the sizes of the underlying bitsets.

Example

```
target datalayout = "e-p:32:32"

@a = internal global i32 0, !type !0
@b = internal global i32 0, !type !0, !type !1
@c = internal global i32 0, !type !1
@d = internal global [2 x i32] [i32 0, i32 0], !type !2

define void @e() !type !3 {
    ret void
}
```

```

define void @f() {
    ret void
}

declare void @g() !type !3

!0 = !{i32 0, !"typeid1"}
!1 = !{i32 0, !"typeid2"}
!2 = !{i32 4, !"typeid2"}
!3 = !{i32 0, !"typeid3"}

declare i1 @llvm.type.test(i8* %ptr, metadata %typeid) nounwind readnone

define i1 @foo(i32* %p) {
    %pi8 = bitcast i32* %p to i8*
    %x = call i1 @llvm.type.test(i8* %pi8, metadata !"typeid1")
    ret i1 %x
}

define i1 @bar(i32* %p) {
    %pi8 = bitcast i32* %p to i8*
    %x = call i1 @llvm.type.test(i8* %pi8, metadata !"typeid2")
    ret i1 %x
}

define i1 @baz(void ()* %p) {
    %pi8 = bitcast void ()* %p to i8*
    %x = call i1 @llvm.type.test(i8* %pi8, metadata !"typeid3")
    ret i1 %x
}

define void @main() {
    %a1 = call i1 @foo(i32* @a) ; returns 1
    %b1 = call i1 @foo(i32* @b) ; returns 1
    %c1 = call i1 @foo(i32* @c) ; returns 0
    %a2 = call i1 @bar(i32* @a) ; returns 0
    %b2 = call i1 @bar(i32* @b) ; returns 1
    %c2 = call i1 @bar(i32* @c) ; returns 1
    %d02 = call i1 @bar(i32* getelementptr ([2 x i32]* @d, i32 0, i32 0)) ; returns 0
    %d12 = call i1 @bar(i32* getelementptr ([2 x i32]* @d, i32 0, i32 1)) ; returns 1
    %e = call i1 @baz(void ()* @e) ; returns 1
    %f = call i1 @baz(void ()* @f) ; returns 0
    %g = call i1 @baz(void ()* @g) ; returns 1
    ret void
}

```

4.32 FaultMaps and implicit checks

- *Motivation*
- *The Fault Map Section*
- *The ImplicitNullChecks pass*

```
- make.implicit.metadata
```

4.32.1 Motivation

Code generated by managed language runtimes tend to have checks that are required for safety but never fail in practice. In such cases, it is profitable to make the non-failing case cheaper even if it makes the failing case significantly more expensive. This asymmetry can be exploited by folding such safety checks into operations that can be made to fault reliably if the check would have failed, and recovering from such a fault by using a signal handler.

For example, Java requires null checks on objects before they are read from or written to. If the object is `null` then a `NullPointerException` has to be thrown, interrupting normal execution. In practice, however, dereferencing a null pointer is extremely rare in well-behaved Java programs, and typically the null check can be folded into a nearby memory operation that operates on the same memory location.

4.32.2 The Fault Map Section

Information about implicit checks generated by LLVM are put in a special “fault map” section. On Darwin this section is named `__llvm_faultmaps`.

The format of this section is

```
Header {
  uint8  : Fault Map Version (current version is 1)
  uint8  : Reserved (expected to be 0)
  uint16 : Reserved (expected to be 0)
}
uint32 : NumFunctions
FunctionInfo[NumFunctions] {
  uint64 : FunctionAddress
  uint32 : NumFaultingPCs
  uint32 : Reserved (expected to be 0)
  FunctionFaultInfo[NumFaultingPCs] {
    uint32 : FaultKind
    uint32 : FaultingPCOffset
    uint32 : HandlerPCOffset
  }
}
```

`FaultKind` describes the reason of expected fault. Currently three kind of faults are supported:

1. `FaultMaps::FaultingLoad` - fault due to load from memory.
2. `FaultMaps::FaultingLoadStore` - fault due to instruction load and store.
3. `FaultMaps::FaultingStore` - fault due to store to memory.

4.32.3 The ImplicitNullChecks pass

The `ImplicitNullChecks` pass transforms explicit control flow for checking if a pointer is null, like:

```
%ptr = call i32* @get_ptr()
%ptr_is_null = icmp i32* %ptr, null
br i1 %ptr_is_null, label %is_null, label %not_null, !make.implicit !0

not_null:
```

```

    %t = load i32, i32* %ptr
    br label %do_something_with_t

is_null:
    call void @HFC()
    unreachable

!0 = !{}

```

to control flow implicit in the instruction loading or storing through the pointer being null checked:

```

    %ptr = call i32* @get_ptr()
    %t = load i32, i32* %ptr ;; handler-pc = label %is_null
    br label %do_something_with_t

is_null:
    call void @HFC()
    unreachable

```

This transform happens at the `MachineInstr` level, not the LLVM IR level (so the above example is only representative, not literal). The `ImplicitNullChecks` pass runs during codegen, if `-enable-implicit-null-checks` is passed to `llc`.

The `ImplicitNullChecks` pass adds entries to the `__llvm_faultmaps` section described above as needed.

`make.implicit` metadata

Making null checks implicit is an aggressive optimization, and it can be a net performance pessimization if too many memory operations end up faulting because of it. A language runtime typically needs to ensure that only a negligible number of implicit null checks actually fault once the application has reached a steady state. A standard way of doing this is by healing failed implicit null checks into explicit null checks via code patching or recompilation. It follows that there are two requirements an explicit null check needs to satisfy for it to be profitable to convert it to an implicit null check:

1. The case where the pointer is actually null (i.e. the “failing” case) is extremely rare.
2. The failing path heals the implicit null check into an explicit null check so that the application does not repeatedly page fault.

The frontend is expected to mark branches that satisfy (1) and (2) using a `!make.implicit` metadata node (the actual content of the metadata node is ignored). Only branches that are marked with `!make.implicit` metadata are considered as candidates for conversion into implicit null checks.

(Note that while we could deal with (1) using profiling data, dealing with (2) requires some information not present in branch profiles.)

4.33 Machine IR (MIR) Format Reference Manual

- *Introduction*
- *Overview*
- *MIR Testing Guide*

- *Testing Individual Code Generation Passes*
 - * *Simplifying MIR files*
- *Limitations*
- *High Level Structure*
 - *Embedded Module*
 - *Machine Functions*
- *Machine Instructions Format Reference*
 - *Machine Basic Blocks*
 - * *Block References*
 - * *Successors*
 - * *Live In Registers*
 - * *Miscellaneous Attributes*
 - *Machine Instructions*
 - * *Instruction Flags*
 - *Registers*
 - *Machine Operands*
 - * *Immediate Operands*
 - * *Register Operands*
 - *Register Flags*
 - *Subregister Indices*
 - * *Global Value Operands*

Warning: This is a work in progress.

4.33.1 Introduction

This document is a reference manual for the Machine IR (MIR) serialization format. MIR is a human readable serialization format that is used to represent LLVM’s *machine specific intermediate representation*.

The MIR serialization format is designed to be used for testing the code generation passes in LLVM.

4.33.2 Overview

The MIR serialization format uses a YAML container. YAML is a standard data serialization language, and the full YAML language spec can be read at yaml.org.

A MIR file is split up into a series of [YAML documents](#). The first document can contain an optional embedded LLVM IR module, and the rest of the documents contain the serialized machine functions.

4.33.3 MIR Testing Guide

You can use the MIR format for testing in two different ways:

- You can write MIR tests that invoke a single code generation pass using the `-run-pass` option in `llc`.
- You can use `llc`'s `-stop-after` option with existing or new LLVM assembly tests and check the MIR output of a specific code generation pass.

Testing Individual Code Generation Passes

The `-run-pass` option in `llc` allows you to create MIR tests that invoke just a single code generation pass. When this option is used, `llc` will parse an input MIR file, run the specified code generation pass(es), and output the resulting MIR code.

You can generate an input MIR file for the test by using the `-stop-after` or `-stop-before` option in `llc`. For example, if you would like to write a test for the post register allocation pseudo instruction expansion pass, you can specify the machine copy propagation pass in the `-stop-after` option, as it runs just before the pass that we are trying to test:

```
llc -stop-after=machine-cp bug-trigger.ll > test.mir
```

After generating the input MIR file, you'll have to add a run line that uses the `-run-pass` option to it. In order to test the post register allocation pseudo instruction expansion pass on X86-64, a run line like the one shown below can be used:

```
# RUN: llc -o - %s -mtriple=x86_64-- -run-pass=postrapseudos |
FileCheck %s
```

The MIR files are target dependent, so they have to be placed in the target specific test directories (`lib/CodeGen/TARGETNAME`). They also need to specify a target triple or a target architecture either in the run line or in the embedded LLVM IR module.

Simplifying MIR files

The MIR code coming out of `-stop-after/-stop-before` is very verbose; Tests are more accessible and future proof when simplified:

- Use the `-simplify-mir` option with `llc`.
- Machine function attributes often have default values or the test works just as well with default values. Typical candidates for this are: *alignment:*, *exposesReturnsTwice*, *legalized*, *regBankSelected*, *selected*. The whole *frameInfo* section is often unnecessary if there is no special frame usage in the function. *tracksRegLiveness* on the other hand is often necessary for some passes that care about block livein lists.
- The (global) *liveins:* list is typically only interesting for early instruction selection passes and can be removed when testing later passes. The per-block *liveins:* on the other hand are necessary if *tracksRegLiveness* is true.
- Branch probability data in block *successors:* lists can be dropped if the test doesn't depend on it. Example: *successors: %bb.1(0x40000000), %bb.2(0x40000000)* can be replaced with *successors: %bb.1, %bb.2*.
- MIR code contains a whole IR module. This is necessary because there are no equivalents in MIR for global variables, references to external functions, function attributes, metadata, debug info. Instead some MIR data references the IR constructs. You can often remove them if the test doesn't depend on them.
- Alias Analysis is performed on IR values. These are referenced by memory operands in MIR. Example: `:: (load 8 from %ir.foobar, !alias.scope !9)`. If the test doesn't depend on (good) alias analysis the references can be dropped: `:: (load 8)`

- MIR blocks can reference IR blocks for debug printing, profile information or debug locations. Example: *bb.42.myblock* in MIR references the IR block *myblock*. It is usually possible to drop the *.myblock* reference and simply use *bb.42*.
- If there are no memory operands or blocks referencing the IR then the IR function can be replaced by a parameterless dummy function like *define @func() { ret void }*.
- It is possible to drop the whole IR section of the MIR file if it only contains dummy functions (see above). The .mir loader will create the IR functions automatically in this case.

Limitations

Currently the MIR format has several limitations in terms of which state it can serialize:

- The target-specific state in the target-specific `MachineFunctionInfo` subclasses isn't serialized at the moment.
- The target-specific `MachineConstantPoolValue` subclasses (in the ARM and SystemZ backends) aren't serialized at the moment.
- The `MCSymbol` machine operands are only printed, they can't be parsed.
- A lot of the state in `MachineModuleInfo` isn't serialized - only the CFI instructions and the variable debug information from MMI is serialized right now.

These limitations impose restrictions on what you can test with the MIR format. For now, tests that would like to test some behaviour that depends on the state of certain `MCSymbol` operands or the exception handling state in MMI, can't use the MIR format. As well as that, tests that test some behaviour that depends on the state of the target specific `MachineFunctionInfo` or `MachineConstantPoolValue` subclasses can't use the MIR format at the moment.

4.33.4 High Level Structure

Embedded Module

When the first YAML document contains a [YAML block literal string](#), the MIR parser will treat this string as an LLVM assembly language string that represents an embedded LLVM IR module. Here is an example of a YAML document that contains an LLVM module:

```
define i32 @inc(i32* %x) {
entry:
  %0 = load i32, i32* %x
  %1 = add i32 %0, 1
  store i32 %1, i32* %x
  ret i32 %1
}
```

Machine Functions

The remaining YAML documents contain the machine functions. This is an example of such YAML document:

```
---
name:          inc
tracksRegLiveness: true
liveins:
- { reg: '%rdi' }
```



```
body: |
  bb.0.entry:
    liveins: %rdi

    %eax = MOV32rm %rdi, 1, __, 0, _
    %eax = INC32r killed %eax, implicit-def dead %eflags
    MOV32mr killed %rdi, 1, __, 0, __, %eax
    RETQ %eax
...
```

The document above consists of attributes that represent the various properties and data structures in a machine function.

The attribute `name` is required, and its value should be identical to the name of a function that this machine function is based on.

The attribute `body` is a [YAML block literal string](#). Its value represents the function's machine basic blocks and their machine instructions.

4.33.5 Machine Instructions Format Reference

The machine basic blocks and their instructions are represented using a custom, human readable serialization language. This language is used in the [YAML block literal string](#) that corresponds to the machine function's body.

A source string that uses this language contains a list of machine basic blocks, which are described in the section below.

Machine Basic Blocks

A machine basic block is defined in a single block definition source construct that contains the block's ID. The example below defines two blocks that have an ID of zero and one:

```
bb.0:
  <instructions>
bb.1:
  <instructions>
```

A machine basic block can also have a name. It should be specified after the ID in the block's definition:

```
bb.0.entry:      ; This block's name is "entry"
  <instructions>
```

The block's name should be identical to the name of the IR block that this machine block is based on.

Block References

The machine basic blocks are identified by their ID numbers. Individual blocks are referenced using the following syntax:

```
%bb.<id>[.<name>]
```

Examples:

```
%bb.0  
%bb.1.then
```

Successors

The machine basic block's successors have to be specified before any of the instructions:

```
bb.0.entry:  
  successors: %bb.1.then, %bb.2.else  
  <instructions>  
bb.1.then:  
  <instructions>  
bb.2.else:  
  <instructions>
```

The branch weights can be specified in brackets after the successor blocks. The example below defines a block that has two successors with branch weights of 32 and 16:

```
bb.0.entry:  
  successors: %bb.1.then(32), %bb.2.else(16)
```

Live In Registers

The machine basic block's live in registers have to be specified before any of the instructions:

```
bb.0.entry:  
  liveins: %edi, %esi
```

The list of live in registers and successors can be empty. The language also allows multiple live in register and successor lists - they are combined into one list by the parser.

Miscellaneous Attributes

The attributes `IsAddressTaken`, `IsLandingPad` and `Alignment` can be specified in brackets after the block's definition:

```
bb.0.entry (address-taken):  
  <instructions>  
bb.2.else (align 4):  
  <instructions>  
bb.3(landing-pad, align 4):  
  <instructions>
```

Machine Instructions

A machine instruction is composed of a name, *machine operands*, *instruction flags*, and machine memory operands.

The instruction's name is usually specified before the operands. The example below shows an instance of the X86 `RETQ` instruction with a single machine operand:

```
RETQ %eax
```

However, if the machine instruction has one or more explicitly defined register operands, the instruction's name has to be specified after them. The example below shows an instance of the AArch64 `LDPXpost` instruction with three defined register operands:

```
%sp, %fp, %lr = LDPXpost %sp, 2
```

The instruction names are serialized using the exact definitions from the target's `*InstrInfo.td` files, and they are case sensitive. This means that similar instruction names like `TSTri` and `tSTri` represent different machine instructions.

Instruction Flags

The flag `frame-setup` can be specified before the instruction's name:

```
%fp = frame-setup ADDXri %sp, 0, 0
```

Registers

Registers are one of the key primitives in the machine instructions serialization language. They are primarily used in the *register machine operands*, but they can also be used in a number of other places, like the *basic block's live in list*.

The physical registers are identified by their name. They use the following syntax:

```
%<name>
```

The example below shows three X86 physical registers:

```
%eax
%r15
%eflags
```

The virtual registers are identified by their ID number. They use the following syntax:

```
%<id>
```

Example:

```
%0
```

The null registers are represented using an underscore (`'_'`). They can also be represented using a `'%noreg'` named register, although the former syntax is preferred.

Machine Operands

There are seventeen different kinds of machine operands, and all of them, except the `MCSymbol` operand, can be serialized. The `MCSymbol` operands are just printed out - they can't be parsed back yet.

Immediate Operands

The immediate machine operands are untyped, 64-bit signed integers. The example below shows an instance of the X86 `MOV32ri` instruction that has an immediate machine operand `-42`:

```
%eax = MOV32ri -42
```

Register Operands

The *register* primitive is used to represent the register machine operands. The register operands can also have optional *register flags*, a *subregister index*, and a reference to the tied register operand. The full syntax of a register operand is shown below:

```
[<flags>] <register> [ :<subregister-idx-name> ] [ (tied-def <tied-op>) ]
```

This example shows an instance of the X86 `XOR32rr` instruction that has 5 register operands with different register flags:

```
dead %eax = XOR32rr undef %eax, undef %eax, implicit-def dead %eflags, implicit-def  
↳ %al
```

Register Flags

The table below shows all of the possible register flags along with the corresponding internal `llvm::RegState` representation:

Flag	Internal Value
<code>implicit</code>	<code>RegState::Implicit</code>
<code>implicit-def</code>	<code>RegState::ImplicitDefine</code>
<code>def</code>	<code>RegState::Define</code>
<code>dead</code>	<code>RegState::Dead</code>
<code>killed</code>	<code>RegState::Kill</code>
<code>undef</code>	<code>RegState::Undef</code>
<code>internal</code>	<code>RegState::InternalRead</code>
<code>early-clobber</code>	<code>RegState::EarlyClobber</code>
<code>debug-use</code>	<code>RegState::Debug</code>

Subregister Indices

The register machine operands can reference a portion of a register by using the subregister indices. The example below shows an instance of the `COPY` pseudo instruction that uses the X86 `sub_8bit` subregister index to copy 8 lower bits from the 32-bit virtual register 0 to the 8-bit virtual register 1:

```
%1 = COPY %0:sub_8bit
```

The names of the subregister indices are target specific, and are typically defined in the target's `*RegisterInfo.td` file.

Global Value Operands

The global value machine operands reference the global values from the *embedded LLVM IR module*. The example below shows an instance of the X86 `MOV64rm` instruction that has a global value operand named `G`:

```
%rax = MOV64rm %rip, 1, __, @G, __
```

The named global values are represented using an identifier with the ‘@’ prefix. If the identifier doesn’t match the regular expression `[-a-zA-Z$._-][-a-zA-Z$._-0-9]*`, then this identifier must be quoted.

The unnamed global values are represented using an unsigned numeric value with the ‘@’ prefix, like in the following examples: `@0`, `@989`.

4.34 Coroutines in LLVM

- *Introduction*
- *Coroutines by Example*
 - *Coroutine Representation*
 - *Coroutine Transformation*
 - *Avoiding Heap Allocations*
 - *Multiple Suspend Points*
 - *Distinct Save and Suspend*
 - *Coroutine Promise*
 - *Final Suspend*
- *Intrinsics*
 - *Coroutine Manipulation Intrinsics*
 - * *‘llvm.coro.destroy’ Intrinsic*
 - * *‘llvm.coro.resume’ Intrinsic*
 - * *‘llvm.coro.done’ Intrinsic*
 - * *‘llvm.coro.promise’ Intrinsic*
 - *Coroutine Structure Intrinsics*
 - * *‘llvm.coro.size’ Intrinsic*
 - * *‘llvm.coro.begin’ Intrinsic*
 - * *‘llvm.coro.free’ Intrinsic*
 - * *‘llvm.coro.alloc’ Intrinsic*
 - * *‘llvm.coro.frame’ Intrinsic*
 - * *‘llvm.coro.id’ Intrinsic*
 - * *‘llvm.coro.end’ Intrinsic*
 - * *‘llvm.coro.suspend’ Intrinsic*

- * *'llvm.coro.save' Intrinsic*
- * *'llvm.coro.param' Intrinsic*

- *Coroutine Transformation Passes*
 - *CoroEarly*
 - *CoroSplit*
 - *CoroElide*
 - *CoroCleanup*
- *Areas Requiring Attention*

Warning: This is a work in progress. Compatibility across LLVM releases is not guaranteed.

4.34.1 Introduction

LLVM coroutines are functions that have one or more *suspend points*. When a suspend point is reached, the execution of a coroutine is suspended and control is returned back to its caller. A suspended coroutine can be resumed to continue execution from the last suspend point or it can be destroyed.

In the following example, we call function *f* (which may or may not be a coroutine itself) that returns a handle to a suspended coroutine (**coroutine handle**) that is used by *main* to resume the coroutine twice and then destroy it:

```
define i32 @main() {
entry:
  %hdl = call i8* @f(i32 4)
  call void @llvm.coro.resume(i8* %hdl)
  call void @llvm.coro.resume(i8* %hdl)
  call void @llvm.coro.destroy(i8* %hdl)
  ret i32 0
}
```

In addition to the function stack frame which exists when a coroutine is executing, there is an additional region of storage that contains objects that keep the coroutine state when a coroutine is suspended. This region of storage is called **coroutine frame**. It is created when a coroutine is called and destroyed when a coroutine runs to completion or destroyed by a call to the *coro.destroy* intrinsic.

An LLVM coroutine is represented as an LLVM function that has calls to *coroutine intrinsics* defining the structure of the coroutine. After lowering, a coroutine is split into several functions that represent three different ways of how control can enter the coroutine:

1. a ramp function, which represents an initial invocation of the coroutine that creates the coroutine frame and executes the coroutine code until it encounters a suspend point or reaches the end of the function;
2. a coroutine resume function that is invoked when the coroutine is resumed;
3. a coroutine destroy function that is invoked when the coroutine is destroyed.

Note: Splitting out resume and destroy functions are just one of the possible ways of lowering the coroutine. We chose it for initial implementation as it matches closely the mental model and results in reasonably nice code.

4.34.2 Coroutines by Example

Coroutine Representation

Let's look at an example of an LLVM coroutine with the behavior sketched by the following pseudo-code.

```
void *f(int n) {
    for(;;) {
        print(n++);
        <suspend> // returns a coroutine handle on first suspend
    }
}
```

This coroutine calls some function *print* with value *n* as an argument and suspends execution. Every time this coroutine resumes, it calls *print* again with an argument one bigger than the last time. This coroutine never completes by itself and must be destroyed explicitly. If we use this coroutine with a *main* shown in the previous section. It will call *print* with values 4, 5 and 6 after which the coroutine will be destroyed.

The LLVM IR for this coroutine looks like this:

```
define i8* @f(i32 %n) {
entry:
    %id = call token @llvm.coro.id(i32 0, i8* null, i8* null, i8* null)
    %size = call i32 @llvm.coro.size.i32()
    %alloc = call i8* @malloc(i32 %size)
    %hdl = call noalias i8* @llvm.coro.begin(token %id, i8* %alloc)
    br label %loop
loop:
    %n.val = phi i32 [ %n, %entry ], [ %inc, %loop ]
    %inc = add nsw i32 %n.val, 1
    call void @print(i32 %n.val)
    %0 = call i8 @llvm.coro.suspend(token none, i1 false)
    switch i8 %0, label %suspend [i8 0, label %loop
                                i8 1, label %cleanup]
cleanup:
    %mem = call i8* @llvm.coro.free(token %id, i8* %hdl)
    call void @free(i8* %mem)
    br label %suspend
suspend:
    %unused = call i1 @llvm.coro.end(i8* %hdl, i1 false)
    ret i8* %hdl
}
```

The *entry* block establishes the coroutine frame. The *coro.size* intrinsic is lowered to a constant representing the size required for the coroutine frame. The *coro.begin* intrinsic initializes the coroutine frame and returns the coroutine handle. The second parameter of *coro.begin* is given a block of memory to be used if the coroutine frame needs to be allocated dynamically. The *coro.id* intrinsic serves as coroutine identity useful in cases when the *coro.begin* intrinsic get duplicated by optimization passes such as jump-threading.

The *cleanup* block destroys the coroutine frame. The *coro.free* intrinsic, given the coroutine handle, returns a pointer of the memory block to be freed or *null* if the coroutine frame was not allocated dynamically. The *cleanup* block is entered when coroutine runs to completion by itself or destroyed via call to the *coro.destroy* intrinsic.

The *suspend* block contains code to be executed when coroutine runs to completion or suspended. The *coro.end* intrinsic marks the point where a coroutine needs to return control back to the caller if it is not an initial invocation of the coroutine.

The *loop* blocks represents the body of the coroutine. The *coro.suspend* intrinsic in combination with the following

switch indicates what happens to control flow when a coroutine is suspended (default case), resumed (case 0) or destroyed (case 1).

Coroutine Transformation

One of the steps of coroutine lowering is building the coroutine frame. The def-use chains are analyzed to determine which objects need be kept alive across suspend points. In the coroutine shown in the previous section, use of virtual register `%n.val` is separated from the definition by a suspend point, therefore, it cannot reside on the stack frame since the latter goes away once the coroutine is suspended and control is returned back to the caller. An i32 slot is allocated in the coroutine frame and `%n.val` is spilled and reloaded from that slot as needed.

We also store addresses of the resume and destroy functions so that the `coro.resume` and `coro.destroy` intrinsics can resume and destroy the coroutine when its identity cannot be determined statically at compile time. For our example, the coroutine frame will be:

```
%f.frame = type { void (%f.frame*)*, void (%f.frame*)*, i32 }
```

After resume and destroy parts are outlined, function `f` will contain only the code responsible for creation and initialization of the coroutine frame and execution of the coroutine until a suspend point is reached:

```
define i8* @f(i32 %n) {
entry:
  %id = call token @llvm.coro.id(i32 0, i8* null, i8* null, i8* null)
  %alloc = call noalias i8* @malloc(i32 24)
  %0 = call noalias i8* @llvm.coro.begin(token %id, i8* %alloc)
  %frame = bitcast i8* %0 to %f.frame*
  %1 = getelementptr %f.frame, %f.frame* %frame, i32 0, i32 0
  store void (%f.frame*)* @f.resume, void (%f.frame*)** %1
  %2 = getelementptr %f.frame, %f.frame* %frame, i32 0, i32 1
  store void (%f.frame*)* @f.destroy, void (%f.frame*)** %2

  %inc = add nsw i32 %n, 1
  %inc.spill.addr = getelementptr inbounds %f.Frame, %f.Frame* %FramePtr, i32 0, i32 2
  store i32 %inc, i32* %inc.spill.addr
  call void @print(i32 %n)

  ret i8* %frame
}
```

Outlined resume part of the coroutine will reside in function `f.resume`:

```
define internal fastcc void @f.resume(%f.frame* %frame.ptr.resume) {
entry:
  %inc.spill.addr = getelementptr %f.frame, %f.frame* %frame.ptr.resume, i64 0, i32 2
  %inc.spill = load i32, i32* %inc.spill.addr, align 4
  %inc = add i32 %n.val, 1
  store i32 %inc, i32* %inc.spill.addr, align 4
  tail call void @print(i32 %inc)
  ret void
}
```

Whereas function `f.destroy` will contain the cleanup code for the coroutine:

```
define internal fastcc void @f.destroy(%f.frame* %frame.ptr.destroy) {
entry:
  %0 = bitcast %f.frame* %frame.ptr.destroy to i8*
  tail call void @free(i8* %0)
}
```



```

    ret void
}

```

Avoiding Heap Allocations

A particular coroutine usage pattern, which is illustrated by the *main* function in the overview section, where a coroutine is created, manipulated and destroyed by the same calling function, is common for coroutines implementing RAII idiom and is suitable for allocation elision optimization which avoid dynamic allocation by storing the coroutine frame as a static *alloca* in its caller.

In the entry block, we will call *coro.alloc* intrinsic that will return *true* when dynamic allocation is required, and *false* if dynamic allocation is elided.

```

entry:
    %id = call token @llvm.coro.id(i32 0, i8* null, i8* null, i8* null)
    %need.dyn.alloc = call i1 @llvm.coro.alloc(token %id)
    br i1 %need.dyn.alloc, label %dyn.alloc, label %coro.begin
dyn.alloc:
    %size = call i32 @llvm.coro.size.i32()
    %alloc = call i8* @CustomAlloc(i32 %size)
    br label %coro.begin
coro.begin:
    %phi = phi i8* [ null, %entry ], [ %alloc, %dyn.alloc ]
    %hdl = call noalias i8* @llvm.coro.begin(token %id, i8* %phi)

```

In the cleanup block, we will make freeing the coroutine frame conditional on *coro.free* intrinsic. If allocation is elided, *coro.free* returns *null* thus skipping the deallocation code:

```

cleanup:
    %mem = call i8* @llvm.coro.free(token %id, i8* %hdl)
    %need.dyn.free = icmp ne i8* %mem, null
    br i1 %need.dyn.free, label %dyn.free, label %if.end
dyn.free:
    call void @CustomFree(i8* %mem)
    br label %if.end
if.end:
    ...

```

With allocations and deallocations represented as described as above, after coroutine heap allocation elision optimization, the resulting main will be:

```

define i32 @main() {
entry:
    call void @print(i32 4)
    call void @print(i32 5)
    call void @print(i32 6)
    ret i32 0
}

```

Multiple Suspend Points

Let's consider the coroutine that has more than one suspend point:

```
void *f(int n) {  
    for(;;) {  
        print(n++);  
        <suspend>  
        print(-n);  
        <suspend>  
    }  
}
```

Matching LLVM code would look like (with the rest of the code remaining the same as the code in the previous section):

```
loop:  
    %n.addr = phi i32 [ %n, %entry ], [ %inc, %loop.resume ]  
    call void @print(i32 %n.addr) #4  
    %2 = call i8 @llvm.coro.suspend(token none, i1 false)  
    switch i8 %2, label %suspend [i8 0, label %loop.resume  
                                i8 1, label %cleanup]  
loop.resume:  
    %inc = add nsw i32 %n.addr, 1  
    %sub = xor i32 %n.addr, -1  
    call void @print(i32 %sub)  
    %3 = call i8 @llvm.coro.suspend(token none, i1 false)  
    switch i8 %3, label %suspend [i8 0, label %loop  
                                i8 1, label %cleanup]
```

In this case, the coroutine frame would include a suspend index that will indicate at which suspend point the coroutine needs to resume. The resume function will use an index to jump to an appropriate basic block and will look as follows:

```
define internal fastcc void @f.Resume(%f.Frame* %FramePtr) {  
entry.Resume:  
    %index.addr = getelementptr inbounds %f.Frame, %f.Frame* %FramePtr, i64 0, i32 2  
    %index = load i8, i8* %index.addr, align 1  
    %switch = icmp eq i8 %index, 0  
    %n.addr = getelementptr inbounds %f.Frame, %f.Frame* %FramePtr, i64 0, i32 3  
    %n = load i32, i32* %n.addr, align 4  
    br i1 %switch, label %loop.resume, label %loop  
loop.resume:  
    %sub = xor i32 %n, -1  
    call void @print(i32 %sub)  
    br label %suspend  
loop:  
    %inc = add nsw i32 %n, 1  
    store i32 %inc, i32* %n.addr, align 4  
    tail call void @print(i32 %inc)  
    br label %suspend  
suspend:  
    %storemerge = phi i8 [ 0, %loop ], [ 1, %loop.resume ]  
    store i8 %storemerge, i8* %index.addr, align 1  
    ret void  
}
```

If different cleanup code needs to get executed for different suspend points, a similar switch will be in the *f.destroy* function.

Note: Using suspend index in a coroutine state and having a switch in *f.resume* and *f.destroy* is one of the possible implementation strategies. We explored another option where a distinct *f.resume1*, *f.resume2*, etc. are created for every suspend point, and instead of storing an index, the resume and destroy function pointers are updated at every suspend. Early testing showed that the current approach is easier on the optimizer than the latter so it is a lowering strategy implemented at the moment.

Distinct Save and Suspend

In the previous example, setting a resume index (or some other state change that needs to happen to prepare a coroutine for resumption) happens at the same time as a suspension of a coroutine. However, in certain cases, it is necessary to control when coroutine is prepared for resumption and when it is suspended.

In the following example, a coroutine represents some activity that is driven by completions of asynchronous operations *async_op1* and *async_op2* which get a coroutine handle as a parameter and resume the coroutine once async operation is finished.

```
void g() {
  for (;;)
    if (cond()) {
      async_op1(<coroutine-handle>); // will resume once async_op1 completes
      <suspend>
      do_one();
    }
    else {
      async_op2(<coroutine-handle>); // will resume once async_op2 completes
      <suspend>
      do_two();
    }
}
```

In this case, coroutine should be ready for resumption prior to a call to *async_op1* and *async_op2*. The *coro.save* intrinsic is used to indicate a point when coroutine should be ready for resumption (namely, when a resume index should be stored in the coroutine frame, so that it can be resumed at the correct resume point):

```
if.true:
  %save1 = call token @llvm.coro.save(i8* %hdl)
  call void @async_op1(i8* %hdl)
  %suspend1 = call i1 @llvm.coro.suspend(token %save1, i1 false)
  switch i8 %suspend1, label %suspend [i8 0, label %resume1
                                     i8 1, label %cleanup]
if.false:
  %save2 = call token @llvm.coro.save(i8* %hdl)
  call void @async_op2(i8* %hdl)
  %suspend2 = call i1 @llvm.coro.suspend(token %save2, i1 false)
  switch i8 %suspend1, label %suspend [i8 0, label %resume2
                                     i8 1, label %cleanup]
```

Coroutine Promise

A coroutine author or a frontend may designate a distinguished *alloca* that can be used to communicate with the coroutine. This distinguished *alloca* is called **coroutine promise** and is provided as the second parameter to the *coro.id* intrinsic.

The following coroutine designates a 32 bit integer *promise* and uses it to store the current value produced by a coroutine.

```
define i8* @f(i32 %n) {
entry:
    %promise = alloca i32
    %pv = bitcast i32* %promise to i8*
    %id = call token @llvm.coro.id(i32 0, i8* %pv, i8* null, i8* null)
    %need.dyn.alloc = call i1 @llvm.coro.alloc(token %id)
    br i1 %need.dyn.alloc, label %dyn.alloc, label %coro.begin
dyn.alloc:
    %size = call i32 @llvm.coro.size.i32()
    %alloc = call i8* @malloc(i32 %size)
    br label %coro.begin
coro.begin:
    %phi = phi i8* [ null, %entry ], [ %alloc, %dyn.alloc ]
    %hdl = call noalias i8* @llvm.coro.begin(token %id, i8* %phi)
    br label %loop
loop:
    %n.val = phi i32 [ %n, %coro.begin ], [ %inc, %loop ]
    %inc = add nsw i32 %n.val, 1
    store i32 %n.val, i32* %promise
    %0 = call i8 @llvm.coro.suspend(token none, i1 false)
    switch i8 %0, label %suspend [i8 0, label %loop
                                i8 1, label %cleanup]
cleanup:
    %mem = call i8* @llvm.coro.free(token %id, i8* %hdl)
    call void @free(i8* %mem)
    br label %suspend
suspend:
    %unused = call i1 @llvm.coro.end(i8* %hdl, i1 false)
    ret i8* %hdl
}
```

A coroutine consumer can rely on the *coro.promise* intrinsic to access the coroutine promise.

```
define i32 @main() {
entry:
    %hdl = call i8* @f(i32 4)
    %promise.addr.raw = call i8* @llvm.coro.promise(i8* %hdl, i32 4, i1 false)
    %promise.addr = bitcast i8* %promise.addr.raw to i32*
    %val0 = load i32, i32* %promise.addr
    call void @print(i32 %val0)
    call void @llvm.coro.resume(i8* %hdl)
    %val1 = load i32, i32* %promise.addr
    call void @print(i32 %val1)
    call void @llvm.coro.resume(i8* %hdl)
    %val2 = load i32, i32* %promise.addr
    call void @print(i32 %val2)
    call void @llvm.coro.destroy(i8* %hdl)
    ret i32 0
}
```

After example in this section is compiled, result of the compilation will be:

```
define i32 @main() {
entry:
    tail call void @print(i32 4)
    tail call void @print(i32 5)
}
```

```
tail call void @print(i32 6)
ret i32 0
}
```

Final Suspend

A coroutine author or a frontend may designate a particular suspend to be final, by setting the second argument of the `coro.suspend` intrinsic to `true`. Such a suspend point has two properties:

- it is possible to check whether a suspended coroutine is at the final suspend point via `coro.done` intrinsic;
- a resumption of a coroutine stopped at the final suspend point leads to undefined behavior. The only possible action for a coroutine at a final suspend point is destroying it via `coro.destroy` intrinsic.

From the user perspective, the final suspend point represents an idea of a coroutine reaching the end. From the compiler perspective, it is an optimization opportunity for reducing number of resume points (and therefore switch cases) in the resume function.

The following is an example of a function that keeps resuming the coroutine until the final suspend point is reached after which point the coroutine is destroyed:

```
define i32 @main() {
entry:
    %hdl = call i8* @f(i32 4)
    br label %while
while:
    call void @llvm.coro.resume(i8* %hdl)
    %done = call i1 @llvm.coro.done(i8* %hdl)
    br i1 %done, label %end, label %while
end:
    call void @llvm.coro.destroy(i8* %hdl)
    ret i32 0
}
```

Usually, final suspend point is a frontend injected suspend point that does not correspond to any explicitly authored suspend point of the high level language. For example, for a Python generator that has only one suspend point:

```
def coroutine(n):
    for i in range(n):
        yield i
```

Python frontend would inject two more suspend points, so that the actual code looks like this:

```
void* coroutine(int n) {
    int current_value;
    <designate current_value to be coroutine promise>
    <SUSPEND> // injected suspend point, so that the coroutine starts suspended
    for (int i = 0; i < n; ++i) {
        current_value = i; <SUSPEND>; // corresponds to "yield i"
    }
    <SUSPEND final=true> // injected final suspend point
}
```

and python iterator `__next__` would look like:

```
int __next__(void* hdl) {
    coro.resume(hdl);
```

```
if (coro.done(hdl)) throw StopIteration();  
return *(int*)coro.promise(hdl, 4, false);  
}
```

4.34.3 Intrinsics

Coroutine Manipulation Intrinsics

Intrinsics described in this section are used to manipulate an existing coroutine. They can be used in any function which happen to have a pointer to a *coroutine frame* or a pointer to a *coroutine promise*.

‘llvm.coro.destroy’ Intrinsic

Syntax:

```
declare void @llvm.coro.destroy(i8* <handle>)
```

Overview:

The ‘llvm.coro.destroy’ intrinsic destroys a suspended coroutine.

Arguments:

The argument is a coroutine handle to a suspended coroutine.

Semantics:

When possible, the *coro.destroy* intrinsic is replaced with a direct call to the coroutine destroy function. Otherwise it is replaced with an indirect call based on the function pointer for the destroy function stored in the coroutine frame. Destroying a coroutine that is not suspended leads to undefined behavior.

‘llvm.coro.resume’ Intrinsic

```
declare void @llvm.coro.resume(i8* <handle>)
```

Overview:

The ‘llvm.coro.resume’ intrinsic resumes a suspended coroutine.

Arguments:

The argument is a handle to a suspended coroutine.

Semantics:

When possible, the *coro.resume* intrinsic is replaced with a direct call to the coroutine resume function. Otherwise it is replaced with an indirect call based on the function pointer for the resume function stored in the coroutine frame. Resuming a coroutine that is not suspended leads to undefined behavior.

'llvm.coro.done' Intrinsic

```
declare i1 @llvm.coro.done(i8* <handle>)
```

Overview:

The 'llvm.coro.done' intrinsic checks whether a suspended coroutine is at the final suspend point or not.

Arguments:

The argument is a handle to a suspended coroutine.

Semantics:

Using this intrinsic on a coroutine that does not have a *final suspend* point or on a coroutine that is not suspended leads to undefined behavior.

'llvm.coro.promise' Intrinsic

```
declare i8* @llvm.coro.promise(i8* <ptr>, i32 <alignment>, i1 <from>)
```

Overview:

The 'llvm.coro.promise' intrinsic obtains a pointer to a *coroutine promise* given a coroutine handle and vice versa.

Arguments:

The first argument is a handle to a coroutine if *from* is false. Otherwise, it is a pointer to a coroutine promise.

The second argument is an alignment requirements of the promise. If a frontend designated *%promise = alloca i32* as a promise, the alignment argument to *coro.promise* should be the alignment of *i32* on the target platform. If a frontend designated *%promise = alloca i32, align 16* as a promise, the alignment argument should be 16. This argument only accepts constants.

The third argument is a boolean indicating a direction of the transformation. If *from* is true, the intrinsic returns a coroutine handle given a pointer to a promise. If *from* is false, the intrinsic return a pointer to a promise from a coroutine handle. This argument only accepts constants.

Semantics:

Using this intrinsic on a coroutine that does not have a coroutine promise leads to undefined behavior. It is possible to read and modify coroutine promise of the coroutine which is currently executing. The coroutine author and a coroutine user are responsible to make sure there is no data races.

Example:

```
define i8* @f(i32 %n) {
entry:
  %promise = alloca i32
  %pv = bitcast i32* %promise to i8*
  ; the second argument to coro.id points to the coroutine promise.
  %id = call token @llvm.coro.id(i32 0, i8* %pv, i8* null, i8* null)
  ...
  %hdl = call noalias i8* @llvm.coro.begin(token %id, i8* %alloc)
  ...
  store i32 42, i32* %promise ; store something into the promise
  ...
  ret i8* %hdl
}

define i32 @main() {
entry:
  %hdl = call i8* @f(i32 4) ; starts the coroutine and returns its handle
  %promise.addr.raw = call i8* @llvm.coro.promise(i8* %hdl, i32 4, i1 false)
  %promise.addr = bitcast i8* %promise.addr.raw to i32*
  %val = load i32, i32* %promise.addr ; load a value from the promise
  call void @print(i32 %val)
  call void @llvm.coro.destroy(i8* %hdl)
  ret i32 0
}
```

Coroutine Structure Intrinsics

Intrinsics described in this section are used within a coroutine to describe the coroutine structure. They should not be used outside of a coroutine.

‘llvm.coro.size’ Intrinsic

```
declare i32 @llvm.coro.size.i32()
declare i64 @llvm.coro.size.i64()
```

Overview:

The ‘llvm.coro.size’ intrinsic returns the number of bytes required to store a *coroutine frame*.

Arguments:

None

Semantics:

The *coro.size* intrinsic is lowered to a constant representing the size of the coroutine frame.

‘llvm.coro.begin’ Intrinsic

```
declare i8* @llvm.coro.begin(token <id>, i8* <mem>)
```

Overview:

The ‘llvm.coro.begin’ intrinsic returns an address of the coroutine frame.

Arguments:

The first argument is a token returned by a call to ‘llvm.coro.id’ identifying the coroutine.

The second argument is a pointer to a block of memory where coroutine frame will be stored if it is allocated dynamically.

Semantics:

Depending on the alignment requirements of the objects in the coroutine frame and/or on the codegen compactness reasons the pointer returned from *coro.begin* may be at offset to the *%mem* argument. (This could be beneficial if instructions that express relative access to data can be more compactly encoded with small positive and negative offsets).

A frontend should emit exactly one *coro.begin* intrinsic per coroutine.

‘llvm.coro.free’ Intrinsic

```
declare i8* @llvm.coro.free(token %id, i8* <frame>)
```

Overview:

The ‘llvm.coro.free’ intrinsic returns a pointer to a block of memory where coroutine frame is stored or *null* if this instance of a coroutine did not use dynamically allocated memory for its coroutine frame.

Arguments:

The first argument is a token returned by a call to ‘llvm.coro.id’ identifying the coroutine.

The second argument is a pointer to the coroutine frame. This should be the same pointer that was returned by prior *coro.begin* call.

Example (custom deallocation function):

```
cleanup:
  %mem = call i8* @llvm.coro.free(token %id, i8* %frame)
  %mem_not_null = icmp ne i8* %mem, null
  br i1 %mem_not_null, label %if.then, label %if.end
if.then:
  call void @CustomFree(i8* %mem)
  br label %if.end
if.end:
  ret void
```

Example (standard deallocation functions):

```
cleanup:
  %mem = call i8* @llvm.coro.free(token %id, i8* %frame)
  call void @free(i8* %mem)
  ret void
```

‘llvm.coro.alloc’ Intrinsic

```
declare i1 @llvm.coro.alloc(token <id>)
```

Overview:

The ‘llvm.coro.alloc’ intrinsic returns *true* if dynamic allocation is required to obtain a memory for the coroutine frame and *false* otherwise.

Arguments:

The first argument is a token returned by a call to ‘llvm.coro.id’ identifying the coroutine.

Semantics:

A frontend should emit at most one *coro.alloc* intrinsic per coroutine. The intrinsic is used to suppress dynamic allocation of the coroutine frame when possible.

Example:

```
entry:
  %id = call token @llvm.coro.id(i32 0, i8* null, i8* null, i8* null)
  %dyn.alloc.required = call i1 @llvm.coro.alloc(token %id)
  br i1 %dyn.alloc.required, label %coro.alloc, label %coro.begin

coro.alloc:
  %frame.size = call i32 @llvm.coro.size()
  %alloc = call i8* @MyAlloc(i32 %frame.size)
```

```
br label %coro.begin

coro.begin:
  %phi = phi i8* [ null, %entry ], [ %alloc, %coro.alloc ]
  %frame = call i8* @llvm.coro.begin(token %id, i8* %phi)
```

'llvm.coro.frame' Intrinsic

```
declare i8* @llvm.coro.frame()
```

Overview:

The 'llvm.coro.frame' intrinsic returns an address of the coroutine frame of the enclosing coroutine.

Arguments:

None

Semantics:

This intrinsic is lowered to refer to the *coro.begin* instruction. This is a frontend convenience intrinsic that makes it easier to refer to the coroutine frame.

'llvm.coro.id' Intrinsic

```
declare token @llvm.coro.id(i32 <align>, i8* <promise>, i8* <coroaddr>,
                           i8* <fnaddrs>)
```

Overview:

The 'llvm.coro.id' intrinsic returns a token identifying a coroutine.

Arguments:

The first argument provides information on the alignment of the memory returned by the allocation function and given to *coro.begin* by the first argument. If this argument is 0, the memory is assumed to be aligned to $2 * \text{sizeof}(i8^*)$. This argument only accepts constants.

The second argument, if not *null*, designates a particular *alloca* instruction to be a *coroutine promise*.

The third argument is *null* coming out of the frontend. The CoroEarly pass sets this argument to point to the function this *coro.id* belongs to.

The fourth argument is *null* before coroutine is split, and later is replaced to point to a private global constant array containing function pointers to outlined resume and destroy parts of the coroutine.

Semantics:

The purpose of this intrinsic is to tie together *coro.id*, *coro.alloc* and *coro.begin* belonging to the same coroutine to prevent optimization passes from duplicating any of these instructions unless entire body of the coroutine is duplicated.

A frontend should emit exactly one *coro.id* intrinsic per coroutine.

'llvm.coro.end' Intrinsic

```
declare i1 @llvm.coro.end(i8* <handle>, i1 <unwind>)
```

Overview:

The 'llvm.coro.end' marks the point where execution of the resume part of the coroutine should end and control should return to the caller.

Arguments:

The first argument should refer to the coroutine handle of the enclosing coroutine. A frontend is allowed to supply null as the first parameter, in this case *coro-early* pass will replace the null with an appropriate coroutine handle value.

The second argument should be *true* if this *coro.end* is in the block that is part of the unwind sequence leaving the coroutine body due to an exception and *false* otherwise.

Semantics:

The purpose of this intrinsic is to allow frontends to mark the cleanup and other code that is only relevant during the initial invocation of the coroutine and should not be present in resume and destroy parts.

This intrinsic is lowered when a coroutine is split into the start, resume and destroy parts. In the start part, it is a no-op, in resume and destroy parts, it is replaced with *ret void* instruction and the rest of the block containing *coro.end* instruction is discarded. In landing pads it is replaced with an appropriate instruction to unwind to caller. The handling of *coro.end* differs depending on whether the target is using landingpad or WinEH exception model.

For landingpad based exception model, it is expected that frontend uses the *coro.end* intrinsic as follows:

```
ehcleanup:
    %InResumePart = call i1 @llvm.coro.end(i8* null, i1 true)
    br i1 %InResumePart, label %eh.resume, label %cleanup.cont

cleanup.cont:
    ; rest of the cleanup

eh.resume:
    %exn = load i8*, i8** %exn.slot, align 8
    %sel = load i32, i32* %ehselector.slot, align 4
    %lpad.val = insertvalue { i8*, i32 } undef, i8* %exn, 0
    %lpad.val29 = insertvalue { i8*, i32 } %lpad.val, i32 %sel, 1
    resume { i8*, i32 } %lpad.val29
```

The *CoroSplit* pass replaces *coro.end* with `True` in the resume functions, thus leading to immediate unwind to the caller, whereas in start function it is replaced with `False`, thus allowing to proceed to the rest of the cleanup code that is only needed during initial invocation of the coroutine.

For Windows Exception handling model, a frontend should attach a funclet bundle referring to an enclosing cleanuppad as follows:

```
ehcleanup:
  %tok = cleanuppad within none []
  %unused = call i1 @llvm.coro.end(i8* null, i1 true) [ "funclet" (token %tok) ]
  cleanupret from %tok unwind label %RestOfTheCleanup
```

The *CoroSplit* pass, if the funclet bundle is present, will insert `cleanupret from %tok unwind to caller` before the *coro.end* intrinsic and will remove the rest of the block.

The following table summarizes the handling of *coro.end* intrinsic.

		In Start Function	In Resume/Destroy Functions
unwind=false		nothing	ret void
	WinEH	nothing	cleanupret unwind to caller
	Landingpad	nothing	nothing

'llvm.coro.suspend' Intrinsic

```
declare i8 @llvm.coro.suspend(token <save>, i1 <final>)
```

Overview:

The `'llvm.coro.suspend'` marks the point where execution of the coroutine need to get suspended and control returned back to the caller. Conditional branches consuming the result of this intrinsic lead to basic blocks where coroutine should proceed when suspended (-1), resumed (0) or destroyed (1).

Arguments:

The first argument refers to a token of *coro.save* intrinsic that marks the point when coroutine state is prepared for suspension. If *none* token is passed, the intrinsic behaves as if there were a *coro.save* immediately preceding the *coro.suspend* intrinsic.

The second argument indicates whether this suspension point is *final*. The second argument only accepts constants. If more than one suspend point is designated as final, the resume and destroy branches should lead to the same basic blocks.

Example (normal suspend point):

```
%0 = call i8 @llvm.coro.suspend(token none, i1 false)
switch i8 %0, label %suspend [i8 0, label %resume
                             i8 1, label %cleanup]
```

Example (final suspend point):

```
while.end:
  %s.final = call i8 @llvm.coro.suspend(token none, i1 true)
  switch i8 %s.final, label %suspend [i8 0, label %trap
                                     i8 1, label %cleanup]
trap:
  call void @llvm.trap()
  unreachable
```

Semantics:

If a coroutine that was suspended at the suspend point marked by this intrinsic is resumed via *coro.resume* the control will transfer to the basic block of the 0-case. If it is resumed via *coro.destroy*, it will proceed to the basic block indicated by the 1-case. To suspend, coroutine proceed to the default label.

If suspend intrinsic is marked as final, it can consider the *true* branch unreachable and can perform optimizations that can take advantage of that fact.

‘llvm.coro.save’ Intrinsic

```
declare token @llvm.coro.save(i8* <handle>)
```

Overview:

The ‘*llvm.coro.save*’ marks the point where a coroutine need to update its state to prepare for resumption to be considered suspended (and thus eligible for resumption).

Arguments:

The first argument points to a coroutine handle of the enclosing coroutine.

Semantics:

Whatever coroutine state changes are required to enable resumption of the coroutine from the corresponding suspend point should be done at the point of *coro.save* intrinsic.

Example:

Separate save and suspend points are necessary when a coroutine is used to represent an asynchronous control flow driven by callbacks representing completions of asynchronous operations.

In such a case, a coroutine should be ready for resumption prior to a call to *async_op* function that may trigger resumption of a coroutine from the same or a different thread possibly prior to *async_op* call returning control back to the coroutine:

```
%save1 = call token @llvm.coro.save(i8* %hdl)
call void @async_op1(i8* %hdl)
%suspend1 = call i1 @llvm.coro.suspend(token %save1, i1 false)
switch i8 %suspend1, label %suspend [i8 0, label %resume1
                                     i8 1, label %cleanup]
```

‘llvm.coro.param’ Intrinsic

```
declare i1 @llvm.coro.param(i8* <original>, i8* <copy>)
```

Overview:

The ‘llvm.coro.param’ is used by a frontend to mark up the code used to construct and destruct copies of the parameters. If the optimizer discovers that a particular parameter copy is not used after any suspends, it can remove the construction and destruction of the copy by replacing corresponding coro.param with *il false* and replacing any use of the *copy* with the *original*.

Arguments:

The first argument points to an *alloca* storing the value of a parameter to a coroutine.

The second argument points to an *alloca* storing the value of the copy of that parameter.

Semantics:

The optimizer is free to always replace this intrinsic with *il true*.

The optimizer is also allowed to replace it with *il false* provided that the parameter copy is only used prior to control flow reaching any of the suspend points. The code that would be DCE’d if the *coro.param* is replaced with *il false* is not considered to be a use of the parameter copy.

The frontend can emit this intrinsic if its language rules allow for this optimization.

Example:

Consider the following example. A coroutine takes two parameters *a* and *b* that has a destructor and a move constructor.

```
struct A { ~A(); A(A&&); bool foo(); void bar(); };

task<int> f(A a, A b) {
    if (a.foo())
        return 42;

    a.bar();
    co_await read_async(); // introduces suspend point
    b.bar();
}
```

Note that, uses of *b* is used after a suspend point and thus must be copied into a coroutine frame, whereas *a* does not have to, since it never used after suspend.

A frontend can create parameter copies for *a* and *b* as follows:

```
task<int> f(A a', A b') {  
    a = alloca A;  
    b = alloca A;  
    // move parameters to its copies  
    if (coro.param(a', a)) A::A(a, A&& a');  
    if (coro.param(b', b)) A::A(b, A&& b');  
    ...  
    // destroy parameters copies  
    if (coro.param(a', a)) A::~~A(a);  
    if (coro.param(b', b)) A::~~A(b);  
}
```

The optimizer can replace `coro.param(a',a)` with *il false* and replace all uses of *a* with *a'*, since it is not used after suspend.

The optimizer must replace `coro.param(b', b)` with *il true*, since *b* is used after suspend and therefore, it has to reside in the coroutine frame.

4.34.4 Coroutine Transformation Passes

CoroEarly

The pass CoroEarly lowers coroutine intrinsics that hide the details of the structure of the coroutine frame, but, otherwise not needed to be preserved to help later coroutine passes. This pass lowers *coro.frame*, *coro.done*, and *coro.promise* intrinsics.

CoroSplit

The pass CoroSplit buides coroutine frame and outlines resume and destroy parts into separate functions.

CoroElide

The pass CoroElide examines if the inlined coroutine is eligible for heap allocation elision optimization. If so, it replaces *coro.begin* intrinsic with an address of a coroutine frame placed on its caller and replaces *coro.alloc* and *coro.free* intrinsics with *false* and *null* respectively to remove the deallocation code. This pass also replaces *coro.resume* and *coro.destroy* intrinsics with direct calls to resume and destroy functions for a particular coroutine where possible.

CoroCleanup

This pass runs late to lower all coroutine related intrinsics not replaced by earlier passes.

4.34.5 Areas Requiring Attention

1. A coroutine frame is bigger than it could be. Adding stack packing and stack coloring like optimization on the coroutine frame will result in tighter coroutine frames.

2. Take advantage of the lifetime intrinsics for the data that goes into the coroutine frame. Leave lifetime intrinsics as is for the data that stays in allocas.
3. The CoroElide optimization pass relies on coroutine ramp function to be inlined. It would be beneficial to split the ramp function further to increase the chance that it will get inlined into its caller.
4. Design a convention that would make it possible to apply coroutine heap elision optimization across ABI boundaries.
5. Cannot handle coroutines with *inalloca* parameters (used in x86 on Windows).
6. Alignment is ignored by *coro.begin* and *coro.free* intrinsics.
7. Make required changes to make sure that coroutine optimizations work with LTO.
8. More tests, more tests, more tests

4.35 Global Instruction Selection

- *Introduction*
- *Generic Machine IR*
- *Core Pipeline*
- *Maintainability*
- *Progress and Future Work*
- *Porting GlobalISel to A New Target*

Warning: This document is a work in progress. It reflects the current state of the implementation, as well as open design and implementation issues.

4.35.1 Introduction

GlobalISel is a framework that provides a set of reusable passes and utilities for instruction selection — translation from LLVM IR to target-specific Machine IR (MIR).

GlobalISel is intended to be a replacement for SelectionDAG and FastISel, to solve three major problems:

- **Performance** — SelectionDAG introduces a dedicated intermediate representation, which has a compile-time cost.

GlobalISel directly operates on the post-isel representation used by the rest of the code generator, MIR. It does require extensions to that representation to support arbitrary incoming IR: *Generic Machine IR*.

- **Granularity** — SelectionDAG and FastISel operate on individual basic blocks, losing some global optimization opportunities.

GlobalISel operates on the whole function.

- **Modularity** — SelectionDAG and FastISel are radically different and share very little code.

GlobalISel is built in a way that enables code reuse. For instance, both the optimized and fast selectors share the *Core Pipeline*, and targets can configure that pipeline to better suit their needs.

4.35.2 Generic Machine IR

Machine IR operates on physical registers, register classes, and (mostly) target-specific instructions.

To bridge the gap with LLVM IR, GlobalISel introduces “generic” extensions to Machine IR:

- *Generic Instructions*
- *Generic Virtual Registers*
- *Register Bank*
- *Low Level Type*

NOTE: The generic MIR (GMIR) representation still contains references to IR constructs (such as `GlobalValue`). Removing those should let us write more accurate tests, or delete IR after building the initial MIR. However, it is not part of the GlobalISel effort.

Generic Instructions

The main addition is support for pre-isel generic machine instructions (e.g., `G_ADD`). Like other target-independent instructions (e.g., `COPY` or `PHI`), these are available on all targets.

TODO: While we’re progressively adding instructions, one kind in particular exposes interesting problems: compares and how to represent condition codes. Some targets (x86, ARM) have generic comparisons setting multiple flags, which are then used by predicated variants. Others (IR) specify the predicate in the comparison and users just get a single bit. SelectionDAG uses `SETCC/CONDBR` vs `BR_CC` (and similar for `select`) to represent this.

The `MachineIRBuilder` class wraps the `MachineInstrBuilder` and provides a convenient way to create these generic instructions.

Generic Virtual Registers

Generic instructions operate on a new kind of register: “generic” virtual registers. As opposed to non-generic vregs, they are not assigned a `Register Class`. Instead, generic vregs have a *Low Level Type*, and can be assigned a *Register Bank*.

`MachineRegisterInfo` tracks the same information that it does for non-generic vregs (e.g., use-def chains). Additionally, it also tracks the *Low Level Type* of the register, and, instead of the `TargetRegisterClass`, its *Register Bank*, if any.

For simplicity, most generic instructions only accept generic vregs:

- instead of immediates, they use a gvgreg defined by an instruction materializing the immediate value (see *Constant Lowering*).
- instead of physical register, they use a gvgreg defined by a `COPY`.

NOTE: We started with an alternative representation, where `MRI` tracks a size for each gvgreg, and instructions have lists of types. That had two flaws: the type and size are redundant, and there was no generic way of getting a given operand’s type (as there was no 1:1 mapping between instruction types and operands). We considered putting the type in some variant of `MCInstrDesc` instead: See [PR26576](#): [GlobalISel] Generic MachineInstrs need a type but this increases the memory footprint of the related objects

Register Bank

A Register Bank is a set of register classes defined by the target. A bank has a size, which is the maximum store size of all covered classes.

In general, cross-class copies inside a bank are expected to be cheaper than copies across banks. They are also coalesceable by the register coalescer, whereas cross-bank copies are not.

Also, equivalent operations can be performed on different banks using different instructions.

For example, X86 can be seen as having 3 main banks: general-purpose, x87, and vector (which could be further split into a bank per domain for single vs double precision instructions).

Register banks are described by a target-provided API, [RegisterBankInfo](#).

Low Level Type

Additionally, every generic virtual register has a type, represented by an instance of the `LLT` class.

Like `EVT/MVT/Type`, it has no distinction between unsigned and signed integer types. Furthermore, it also has no distinction between integer and floating-point types: it mainly conveys absolutely necessary information, such as size and number of vector lanes:

- `sN` for scalars
- `pN` for pointers
- `<N x sM>` for vectors
- `unsized` for labels, etc..

`LLT` is intended to replace the usage of `EVT` in `SelectionDAG`.

Here are some `LLT` examples and their `EVT` and `Type` equivalents:

LLT	EVT	IR Type
<code>s1</code>	<code>i1</code>	<code>i1</code>
<code>s8</code>	<code>i8</code>	<code>i8</code>
<code>s32</code>	<code>i32</code>	<code>i32</code>
<code>s32</code>	<code>f32</code>	<code>float</code>
<code>s17</code>	<code>i17</code>	<code>i17</code>
<code>s16</code>	<code>N/A</code>	<code>{i8, i8}</code>
<code>s32</code>	<code>N/A</code>	<code>[4 x i8]</code>
<code>p0</code>	<code>iPTR</code>	<code>i8*, i32*, %opaque*</code>
<code>p2</code>	<code>iPTR</code>	<code>i8 addrspace(2) *</code>
<code><4 x s32></code>	<code>v4f32</code>	<code><4 x float></code>
<code>s64</code>	<code>v1f64</code>	<code><1 x double></code>
<code><3 x s32></code>	<code>v3i32</code>	<code><3 x i32></code>
<code>unsized</code>	<code>Other</code>	<code>label</code>

Rationale: instructions already encode a specific interpretation of types (e.g., `add` vs. `fadd`, or `sdiv` vs. `udiv`). Also encoding that information in the type system requires introducing bitcast with no real advantage for the selector.

Pointer types are distinguished by address space. This matches IR, as opposed to `SelectionDAG` where address space is an attribute on operations. This representation better supports pointers having different sizes depending on their addressspace.

NOTE: Currently, `LLT` requires at least 2 elements in vectors, but some targets have the concept of a ‘1-element vector’. Representing them as their underlying scalar type is a nice simplification.

TODO: Currently, non-generic virtual registers, defined by non-pre-isel-generic instructions, cannot have a type, and thus cannot be used by a pre-isel generic instruction. Instead, they are given a type using a COPY. We could relax that and allow types on all vregs: this would reduce the number of MI required when emitting target-specific MIR early in the pipeline. This should purely be a compile-time optimization.

4.35.3 Core Pipeline

There are four required passes, regardless of the optimization mode:

- *IRTranslator*
 - *API: CallLowering*
 - *Aggregates*
 - *Constant Lowering*
- *Legalizer*
 - *API: LegalizerInfo*
 - *Non-power of 2 types*
 - *Vector types*
- *RegBankSelect*
 - *API: RegisterBankInfo*
 - *RegBankSelect Modes*
- *InstructionSelect*
 - *API: InstructionSelector*

Additional passes can then be inserted at higher optimization levels or for specific targets. For example, to match the current SelectionDAG set of transformations: MachineCSE and a better MachineCombiner between every pass.

NOTE: In theory, not all passes are always necessary. As an additional compile-time optimization, we could skip some of the passes by setting the relevant MachineFunction properties. For instance, if the IRTranslator did not encounter any illegal instruction, it would set the `legalized` property to avoid running the *Legalizer*. Similarly, we considered specializing the IRTranslator per-target to directly emit target-specific MI. However, we instead decided to keep the core pipeline simple, and focus on minimizing the overhead of the passes in the no-op cases.

IRTranslator

This pass translates the input LLVM IR `Function` to a GMIR `MachineFunction`.

TODO: This currently doesn't support the more complex instructions, in particular those involving control flow (`switch`, `invoke`, ...). For `switch` in particular, we can initially use the `LowerSwitch` pass.

API: CallLowering

The IRTranslator (using the CallLowering target-provided utility) also implements the ABI's calling convention by lowering calls, returns, and arguments to the appropriate physical register usage and instruction sequences.

Aggregates

Aggregates are lowered to a single scalar vreg. This differs from SelectionDAG's multiple vregs via `GetValueVTs`.

TODO: As some of the bits are undef (padding), we should consider augmenting the representation with additional metadata (in effect, caching `computeKnownBits` information on vregs). See [PR26161: \[GlobalISel\] Value to vreg](#) during IR to MachineInstr translation for aggregate type

Constant Lowering

The `IRTranslator` lowers `Constant` operands into uses of gvregs defined by `G_CONSTANT` or `G_FCONSTANT` instructions. Currently, these instructions are always emitted in the entry basic block. In a `MachineFunction`, each `Constant` is materialized by a single vreg.

This is beneficial as it allows us to fold constants into immediate operands during *InstructionSelect*, while still avoiding redundant materializations for expensive non-foldable constants. However, this can lead to unnecessary spills and reloads in an -O0 pipeline, as these vregs can have long live ranges.

TODO: We're investigating better placement of these instructions, in fast and optimized modes.

Legalizer

This pass transforms the generic machine instructions such that they are legal.

A legal instruction is defined as:

- **selectable** — the target will later be able to select it to a target-specific (non-generic) instruction.
- operating on **vregs that can be loaded and stored** – if necessary, the target can select a `G_LOAD/G_STORE` of each vreg operand.

As opposed to SelectionDAG, there are no legalization phases. In particular, 'type' and 'operation' legalization are not separate.

Legalization is iterative, and all state is contained in GMIR. To maintain the validity of the intermediate code, instructions are introduced:

- `G_SEQUENCE` — concatenate multiple registers into a single wider register.
- `G_EXTRACT` — extract multiple registers (as contiguous sequences of bits) from a single wider register.

As they are expected to be temporary byproducts of the legalization process, they are combined at the end of the *Legalizer* pass. If any remain, they are expected to always be selectable, using loads and stores if necessary.

API: LegalizerInfo

Currently the API is broadly similar to SelectionDAG/TargetLowering, but extended in two ways:

- The set of available actions is wider, avoiding the currently very overloaded `Expand` (which can cover everything from libcalls to scalarization depending on the node's opcode).
- Since there's no separate type legalization, independently varying types on an instruction can have independent actions. For example a `G_ICMP` has 2 independent types: the result and the inputs; we need to be able to say that comparing 2 s32s is OK, but the s1 result must be dealt with in another way.

As such, the primary key when deciding what to do is the `InstrAspect`, essentially a tuple consisting of (`Opcode`, `TypeIdx`, `Type`) and mapping to a suggested course of action.

An example use might be:

```
// The CPU can't deal with an s1 result, do something about it.
setAction({G_ICMP, 0, s1}, WidenScalar);
// An s32 input (the second type) is fine though.
setAction({G_ICMP, 1, s32}, Legal);
```

TODO: An alternative worth investigating is to generalize the API to represent actions using `std::function` that implements the action, instead of explicit enum tokens (`Legal`, `WidenScalar`, ...).

TODO: Moreover, we could use TableGen to initially infer legality of operation from existing patterns (as any pattern we can select is by definition legal). Expanding that to describe legalization actions is a much larger but potentially useful project.

Non-power of 2 types

TODO: Types which have a size that isn't a power of 2 aren't currently supported. The `setAction` API will probably require changes to support them. Even notionally explicitly specified operations only make suggestions like "Widen" or "Narrow". The eventual type is still unspecified and a search is performed by repeated doubling/halving of the type's size. This is incorrect for types that aren't a power of 2. It's reasonable to expect we could construct an efficient set of side-tables for more general lookups though, encoding a map from the integers (i.e. the size of the current type) to types (the legal size).

Vector types

Vectors first get their element type legalized: `<A x sB>` becomes `<A x sC>` such that at least one operation is legal with `sC`.

This is currently specified by the function `setScalarInVectorAction`, called for example as:

```
setScalarInVectorAction(G_ICMP, s1, WidenScalar);
```

Next the number of elements is chosen so that the entire operation is legal. This aspect is not controllable at the moment, but probably should be (you could imagine disagreements on whether a `<2 x s8>` operation should be scalarized or extended to `<8 x s8>`).

RegBankSelect

This pass constrains the *Generic Virtual Registers* operands of generic instructions to some *Register Bank*.

It iteratively maps instructions to a set of per-operand bank assignment. The possible mappings are determined by the target-provided *RegisterBankInfo*. The mapping is then applied, possibly introducing `COPY` instructions if necessary.

It traverses the `MachineFunction` top down so that all operands are already mapped when analyzing an instruction.

This pass could also remap target-specific instructions when beneficial. In the future, this could replace the `ExeDeps-Fix` pass, as we can directly select the best variant for an instruction that's available on multiple banks.

API: RegisterBankInfo

The `RegisterBankInfo` class describes multiple aspects of register banks.

- **Banks:** `addRegBankCoverage` — which register bank covers each register class.
- **Cross-Bank Copies:** `copyCost` — the cost of a `COPY` from one bank to another.
- **Default Mapping:** `getInstrMapping` — the default bank assignments for a given instruction.

- **Alternative Mapping:** `getInstrAlternativeMapping` — the other possible bank assignments for a given instruction.

TODO: All this information should eventually be static and generated by TableGen, mostly using existing information augmented by bank descriptions.

TODO: `getInstrMapping` is currently separate from `getInstrAlternativeMapping` because the latter is more expensive: as we move to static mapping info, both methods should be free, and we should merge them.

RegBankSelect Modes

`RegBankSelect` currently has two modes:

- **Fast** — For each instruction, pick a target-provided “default” bank assignment. This is the default at -O0.
- **Greedy** — For each instruction, pick the cheapest of several target-provided bank assignment alternatives.

We intend to eventually introduce an additional optimizing mode:

- **Global** — Across multiple instructions, pick the cheapest combination of bank assignments.

NOTE: On AArch64, we are considering using the Greedy mode even at -O0 (or perhaps at backend -O1): because *Low Level Type* doesn’t distinguish floating point from integer scalars, the default assignment for loads and stores is the integer bank, introducing cross-bank copies on most floating point operations.

InstructionSelect

This pass transforms generic machine instructions into equivalent target-specific instructions. It traverses the `MachineFunction` bottom-up, selecting uses before definitions, enabling trivial dead code elimination.

API: InstructionSelector

The target implements the `InstructionSelector` class, containing the target-specific selection logic proper.

The instance is provided by the subtarget, so that it can specialize the selector by subtarget feature (with, e.g., a vector selector overriding parts of a general-purpose common selector). We might also want to parameterize it by `MachineFunction`, to enable selector variants based on function attributes like `optsize`.

The simple API consists of:

```
virtual bool select (MachineInstr &MI)
```

This target-provided method is responsible for mutating (or replacing) a possibly-generic MI into a fully target-specific equivalent. It is also responsible for doing the necessary constraining of gvregs into the appropriate register classes.

The `InstructionSelector` can fold other instructions into the selected MI, by walking the use-def chain of the vreg operands. As `GlobalISel` is Global, this folding can occur across basic blocks.

TODO: Currently, the `Select` pass is implemented with hand-written c++, similar to `FastISel`, rather than backed by tblgen’erated pattern-matching. We intend to eventually reuse `SelectionDAG` patterns.

4.35.4 Maintainability

Iterative Transformations

Passes are split into small, iterative transformations, with all state represented in the MIR.

This differs from SelectionDAG (in particular, the legalizer) using various in-memory side-tables.

MIR Serialization

Generic Machine IR is serializable (see *Machine IR (MIR) Format Reference Manual*). Combined with *Iterative Transformations*, this enables much finer-grained testing, rather than requiring large and fragile IR-to-assembly tests.

The current “stage” in the *Core Pipeline* is represented by a set of `MachineFunctionProperties`:

- `legalized`
- `regBankSelected`
- `selected`

MachineVerifier

The pass approach lets us use the `MachineVerifier` to enforce invariants. For instance, a `regBankSelected` function may not have `gvregs` without a bank.

TODO: The `MachineVerifier` being monolithic, some of the checks we want to do can’t be integrated to it: `GlobalISel` is a separate library, so we can’t directly reference it from `CodeGen`. For instance, legality checks are currently done in `RegBankSelect/InstructionSelect` proper. We could `#ifdef` out the checks, or we could add some sort of verifier API.

4.35.5 Progress and Future Work

The initial goal is to replace `FastISel` on AArch64. The next step will be to replace SelectionDAG as the optimized ISel.

NOTE: While we iterate on `GlobalISel`, we strive to avoid affecting the performance of SelectionDAG, `FastISel`, or the other MIR passes. For instance, the types of *Generic Virtual Registers* are stored in a separate table in `MachineRegisterInfo`, that is destroyed after *InstructionSelect*.

FastISel Replacement

For the initial `FastISel` replacement, we intend to fallback to SelectionDAG on selection failures.

Currently, compile-time of the fast pipeline is within 1.5x of `FastISel`. We’re optimistic we can get to within 1.1/1.2x, but beating `FastISel` will be challenging given the multi-pass approach. Still, supporting all IR (via a complete legalizer) and avoiding the fallback to SelectionDAG in the worst case should enable better amortized performance than SelectionDAG+`FastISel`.

NOTE: We considered never having a fallback to SelectionDAG, instead deciding early whether a given function is supported by `GlobalISel` or not. The decision would be based on *Legalizer* queries. We abandoned that for two reasons: a) on IR inputs, we’d need to basically simulate the *IRTranslator*; b) to be robust against unforeseen failures and to enable iterative improvements.

Support For Other Targets

In parallel, we’re investigating adding support for other - ideally quite different - targets. For instance, there is some initial AMDGPU support.

4.35.6 Porting GlobalISel to A New Target

There are four major classes to implement by the target:

- *CallLowering* — lower calls, returns, and arguments according to the ABI.
- *RegisterBankInfo* — describe *Register Bank* coverage, cross-bank copy cost, and the mapping of operands onto banks for each instruction.
- *LegalizerInfo* — describe what is legal, and how to legalize what isn't.
- *InstructionSelector* — select generic MIR to target-specific MIR.

Additionally:

- *TargetPassConfig* — create the passes constituting the pipeline, including additional passes not included in the *Core Pipeline*.
- *GISelAccessor* — setup the various subtarget-provided classes, with a graceful fallback to no-op when GlobalISel isn't enabled.

4.36 XRay Instrumentation

Version 1 as of 2016-11-08

- *Introduction*
- *XRay in LLVM*
- *Using XRay*
 - *Instrumenting your C/C++/Objective-C Application*
 - *LLVM Function Attribute*
 - *XRay Runtime Library*
 - *Flight Data Recorder Mode*
 - *Trace Analysis Tools*
- *Future Work*
 - *Trace Analysis*
 - *More Platforms*

4.36.1 Introduction

XRay is a function call tracing system which combines compiler-inserted instrumentation points and a runtime library that can dynamically enable and disable the instrumentation.

More high level information about XRay can be found in the [XRay whitepaper](#).

This document describes how to use XRay as implemented in LLVM.

4.36.2 XRay in LLVM

XRay consists of three main parts:

- Compiler-inserted instrumentation points.
- A runtime library for enabling/disabling tracing at runtime.
- A suite of tools for analysing the traces.

NOTE: As of February 27, 2017, XRay is only available for the following architectures running Linux: x86_64, arm7 (no thumb), aarch64, powerpc64le, mips, mipsel, mips64, mips64el.

The compiler-inserted instrumentation points come in the form of nop-sleds in the final generated binary, and an ELF section named `xray_instr_map` which contains entries pointing to these instrumentation points. The runtime library relies on being able to access the entries of the `xray_instr_map`, and overwrite the instrumentation points at runtime.

4.36.3 Using XRay

You can use XRay in a couple of ways:

- Instrumenting your C/C++/Objective-C/Objective-C++ application.
- Generating LLVM IR with the correct function attributes.

The rest of this section covers these main ways and later on how to customise what XRay does in an XRay-instrumented binary.

Instrumenting your C/C++/Objective-C Application

The easiest way of getting XRay instrumentation for your application is by enabling the `-fxray-instrument` flag in your clang invocation.

For example:

```
clang -fxray-instrument ..
```

By default, functions that have at least 200 instructions will get XRay instrumentation points. You can tweak that number through the `-fxray-instruction-threshold=` flag:

```
clang -fxray-instrument -fxray-instruction-threshold=1 ..
```

You can also specifically instrument functions in your binary to either always or never be instrumented using source-level attributes. You can do it using the GCC-style attributes or C++11-style attributes.

```
[[clang::xray_always_instrument]] void always_instrumented();  
[[clang::xray_never_instrument]] void never_instrumented();  
  
void alt_always_instrumented() __attribute__((xray_always_instrument));  
void alt_never_instrumented() __attribute__((xray_never_instrument));
```

When linking a binary, you can either manually link in the *XRay Runtime Library* or use `clang` to link it in automatically with the `-fxray-instrument` flag. Alternatively, you can statically link-in the XRay runtime library from `compiler-rt` – those archive files will take the name of `libclang_rt.xray-{arch}` where `{arch}` is the mnemonic supported by clang (x86_64, arm7, etc.).

LLVM Function Attribute

If you're using LLVM IR directly, you can add the `function-instrument` string attribute to your functions, to get the similar effect that the C/C++/Objective-C source-level attributes would get:

```
define i32 @always_instrument() uwtable "function-instrument"="xray-always" {
    ; ...
}

define i32 @never_instrument() uwtable "function-instrument"="xray-never" {
    ; ...
}
```

You can also set the `xray-instruction-threshold` attribute and provide a numeric string value for how many instructions should be in the function before it gets instrumented.

```
define i32 @maybe_instrument() uwtable "xray-instruction-threshold"="2" {
    ; ...
}
```

XRay Runtime Library

The XRay Runtime Library is part of the `compiler-rt` project, which implements the runtime components that perform the patching and unpatching of inserted instrumentation points. When you use `clang` to link your binaries and the `-fxray-instrument` flag, it will automatically link in the XRay runtime.

The default implementation of the XRay runtime will enable XRay instrumentation before `main` starts, which works for applications that have a short lifetime. This implementation also records all function entry and exit events which may result in a lot of records in the resulting trace.

Also by default the filename of the XRay trace is `xray-log.XXXXXXX` where the `XXXXXX` part is randomly generated.

These options can be controlled through the `XRAY_OPTIONS` environment variable, where we list down the options and their defaults below.

Option	Type	Default	Description
<code>patch_premain</code>	<code>bool</code>	<code>false</code>	Whether to patch instrumentation points before <code>main</code> .
<code>xray_naive_log</code>	<code>bool</code>	<code>true</code>	Whether to install the naive log implementation.
<code>xray_logfile_base</code>	<code>const char*</code>	<code>xray-log.</code>	Filename base for the XRay logfile.
<code>xray_fdr_log</code>	<code>bool</code>	<code>false</code>	Whether to install the Flight Data Recorder (FDR) mode.

If you choose to not use the default logging implementation that comes with the XRay runtime and/or control when/how the XRay instrumentation runs, you may use the XRay APIs directly for doing so. To do this, you'll need to include the `xray_interface.h` from the `compiler-rt/xray` directory. The important API functions we list below:

- `__xray_set_handler(void (*entry)(int32_t, XRayEntryType))`: Install your own logging handler for when an event is encountered. See `xray/xray_interface.h` for more details.
- `__xray_remove_handler()`: Removes whatever the installed handler is.
- `__xray_patch()`: Patch all the instrumentation points defined in the binary.
- `__xray_unpatch()`: Unpatch the instrumentation points defined in the binary.

There are some requirements on the logging handler to be installed for the thread-safety of operations to be performed by the XRay runtime library:

- The function should be thread-safe, as multiple threads may be invoking the function at the same time. If the logging function needs to do synchronisation, it must do so internally as XRay does not provide any synchronisation guarantees outside from the atomicity of updates to the pointer.
- The pointer provided to `__xray_set_handler(...)` must be live even after calls to `__xray_remove_handler()` and `__xray_unpatch()` have succeeded. XRay cannot guarantee that all threads that have ever gotten a copy of the pointer will not invoke the function.

Flight Data Recorder Mode

XRay supports a logging mode which allows the application to only capture a fixed amount of memory's worth of events. Flight Data Recorder (FDR) mode works very much like a plane's "black box" which keeps recording data to memory in a fixed-size circular queue of buffers, and have the data available programmatically until the buffers are finalized and flushed. To use FDR mode on your application, you may set the `xray_fdr_log` option to `true` in the `XRAY_OPTIONS` environment variable (while also optionally setting the `xray_naive_log` to `false`).

When FDR mode is on, it will keep writing and recycling memory buffers until the logging implementation is finalized – at which point it can be flushed and re-initialised later. To do this programmatically, we follow the workflow provided below:

```
// Patch the sleds, if we haven't yet.
auto patch_status = __xray_patch();

// Maybe handle the patch_status errors.

// When we want to flush the log, we need to finalize it first, to give
// threads a chance to return buffers to the queue.
auto finalize_status = __xray_log_finalize();
if (finalize_status != XRAY_LOG_FINALIZED) {
    // maybe retry, or bail out.
}

// At this point, we are sure that the log is finalized, so we may try
// flushing the log.
auto flush_status = __xray_log_flushLog();
if (flush_status != XRAY_LOG_FLUSHED) {
    // maybe retry, or bail out.
}
```

The default settings for the FDR mode implementation will create logs named similarly to the naive log implementation, but will have a different log format. All the trace analysis tools (and the trace reading library) will support all versions of the FDR mode format as we add more functionality and record types in the future.

NOTE: We do not however promise perpetual support for when we update the log versions we support going forward. Deprecation of the formats will be announced and discussed on the developers mailing list.

XRay allows for replacing the default FDR mode logging implementation using the following API:

- `__xray_set_log_impl(...)`: This function takes a struct of type `XRayLogImpl`, which is defined in `xray/xray_log_interface.h`, part of the XRay compiler-rt installation.
- `__xray_log_init(...)`: This function allows for initializing and re-initializing an installed logging implementation. See `xray/xray_log_interface.h` for details, part of the XRay compiler-rt installation.

Trace Analysis Tools

We currently have the beginnings of a trace analysis tool in LLVM, which can be found in the `tools/llvm-xray` directory. The `llvm-xray` tool currently supports the following subcommands:

- `extract`: Extract the instrumentation map from a binary, and return it as YAML.
- `account`: Performs basic function call accounting statistics with various options for sorting, and output formats (supports CSV, YAML, and console-friendly TEXT).
- `convert`: Converts an XRay log file from one format to another. Currently only converts to YAML.
- `graph`: Generates a DOT graph of the function call relationships between functions found in an XRay trace.

These subcommands use various library components found as part of the XRay libraries, distributed with the LLVM distribution. These are:

- `llvm/XRay/Trace.h`: A trace reading library for conveniently loading an XRay trace of supported forms, into a convenient in-memory representation. All the analysis tools that deal with traces use this implementation.
- `llvm/XRay/Graph.h`: A semi-generic graph type used by the `graph` subcommand to conveniently represent a function call graph with statistics associated with edges and vertices.
- `llvm/XRay/InstrumentationMap.h`: A convenient tool for analyzing the instrumentation map in XRay-instrumented object files and binaries. The `extract` subcommand uses this particular library.

4.36.4 Future Work

There are a number of ongoing efforts for expanding the toolset building around the XRay instrumentation system.

Trace Analysis

We have more subcommands and modes that we're thinking of developing, in the following forms:

- `stack`: Reconstruct the function call stacks in a timeline.

More Platforms

We're looking forward to contributions to port XRay to more architectures and operating systems.

4.37 Debugging with XRay

This document shows an example of how you would go about analyzing applications built with XRay instrumentation. Here we will attempt to debug `libc` compiling some sample LLVM IR generated by Clang.

- *Building with XRay*
- *Getting Traces*
- *The `llvm-xray` Tool*
- *Controlling Fidelity*
 - *Instruction Threshold*

– *Instrumentation Attributes*

- *Further Exploration*
- *Next Steps*

4.37.1 Building with XRay

To debug an application with XRay instrumentation, we need to build it with a Clang that supports the `-fxray-instrument` option. See [XRay](#) for more technical details of how XRay works for background information.

In our example, we need to add `-fxray-instrument` to the list of flags passed to Clang when building a binary. Note that we need to link with Clang as well to get the XRay runtime linked in appropriately. For building `llc` with XRay, we do something similar below for our LLVM build:

```
$ mkdir -p llvm-build && cd llvm-build
# Assume that the LLVM sources are at ../llvm
$ cmake -GNinja ../llvm -DCMAKE_BUILD_TYPE=Release \
    -DCMAKE_C_FLAGS_RELEASE="-fxray-instrument" -DCMAKE_CXX_FLAGS="-fxray-instrument" \
    ↪ \
# Once this finishes, we should build llc
$ ninja llc
```

To verify that we have an XRay instrumented binary, we can use `objdump` to look for the `xray_instr_map` section.

```
$ objdump -h -j xray_instr_map ./bin/llc
./bin/llc:      file format elf64-x86-64

Sections:
Idx Name          Size      VMA           LMA             File off  Algn
 14 xray_instr_map 00002fc0  00000000041516c6  00000000041516c6  03d516c6  2**0
                CONTENTS, ALLOC, LOAD, READONLY, DATA
```

4.37.2 Getting Traces

By default, XRay does not write out the trace files or patch the application before main starts. If we just run `llc` it should just work like a normally built binary. However, if we want to get a full trace of the application's operations (of the functions we do end up instrumenting with XRay) then we need to enable XRay at application start. To do this, XRay checks the `XRAY_OPTIONS` environment variable.

```
# The following doesn't create an XRay trace by default.
$ ./bin/llc input.ll

# We need to set the XRAY_OPTIONS to enable some features.
$ XRAY_OPTIONS="patch_premain=true" ./bin/llc input.ll
==69819==XRay: Log file in 'xray-log.llc.m35qPB'
```

At this point we now have an XRay trace we can start analysing.

4.37.3 The `llvm-xray` Tool

Having a trace then allows us to do basic accounting of the functions that were instrumented, and how much time we're spending in parts of the code. To make sense of this data, we use the `llvm-xray` tool which has a few subcommands to help us understand our trace.

One of the simplest things we can do is to get an accounting of the functions that have been instrumented. We can see an example accounting with `llvm-xray account`:

```
$ llvm-xray account xray-log.llc.m35qPB -top=10 -sort=sum -sortorder=dsc -instr_map ./
↳ bin/llc
Functions with latencies: 29
  funcid      count [      min,          med,          90p,          99p,          max]
↳ sum function
  187          360 [ 0.000000, 0.000001, 0.000014, 0.000032, 0.000075] 0.
↳ 001596 LLVM.cpp:446:0: llvm::LLLexer::LexIdentifier()
  85           130 [ 0.000000, 0.000000, 0.000018, 0.000023, 0.000156] 0.
↳ 000799 X86ISelDAGToDAG.cpp:1984:0: (anonymous
↳ namespace)::X86DAGToDAGISel::Select(llvm::SDNode*)
  138           130 [ 0.000000, 0.000000, 0.000017, 0.000155, 0.000155] 0.
↳ 000774 SelectionDAGISel.cpp:2963:0:
↳ llvm::SelectionDAGISel::SelectCodeCommon(llvm::SDNode*, unsigned char const*,
↳ unsigned int)
  188           103 [ 0.000000, 0.000000, 0.000003, 0.000123, 0.000214] 0.
↳ 000737 LLVM.cpp:2692:0: llvm::LLParser::ParseValID(llvm::ValID&,
↳ llvm::LLParser::PerFunctionState*)
  88            1 [ 0.000562, 0.000562, 0.000562, 0.000562, 0.000562] 0.
↳ 000562 X86ISelLowering.cpp:83:0:
↳ llvm::X86TargetLowering::X86TargetLowering(llvm::X86TargetMachine const&,
↳ llvm::X86Subtarget const&)
  125           102 [ 0.000001, 0.000003, 0.000010, 0.000017, 0.000049] 0.
↳ 000471 Verifier.cpp:3714:0: (anonymous
↳ namespace)::Verifier::visitInstruction(llvm::Instruction&)
  90            8 [ 0.000023, 0.000035, 0.000106, 0.000106, 0.000106] 0.
↳ 000342 X86ISelLowering.cpp:3363:0:
↳ llvm::X86TargetLowering::LowerCall(llvm::TargetLowering::CallLoweringInfo&,
↳ llvm::SmallVectorImpl<llvm::SDValue>&) const
  124           32 [ 0.000003, 0.000007, 0.000016, 0.000041, 0.000041] 0.
↳ 000310 Verifier.cpp:1967:0: (anonymous
↳ namespace)::Verifier::visitFunction(llvm::Function const&)
  123            1 [ 0.000302, 0.000302, 0.000302, 0.000302, 0.000302] 0.
↳ 000302 LLVMContextImpl.cpp:54:0: llvm::LLVMContextImpl::~~LLVMContextImpl()
  139           46 [ 0.000000, 0.000002, 0.000006, 0.000008, 0.000019] 0.
↳ 000138 TargetLowering.cpp:506:0:
↳ llvm::TargetLowering::SimplifyDemandedBits(llvm::SDValue, llvm::APInt const&,
↳ llvm::APInt&, llvm::APInt&, llvm::TargetLowering::TargetLoweringOpt&, unsigned int,
↳ bool) const
```

This shows us that for our input file, `llc` spent the most cumulative time in the lexer (a total of 1 millisecond). If we wanted for example to work with this data in a spreadsheet, we can output the results as CSV using the `-format=csv` option to the command for further analysis.

If we want to get a textual representation of the raw trace we can use the `llvm-xray convert` tool to get YAML output. The first few lines of that output for an example trace would look like the following:

```
$ llvm-xray convert -f yaml -symbolize -instr_map=./bin/llc xray-log.llc.m35qPB
---
header:
  version: 1
```

```

type:          0
constant-tsc:   true
nonstop-tsc:    true
cycle-frequency: 2601000000
records:
- { type: 0, func-id: 110, function: __cxx_global_var_init.8, cpu: 37, thread: 69819, kind: function-enter, tsc: 5434426023268520 }
- { type: 0, func-id: 110, function: __cxx_global_var_init.8, cpu: 37, thread: 69819, kind: function-exit, tsc: 5434426023523052 }
- { type: 0, func-id: 164, function: __cxx_global_var_init, cpu: 37, thread: 69819, kind: function-enter, tsc: 5434426029925386 }
- { type: 0, func-id: 164, function: __cxx_global_var_init, cpu: 37, thread: 69819, kind: function-exit, tsc: 5434426030031128 }
- { type: 0, func-id: 142, function: '(anonymous namespace)::CommandLineParser::ParseCommandLineOptions(int, char const* const*, llvm::StringRef, llvm::raw_ostream*)', cpu: 37, thread: 69819, kind: function-enter, tsc: 5434426046951388 }
- { type: 0, func-id: 142, function: '(anonymous namespace)::CommandLineParser::ParseCommandLineOptions(int, char const* const*, llvm::StringRef, llvm::raw_ostream*)', cpu: 37, thread: 69819, kind: function-exit, tsc: 5434426047282020 }
- { type: 0, func-id: 187, function: 'llvm::LLLexer::LexIdentifier()', cpu: 37, thread: 69819, kind: function-enter, tsc: 5434426047857332 }
- { type: 0, func-id: 187, function: 'llvm::LLLexer::LexIdentifier()', cpu: 37, thread: 69819, kind: function-exit, tsc: 5434426047984152 }
- { type: 0, func-id: 187, function: 'llvm::LLLexer::LexIdentifier()', cpu: 37, thread: 69819, kind: function-enter, tsc: 5434426048036584 }
- { type: 0, func-id: 187, function: 'llvm::LLLexer::LexIdentifier()', cpu: 37, thread: 69819, kind: function-exit, tsc: 5434426048042292 }
- { type: 0, func-id: 187, function: 'llvm::LLLexer::LexIdentifier()', cpu: 37, thread: 69819, kind: function-enter, tsc: 5434426048055056 }
- { type: 0, func-id: 187, function: 'llvm::LLLexer::LexIdentifier()', cpu: 37, thread: 69819, kind: function-exit, tsc: 5434426048067316 }

```

4.37.4 Controlling Fidelity

So far in our examples, we haven't been getting full coverage of the functions we have in the binary. To get that, we need to modify the compiler flags so that we can instrument more (if not all) the functions we have in the binary. We have two options for doing that, and we explore both of these below.

Instruction Threshold

The first “blunt” way of doing this is by setting the minimum threshold for function bodies to 1. We can do that with the `-fxray-instruction-threshold=N` flag when building our binary. We rebuild `llc` with this option and observe the results:

```

$ rm CMakeCache.txt
$ cmake -GNinja ../llvm -DCMAKE_BUILD_TYPE=Release \
  -DCMAKE_C_FLAGS_RELEASE="-fxray-instrument -fxray-instruction-threshold=1" \
  -DCMAKE_CXX_FLAGS="-fxray-instrument -fxray-instruction-threshold=1"
$ ninja llc
$ XRAY_OPTIONS="patch_premain=true" ./bin/llc input.ll
==69819==XRay: Log file in 'xray-log.llc.5rqxkU'

$ llvm-xray account xray-log.llc.5rqxkU -top=10 -sort=sum -sortorder=dsc -instr_map ./
bin/llc

```


Functions with latencies: 36652								
funcid	count	[min,	med,	90p,	99p,	max]	sum
↪function								
75	1	[0.672368,	0.672368,	0.672368,	0.672368,	0.672368]	0.672368
↪llc.cpp:271:0: main								
78	1	[0.626455,	0.626455,	0.626455,	0.626455,	0.626455]	0.626455
↪llc.cpp:381:0: compileModule(char**, llvm::LLVMContext&)								
139617	1	[0.472618,	0.472618,	0.472618,	0.472618,	0.472618]	0.472618
↪LegacyPassManager.cpp:1723:0: llvm::legacy::PassManager::run(llvm::Module&)								
139610	1	[0.472618,	0.472618,	0.472618,	0.472618,	0.472618]	0.472618
↪LegacyPassManager.cpp:1681:0: llvm::legacy::PassManagerImpl::run(llvm::Module&)								
139612	1	[0.470948,	0.470948,	0.470948,	0.470948,	0.470948]	0.470948
↪LegacyPassManager.cpp:1564:0: (anonymous								
↪namespace)::MPPassManager::runOnModule(llvm::Module&)								
139607	2	[0.147345,	0.315994,	0.315994,	0.315994,	0.315994]	0.463340
↪LegacyPassManager.cpp:1530:0: llvm::FPPassManager::runOnModule(llvm::Module&)								
139605	21	[0.000002,	0.000002,	0.102593,	0.213336,	0.213336]	0.463331
↪LegacyPassManager.cpp:1491:0: llvm::FPPassManager::runOnFunction(llvm::Function&)								
139563	26096	[0.000002,	0.000002,	0.000037,	0.000063,	0.000215]	0.225708
↪LegacyPassManager.cpp:1083:0: llvm::PMDDataManager::findAnalysisPass(void const*,								
↪bool)								
108055	188	[0.000002,	0.000120,	0.001375,	0.004523,	0.062624]	0.159279
↪MachineFunctionPass.cpp:38:0:								
↪llvm::MachineFunctionPass::runOnFunction(llvm::Function&)								
62635	22	[0.000041,	0.000046,	0.000050,	0.126744,	0.126744]	0.127715
↪X86TargetMachine.cpp:242:0: llvm::X86TargetMachine::getSubtargetImpl(llvm::Function								
↪const&) const								

Instrumentation Attributes

The other way is to use configuration files for selecting which functions should always be instrumented by the compiler. This gives us a way of ensuring that certain functions are either always or never instrumented by not having to add the attribute to the source.

To use this feature, you can define one file for the functions to always instrument, and another for functions to never instrument. The format of these files are exactly the same as the SanitizerLists files that control similar things for the sanitizer implementations. For example, we can have two different files like below:

```
# always-instrument.txt
# always instrument functions that match the following filters:
fun:main

# never-instrument.txt
# never instrument functions that match the following filters:
fun:__cxx_*
```

Given the above two files we can re-build by providing those two files as arguments to clang as `-fxray-always-instrument=always-instrument.txt` or `-fxray-never-instrument=never-instrument.txt`.

4.37.5 Further Exploration

The `llvm-xray` tool has a few other subcommands that are in various stages of being developed. One interesting subcommand that can highlight a few interesting things is the `graph` subcommand. Given for example the following

toy program that we build with XRay instrumentation, we can see how the generated graph may be a helpful indicator of where time is being spent for the application.

```
// sample.cc
#include <iostream>
#include <thread>

[[clang::xray_always_instrument]] void f() {
    std::cerr << '.';
}

[[clang::xray_always_instrument]] void g() {
    for (int i = 0; i < 10; ++i) {
        std::cerr << '-';
    }
}

int main(int argc, char* argv[]) {
    std::thread t1([] {
        for (int i = 0; i < 10; ++i)
            f();
    });
    std::thread t2([] {
        g();
    });
    t1.join();
    t2.join();
    std::cerr << '\n';
}
```

We then build the above with XRay instrumentation:

```
$ clang++ -o sample -O3 sample.cc -std=c++11 -fxray-instrument -fxray-instruction-
→threshold=1
$ XRAY_OPTIONS="patch_premain=true" ./sample
```

We can then explore the graph rendering of the trace generated by this sample application. We assume you have the graphviz tools available in your system, including both `unflatten` and `dot`. If you prefer rendering or exploring the graph using another tool, then that should be feasible as well. `llvm-xray graph` will create DOT format graphs which should be usable in most graph rendering applications. One example invocation of the `llvm-xray graph` command should yield some interesting insights to the workings of C++ applications:

```
$ llvm-xray graph xray-log.sample.* -m sample -color-edges=sum -edge-label=sum \
  | unflatten -f -l10 | dot -Tsvg -o sample.svg
```

4.37.6 Next Steps

If you have some interesting analyses you'd like to implement as part of the `llvm-xray` tool, please feel free to propose them on the `llvm-dev@` mailing list. The following are some ideas to inspire you in getting involved and potentially making things better.

- Implement a query/filtering library that allows for finding patterns in the XRay traces.
- A conversion from the XRay trace onto something that can be visualised better by other tools (like the Chrome trace viewer for example).
- Collecting function call stacks and how often they're encountered in the XRay trace.

4.38 The PDB File Format

- *Introduction*
- *File Layout*
 - *The MSF Container*
 - *Streams*
- *CodeView*

4.38.1 Introduction

PDB (Program Database) is a file format invented by Microsoft and which contains debug information that can be consumed by debuggers and other tools. Since officially supported APIs exist on Windows for querying debug information from PDBs even without the user understanding the internals of the file format, a large ecosystem of tools has been built for Windows to consume this format. In order for Clang to be able to generate programs that can interoperate with these tools, it is necessary for us to generate PDB files ourselves.

At the same time, LLVM has a long history of being able to cross-compile from any platform to any platform, and we wish for the same to be true here. So it is necessary for us to understand the PDB file format at the byte-level so that we can generate PDB files entirely on our own.

This manual describes what we know about the PDB file format today. The layout of the file, the various streams contained within, the format of individual records within, and more.

We would like to extend our heartfelt gratitude to Microsoft, without whom we would not be where we are today. Much of the knowledge contained within this manual was learned through reading code published by Microsoft on their [GitHub repo](#).

4.38.2 File Layout

Important: Unless otherwise specified, all numeric values are encoded in little endian. If you see a type such as `uint16_t` or `uint64_t` going forward, always assume it is little endian!

The MSF File Format

- *The Superblock*
- *The Stream Directory*
- *Alignment and Block Boundaries*

The Superblock

At file offset 0 in an MSF file is the MSF *SuperBlock*, which is laid out as follows:

```
struct SuperBlock {
    char FileMagic[sizeof(Magic)];
    ulittle32_t BlockSize;
    ulittle32_t FreeBlockMapBlock;
    ulittle32_t NumBlocks;
    ulittle32_t NumDirectoryBytes;
    ulittle32_t Unknown;
    ulittle32_t BlockMapAddr;
};
```

- **FileMagic** - Must be equal to "Microsoft C / C++ MSF 7.00\\r\\n" followed by the bytes 1A 44 53 00 00 00.
- **BlockSize** - The block size of the internal file system. Valid values are 512, 1024, 2048, and 4096 bytes. Certain aspects of the MSF file layout vary depending on the block sizes. For the purposes of LLVM, we handle only block sizes of 4KiB, and all further discussion assumes a block size of 4KiB.
- **FreeBlockMapBlock** - The index of a block within the file, at which begins a bitfield representing the set of all blocks within the file which are “free” (i.e. the data within that block is not used). This bitfield is spread across the MSF file at `BlockSize` intervals. **Important:** `FreeBlockMapBlock` can only be 1 or 2! This field is designed to support incremental and atomic updates of the underlying MSF file. While writing to an MSF file, if the value of this field is 1, you can write your new modified bitfield to page 2, and vice versa. Only when you commit the file to disk do you need to swap the value in the `SuperBlock` to point to the new `FreeBlockMapBlock`.
- **NumBlocks** - The total number of blocks in the file. `NumBlocks * BlockSize` should equal the size of the file on disk.
- **NumDirectoryBytes** - The size of the stream directory, in bytes. The stream directory contains information about each stream’s size and the set of blocks that it occupies. It will be described in more detail later.
- **BlockMapAddr** - The index of a block within the MSF file. At this block is an array of `ulittle32_t`’s listing the blocks that the stream directory resides on. For large MSF files, the stream directory (which describes the block layout of each stream) may not fit entirely on a single block. As a result, this extra layer of indirection is introduced, whereby this block contains the list of blocks that the stream directory occupies, and the stream directory itself can be stitched together accordingly. The number of `ulittle32_t`’s in this array is given by `ceil(NumDirectoryBytes / BlockSize)`.

The Stream Directory

The Stream Directory is the root of all access to the other streams in an MSF file. Beginning at byte 0 of the stream directory is the following structure:

```
struct StreamDirectory {
    ulittle32_t NumStreams;
    ulittle32_t StreamSizes[NumStreams];
    ulittle32_t StreamBlocks[NumStreams][];
};
```

And this structure occupies exactly `SuperBlock->NumDirectoryBytes` bytes. Note that each of the last two arrays is of variable length, and in particular that the second array is jagged.

Example: Suppose a hypothetical PDB file with a 4KiB block size, and 4 streams of lengths {1000 bytes, 8000 bytes, 16000 bytes, 9000 bytes}.

Stream 0: `ceil(1000 / 4096) = 1` block

Stream 1: `ceil(8000 / 4096) = 2` blocks

Stream 2: $\text{ceil}(16000 / 4096) = 4$ blocks

Stream 3: $\text{ceil}(9000 / 4096) = 3$ blocks

In total, 10 blocks are used. Let's see what the stream directory might look like:

```
struct StreamDirectory {
    ulittle32_t NumStreams = 4;
    ulittle32_t StreamSizes[] = {1000, 8000, 16000, 9000};
    ulittle32_t StreamBlocks[][] = {
        {4},
        {5, 6},
        {11, 9, 7, 8},
        {10, 15, 12}
    };
};
```

In total, this occupies $15 * 4 = 60$ bytes, so `SuperBlock->NumDirectoryBytes` would equal 60, and `SuperBlock->BlockMapAddr` would be an array of one `ulittle32_t`, since $60 \leq \text{SuperBlock->BlockSize}$.

Note also that the streams are discontinuous, and that part of stream 3 is in the middle of part of stream 2. You cannot assume anything about the layout of the blocks!

Alignment and Block Boundaries

As may be clear by now, it is possible for a single field (whether it be a high level record, a long string field, or even a single `uint16`) to begin and end in separate blocks. For example, if the block size is 4096 bytes, and a `uint16` field begins at the last byte of the current block, then it would need to end on the first byte of the next block. Since blocks are not necessarily contiguously laid out in the file, this means that both the consumer and the producer of an MSF file must be prepared to split data apart accordingly. In the aforementioned example, the high byte of the `uint16` would be written to the last byte of block N, and the low byte would be written to the first byte of block N+1, which could be tens of thousands of bytes later (or even earlier!) in the file, depending on what the stream directory says.

The PDB Info Stream (aka the PDB Stream)

- *Stream Header*
- *Matching a PDB to its executable*

Stream Header

At offset 0 of the PDB Stream is a header with the following layout:

```
struct PdbStreamHeader {
    ulittle32_t Version;
    ulittle32_t Signature;
    ulittle32_t Age;
    Guid UniqueId;
};
```

- **Version** - A Value from the following enum:

```
enum class PdbStreamVersion : uint32_t {  
    VC2 = 19941610,  
    VC4 = 19950623,  
    VC41 = 19950814,  
    VC50 = 19960307,  
    VC98 = 19970604,  
    VC70Dep = 19990604,  
    VC70 = 20000404,  
    VC80 = 20030901,  
    VC110 = 20091201,  
    VC140 = 20140508,  
};
```

While the meaning of this field appears to be obvious, in practice we have never observed a value other than `VC70`, even with modern versions of the toolchain, and it is unclear why the other values exist. It is assumed that certain aspects of the PDB stream's layout, and perhaps even that of the other streams, will change if the value is something other than `VC70`.

- **Signature** - A 32-bit time-stamp generated with a call to `time()` at the time the PDB file is written. Note that due to the inherent uniqueness problems of using a timestamp with 1-second granularity, this field does not really serve its intended purpose, and as such is typically ignored in favor of the `Guid` field, described below.
- **Age** - The number of times the PDB file has been written. This can be used along with `Guid` to match the PDB to its corresponding executable.
- **Guid** - A 128-bit identifier guaranteed to be unique across space and time. In general, this can be thought of as the result of calling the Win32 API `UuidCreate`, although LLVM cannot rely on that, as it must work on non-Windows platforms.

Matching a PDB to its executable

The linker is responsible for writing both the PDB and the final executable, and as a result is the only entity capable of writing the information necessary to match the PDB to the executable.

In order to accomplish this, the linker generates a guid for the PDB (or re-uses the existing guid if it is linking incrementally) and increments the Age field.

The executable is a PE/COFF file, and part of a PE/COFF file is the presence of number of “directories”. For our purposes here, we are interested in the “debug directory”. The exact format of a debug directory is described by the `IMAGE_DEBUG_DIRECTORY` structure. For this particular case, the linker emits a debug directory of type `IMAGE_DEBUG_TYPE_CODEVIEW`. The format of this record is defined in `llvm/DebugInfo/CodeView/CVDebugRecord.h`, but it suffices to say here only that it includes the same `Guid` and `Age` fields. At runtime, a debugger or tool can scan the COFF executable image for the presence of a debug directory of the correct type and verify that the `Guid` and `Age` match.

The PDB TPI Stream

The PDB DBI (Debug Info) Stream

- *Introduction*
- *Stream Header*
- *Substreams*

- *Module Info Substream*
- *Section Contribution Substream*
- *Section Map Substream*
- *File Info Substream*
- *Type Server Substream*
- *EC Substream*
- *Optional Debug Header Stream*

Introduction

The PDB DBI Stream (Index 3) is one of the largest and most important streams in a PDB file. It contains information about how the program was compiled, (e.g. compilation flags, etc), the compilands (e.g. object files) that were used to link together the program, the source files which were used to build the program, as well as references to other streams that contain more detailed information about each compiland, such as the CodeView symbol records contained within each compiland and the source and line information for functions and other symbols within each compiland.

Stream Header

At offset 0 of the DBI Stream is a header with the following layout:

```
struct DbiStreamHeader {
    int32_t VersionSignature;
    uint32_t VersionHeader;
    uint32_t Age;
    uint16_t GlobalStreamIndex;
    uint16_t BuildNumber;
    uint16_t PublicStreamIndex;
    uint16_t PdbDllVersion;
    uint16_t SymRecordStream;
    uint16_t PdbDllRbld;
    int32_t ModInfoSize;
    int32_t SectionContributionSize;
    int32_t SectionMapSize;
    int32_t SourceInfoSize;
    int32_t TypeServerSize;
    uint32_t MFCTypeServerIndex;
    int32_t OptionalDbgHeaderSize;
    int32_t ECSubstreamSize;
    uint16_t Flags;
    uint16_t Machine;
    uint32_t Padding;
};
```

- **VersionSignature** - Unknown meaning. Appears to always be -1.
- **VersionHeader** - A value from the following enum.

```
enum class DbiStreamVersion : uint32_t {
    VC41 = 930803,
    V50 = 19960307,
    V60 = 19970606,
```

```
V70 = 19990903,
V110 = 20091201
};
```

Similar to the *PDB Stream*, this value always appears to be V70, and it is not clear what the other values are for.

- **Age** - The number of times the PDB has been written. Equal to the same field from the *PDB Stream header*.
- **GlobalStreamIndex** - The index of the *Global Symbol Stream*, which contains CodeView symbol records for all global symbols. Actual records are stored in the symbol record stream, and are referenced from this stream.
- **BuildNumber** - A bitfield containing values representing the major and minor version number of the toolchain (e.g. 12.0 for MSVC 2013) used to build the program, with the following layout:

```
uint16_t MinorVersion : 8;
uint16_t MajorVersion : 7;
uint16_t NewVersionFormat : 1;
```

For the purposes of LLVM, we assume `NewVersionFormat` to be always true. If it is false, the layout above does not apply and the reader should consult the [Microsoft Source Code](#) for further guidance.

- **PublicStreamIndex** - The index of the *Public Symbol Stream*, which contains CodeView symbol records for all public symbols. Actual records are stored in the symbol record stream, and are referenced from this stream.
- **PdbDllVersion** - The version number of `mspdbXXXX.dll` used to produce this PDB. Note this obviously does not apply for LLVM as LLVM does not use `mspdb.dll`.
- **SymRecordStream** - The stream containing all CodeView symbol records used by the program. This is used for deduplication, so that many different compilands can refer to the same symbols without having to include the full record content inside of each module stream.
- **PdbDllRbld** - Unknown
- **MFCTypeServerIndex** - The length of the `:ref:dbi_mfc_type_server_substream`
- **Flags** - A bitfield with the following layout, containing various information about how the program was built:

```
uint16_t WasIncrementallyLinked : 1;
uint16_t ArePrivateSymbolsStripped : 1;
uint16_t HasConflictingTypes : 1;
uint16_t Reserved : 13;
```

The only one of these that is not self-explanatory is `HasConflictingTypes`. Although undocumented, `link.exe` contains a hidden flag `/DEBUG:CTYPES`. If it is passed to `link.exe`, this field will be set. Otherwise it will not be set. It is unclear what this flag does, although it seems to have subtle implications on the algorithm used to look up type records.

- **Machine** - A value from the `CV_CPU_TYPE_e` enumeration. Common values are `0x8664` (x86-64) and `0x14C` (x86).

Immediately after the fixed-size DBI Stream header are 7 variable-length *substreams*. The following 7 fields of the DBI Stream header specify the number of bytes of the corresponding substream. Each substream's contents will be described in detail *below*. The length of the entire DBI Stream should equal 64 (the length of the header above) plus the value of each of the following 7 fields.

- **ModInfoSize** - The length of the *Module Info Substream*.
- **SectionContributionSize** - The length of the *Section Contribution Substream*.
- **SectionMapSize** - The length of the *Section Map Substream*.
- **SourceInfoSize** - The length of the *File Info Substream*.

- **TypeServerSize** - The length of the *Type Server Substream*.
- **OptionalDbgHeaderSize** - The length of the *Optional Debug Header Stream*.
- **ECSubstreamSize** - The length of the *EC Substream*.

Substreams

Module Info Substream

Begins at offset 0 immediately after the *header*. The module info substream is an array of variable-length records, each one describing a single module (e.g. object file) linked into the program. Each record in the array has the format:

```
struct SectionContribEntry {
    uint16_t Section;
    char Padding1[2];
    int32_t Offset;
    int32_t Size;
    uint32_t Characteristics;
    uint16_t ModuleIndex;
    char Padding2[2];
    uint32_t DataCrc;
    uint32_t RelocCrc;
};
```

While most of these are self-explanatory, the *Characteristics* field warrants some elaboration. It corresponds to the *Characteristics* field of the `IMAGE_SECTION_HEADER` structure.

```
struct ModInfo {
    uint32_t Unused1;
    SectionContribEntry SectionContr;
    uint16_t Flags;
    uint16_t ModuleSymStream;
    uint32_t SymByteSize;
    uint32_t C11ByteSize;
    uint32_t C13ByteSize;
    uint16_t SourceFileCount;
    char Padding[2];
    uint32_t Unused2;
    uint32_t SourceFileNameIndex;
    uint32_t PdbFilePathNameIndex;
    char ModuleName[];
    char ObjFileName[];
};
```

- **SectionContr** - Describes the properties of the section in the final binary which contain the code and data from this module.
- **Flags** - A bitfield with the following format:

```
uint16_t Dirty : 1; // ``true`` if this ModInfo has been written since reading the_
↳PDB.
uint16_t EC : 1; // ``true`` if EC information is present for this module. It is_
↳unknown what EC actually is.
uint16_t Unused : 6;
uint16_t TSM : 8; // Type Server Index for this module. It is unknown what this_
↳is used for, but it is not used by LLVM.
```

- **ModuleSymStream** - The index of the stream that contains symbol information for this module. This includes CodeView symbol information as well as source and line information.
- **SymByteSize** - The number of bytes of data from the stream identified by `ModuleSymStream` that represent CodeView symbol records.
- **C11ByteSize** - The number of bytes of data from the stream identified by `ModuleSymStream` that represent C11-style CodeView line information.
- **C13ByteSize** - The number of bytes of data from the stream identified by `ModuleSymStream` that represent C13-style CodeView line information. At most one of `C11ByteSize` and `C13ByteSize` will be non-zero.
- **SourceFileCount** - The number of source files that contributed to this module during compilation.
- **SourceFileNameIndex** - The offset in the names buffer of the primary translation unit used to build this module. All PDB files observed to date always have this value equal to 0.
- **PdbFilePathNameIndex** - The offset in the names buffer of the PDB file containing this module's symbol information. This has only been observed to be non-zero for the special `* Linker *` module.
- **ModuleName** - The module name. This is usually either a full path to an object file (either directly passed to `link.exe` or from an archive) or a string of the form `Import:<dll name>`.
- **ObjFileName** - The object file name. In the case of an module that is linked directly passed to `link.exe`, this is the same as **ModuleName**. In the case of a module that comes from an archive, this is usually the full path to the archive.

Section Contribution Substream

Begins at offset 0 immediately after the *Module Info Substream* ends, and consumes `Header->SectionContributionSize` bytes. This substream begins with a single `uint32_t` which will be one of the following values:

```
enum class SectionContrSubstreamVersion : uint32_t {  
    Ver60 = 0xeffe0000 + 19970605,  
    V2 = 0xeffe0000 + 20140516  
};
```

`Ver60` is the only value which has been observed in a PDB so far. Following this 4 byte field is an array of fixed-length structures. If the version is `Ver60`, it is an array of `SectionContribEntry` structures. If the version is `V2`, it is an array of `SectionContribEntry2` structures, defined as follows:

```
struct SectionContribEntry2 {  
    SectionContribEntry SC;  
    uint32_t ISectCoff;  
};
```

The purpose of the second field is not well understood.

Section Map Substream

Begins at offset 0 immediately after the *Section Contribution Substream* ends, and consumes `Header->SectionMapSize` bytes. This substream begins with an 8 byte header followed by an array of fixed-length records. The header and records have the following layout:

```

struct SectionMapHeader {
    uint16_t Count;      // Number of segment descriptors
    uint16_t LogCount;  // Number of logical segment descriptors
};

struct SectionMapEntry {
    uint16_t Flags;      // See the SectionMapEntryFlags enum below.
    uint16_t Ovl;       // Logical overlay number
    uint16_t Group;     // Group index into descriptor array.
    uint16_t Frame;
    uint16_t SectionName; // Byte index of segment / group name in string table, or ↵
    ↵ 0xFFFF.
    uint16_t ClassName;  // Byte index of class in string table, or 0xFFFF.
    uint32_t Offset;     // Byte offset of the logical segment within physical ↵
    ↵ segment. If group is set in flags, this is the offset of the group.
    uint32_t SectionLength; // Byte count of the segment or group.
};

enum class SectionMapEntryFlags : uint16_t {
    Read = 1 << 0,      // Segment is readable.
    Write = 1 << 1,     // Segment is writable.
    Execute = 1 << 2,   // Segment is executable.
    AddressIs32Bit = 1 << 3, // Descriptor describes a 32-bit linear address.
    IsSelector = 1 << 8,  // Frame represents a selector.
    IsAbsoluteAddress = 1 << 9, // Frame represents an absolute address.
    IsGroup = 1 << 10    // If set, descriptor represents a group.
};

```

Many of these fields are not well understood, so will not be discussed further.

File Info Substream

Begins at offset 0 immediately after the *Section Map Substream* ends, and consumes Header->SourceInfoSize bytes. This substream defines the mapping from module to the source files that contribute to that module. Since multiple modules can use the same source file (for example, a header file), this substream uses a string table to store each unique file name only once, and then have each module use offsets into the string table rather than embedding the string's value directly. The format of this substream is as follows:

```

struct FileInfoSubstream {
    uint16_t NumModules;
    uint16_t NumSourceFiles;

    uint16_t ModIndices[NumModules];
    uint16_t ModFileCounts[NumModules];
    uint32_t FileNameOffsets[NumSourceFiles];
    char NamesBuffer[][NumSourceFiles];
};

```

NumModules - The number of modules for which source file information is contained within this substream. Should match the corresponding value from the ref:dbi_header.

NumSourceFiles: In theory this is supposed to contain the number of source files for which this substream contains information. But that would present a problem in that the width of this field being 16-bits would prevent one from having more than 64K source files in a program. In early versions of the file format, this seems to have been the case. In order to support more than this, this field of the is simply ignored, and computed dynamically by summing up the values of the ModFileCounts array (discussed below). In short, this value should be ignored.

ModIndices - This array is present, but does not appear to be useful.

ModFileCountArray - An array of `NumModules` integers, each one containing the number of source files which contribute to the module at the specified index. While each individual module is limited to 64K contributing source files, the union of all modules' source files may be greater than 64K. The real number of source files is thus computed by summing this array. Note that summing this array does not give the number of *unique* source files, only the total number of source file contributions to modules.

FileNameOffsets - An array of **NumSourceFiles** integers (where **NumSourceFiles** here refers to the 32-bit value obtained from summing **ModFileCountArray**), where each integer is an offset into **NamesBuffer** pointing to a null terminated string.

NamesBuffer - An array of null terminated strings containing the actual source file names.

Type Server Substream

Begins at offset 0 immediately after the *File Info Substream* ends, and consumes `Header->TypeServerSize` bytes. Neither the purpose nor the layout of this substream is understood, although it is assumed to related somehow to the usage of `/Zi` and `mspdbsrv.exe`. This substream will not be discussed further.

EC Substream

Begins at offset 0 immediately after the *Type Server Substream* ends, and consumes `Header->ECSubstreamSize` bytes. Neither the purpose nor the layout of this substream is understood, and it will not be discussed further.

Optional Debug Header Stream

Begins at offset 0 immediately after the *EC Substream* ends, and consumes `Header->OptionalDbgHeaderSize` bytes. This field is an array of stream indices (e.g. `uint16_t`'s), each of which identifies a stream index in the larger MSF file which contains some additional debug information. Each position of this array has a special meaning, allowing one to determine what kind of debug information is at the referenced stream. 11 indices are currently understood, although it's possible there may be more. The layout of each stream generally corresponds exactly to a particular type of debug data directory from the PE/COFF file. The format of these fields can be found in the [Microsoft PE/COFF Specification](#).

FPO Data - `DbgStreamArray[0]`. The data in the referenced stream is a debug data directory of type `IMAGE_DEBUG_TYPE_FPO`

Exception Data - `DbgStreamArray[1]`. The data in the referenced stream is a debug data directory of type `IMAGE_DEBUG_TYPE_EXCEPTION`.

Fixup Data - `DbgStreamArray[2]`. The data in the referenced stream is a debug data directory of type `IMAGE_DEBUG_TYPE_FIXUP`.

Omap To Src Data - `DbgStreamArray[3]`. The data in the referenced stream is a debug data directory of type `IMAGE_DEBUG_TYPE_OMAP_TO_SRC`. This is used for mapping addresses between instrumented and uninstrumented code.

Omap From Src Data - `DbgStreamArray[4]`. The data in the referenced stream is a debug data directory of type `IMAGE_DEBUG_TYPE_OMAP_FROM_SRC`. This is used for mapping addresses between instrumented and uninstrumented code.

Section Header Data - `DbgStreamArray[5]`. A dump of all section headers from the original executable.

Token / RID Map - `DbgStreamArray[6]`. The layout of this stream is not understood, but it is assumed to be a mapping from CLR Token to CLR Record ID. Refer to [ECMA 335](#) for more information.

Xdata - DbgStreamArray[7]. A copy of the .xdata section from the executable.

Pdata - DbgStreamArray[8]. This is assumed to be a copy of the .pdata section from the executable, but that would make it identical to DbgStreamArray[1]. The difference between these two indices is not well understood.

New FPO Data - DbgStreamArray[9]. The data in the referenced stream is a debug data directory of type IMAGE_DEBUG_TYPE_FPO. It is not clear how this differs from DbgStreamArray[0], but in practice all observed PDB files have used the “new” format rather than the “old” format.

Original Section Header Data - DbgStreamArray[10]. Assumed to be similar to DbgStreamArray[5], but has not been observed in practice.

The Module Information Stream

- *Introduction*
- *Stream Layout*
- *The CodeView Symbol Substream*

Introduction

The Module Info Stream (henceforth referred to as the Modi stream) contains information about a single module (object file, import library, etc that contributes to the binary this PDB contains debug information about. There is one modi stream for each module, and the mapping between modi stream index and module is contained in the *DBI Stream*. The modi stream for a single module contains line information for the compiland, as well as all CodeView information for the symbols defined in the compiland. Finally, there is a “global refs” substream which is not well understood.

Stream Layout

A modi stream is laid out as follows:

```
struct ModiStream {
    uint32_t Signature;
    uint8_t Symbols[SymbolSize-4];
    uint8_t C11LineInfo[C11Size];
    uint8_t C13LineInfo[C13Size];

    uint32_t GlobalRefsSize;
    uint8_t GlobalRefs[GlobalRefsSize];
};
```

- **Signature** - Unknown. In practice only the value of 4 has been observed. It is hypothesized that this value corresponds to the set of CV_SIGNATURE_xx defines in cvinfo.h, with the value of 4 meaning that this module has C13 line information (as opposed to C11 line information). A corollary of this is that we expect to only ever see C13 line info, and that we do not understand the format of C11 line info.
- **Symbols** - The *CodeView Symbol Substream*. SymbolSize is equal to the value of SymByteSize for the corresponding module’s entry in the *Module Info Substream* of the *DBI Stream*.
- **C11LineInfo** - A block containing CodeView line information in C11 format. C11Size is equal to the value of C11ByteSize from the *Module Info Substream* of the *DBI Stream*. If this value is 0, then C11 line information

is not present. As mentioned previously, the format of C11 line info is not understood and we assume all line in modern PDBs to be in C13 format.

- **C13LineInfo** - A block containing CodeView line information in C13 format. `C13Size` is equal to the value of `C13ByteSize` from the *Module Info Substream* of the *DBI Stream*. If this value is 0, then C13 line information is not present.
- **GlobalRefs** - The meaning of this substream is not understood.

The CodeView Symbol Substream

The CodeView Symbol Substream. This is an array of variable length records describing the functions, variables, inlining information, and other symbols defined in the compiland. The entire array consumes `SymbolSize-4` bytes. The format of a CodeView Symbol Record (and thusly, an array of CodeView Symbol Records) is described in *CodeView Symbol Records*.

The PDB Public Symbol Stream

The PDB Global Symbol Stream

The TPI & IPI Hash Streams

CodeView Symbol Records

CodeView Type Records

The MSF Container

A PDB file is really just a special case of an MSF (Multi-Stream Format) file. An MSF file is actually a miniature “file system within a file”. It contains multiple streams (aka files) which can represent arbitrary data, and these streams are divided into blocks which may not necessarily be contiguously laid out within the file (aka fragmented). Additionally, the MSF contains a stream directory (aka MFT) which describes how the streams (files) are laid out within the MSF.

For more information about the MSF container format, stream directory, and block layout, see *The MSF File Format*.

Streams

The PDB format contains a number of streams which describe various information such as the types, symbols, source files, and compilands (e.g. object files) of a program, as well as some additional streams containing hash tables that are used by debuggers and other tools to provide fast lookup of records and types by name, and various other information about how the program was compiled such as the specific toolchain used, and more. A summary of streams contained in a PDB file is as follows:

Name	Stream Index	Contents
Old Directory	<ul style="list-style-type: none"> Fixed Stream Index 0 	<ul style="list-style-type: none"> Previous MSF Stream Directory
PDB Stream	<ul style="list-style-type: none"> Fixed Stream Index 1 	<ul style="list-style-type: none"> Basic File Information Fields to match EXE to this PDB Map of named streams to stream indices
TPI Stream	<ul style="list-style-type: none"> Fixed Stream Index 2 	<ul style="list-style-type: none"> CodeView Type Records Index of TPI Hash Stream
DBI Stream	<ul style="list-style-type: none"> Fixed Stream Index 3 	<ul style="list-style-type: none"> Module/Compiland Information Indices of individual module streams Indices of public / global streams Section Contribution Information Source File Information FPO / PGO Data
IPI Stream	<ul style="list-style-type: none"> Fixed Stream Index 4 	<ul style="list-style-type: none"> CodeView Type Records Index of IPI Hash Stream
/LinkInfo	<ul style="list-style-type: none"> Contained in PDB Stream Named Stream map 	<ul style="list-style-type: none"> Unknown
/src/headerblock	<ul style="list-style-type: none"> Contained in PDB Stream Named Stream map 	<ul style="list-style-type: none"> Unknown
/names	<ul style="list-style-type: none"> Contained in PDB Stream Named Stream map 	<ul style="list-style-type: none"> PDB-wide global string table used for string de-duplication
Module Info Stream	<ul style="list-style-type: none"> Contained in DBI Stream One for each compiland 	<ul style="list-style-type: none"> CodeView Symbol Records for this module Line Number Information
Public Stream	<ul style="list-style-type: none"> Contained in DBI Stream 	<ul style="list-style-type: none"> Public (Exported) Symbol Records Index of Public Hash Stream
Global Stream	<ul style="list-style-type: none"> Contained in DBI Stream 	<ul style="list-style-type: none"> Global Symbol Records Index of Global Hash Stream
TPI Hash Stream	<ul style="list-style-type: none"> Contained in TPI Stream 	<ul style="list-style-type: none"> Hash table for looking up TPI records by name
4.38. The PDB File Format		1401
IPI Hash Stream	<ul style="list-style-type: none"> Contained in IPI Stream 	<ul style="list-style-type: none"> Hash table for looking up IPI records by name

More information about the structure of each of these can be found on the following pages:

The PDB Info Stream (aka the PDB Stream) Information about the PDB Info Stream and how it is used to match PDBs to EXEs.

The PDB TPI Stream Information about the TPI stream and the CodeView records contained within.

The PDB DBI (Debug Info) Stream Information about the DBI stream and relevant substreams including the Module Substreams, source file information, and CodeView symbol records contained within.

The Module Information Stream Information about the Module Information Stream, of which there is one for each compilation unit and the format of symbols contained within.

The PDB Public Symbol Stream Information about the Public Symbol Stream.

The PDB Global Symbol Stream Information about the Global Symbol Stream.

The TPI & IPI Hash Streams Information about the Hash Table stream, and how it can be used to quickly look up records by name.

4.38.3 CodeView

CodeView is another format which comes into the picture. While MSF defines the structure of the overall file, and PDB defines the set of streams that appear within the MSF file and the format of those streams, CodeView defines the format of **symbol and type records** that appear within specific streams. Refer to the pages on *CodeView Symbol Records* and *CodeView Type Records* for more information about the CodeView format.

Writing an LLVM Pass Information on how to write LLVM transformations and analyses.

Writing an LLVM Backend Information on how to write LLVM backends for machine targets.

The LLVM Target-Independent Code Generator The design and implementation of the LLVM code generator. Useful if you are working on retargeting LLVM to a new architecture, designing a new codegen pass, or enhancing existing components.

Machine IR (MIR) Format Reference Manual A reference manual for the MIR serialization format, which is used to test LLVM's code generation passes.

TableGen Describes the TableGen tool, which is used heavily by the LLVM code generator.

LLVM Alias Analysis Infrastructure Information on how to write a new alias analysis implementation or how to use existing analyses.

MemorySSA Information about the MemorySSA utility in LLVM, as well as how to use it.

Garbage Collection with LLVM The interfaces source-language compilers should use for compiling GC'd programs.

Source Level Debugging with LLVM This document describes the design and philosophy behind the LLVM source-level debugger.

Auto-Vectorization in LLVM This document describes the current status of vectorization in LLVM.

Exception Handling in LLVM This document describes the design and implementation of exception handling in LLVM.

LLVM bugpoint tool: design and usage Automatic bug finder and test-case reducer description and usage information.

LLVM Bitcode File Format This describes the file format and encoding used for LLVM "bc" files.

System Library This document describes the LLVM System Library (`lib/System`) and how to keep LLVM source code portable

LLVM Link Time Optimization: Design and Implementation This document describes the interface between LLVM intermodular optimizer and the linker and its design

The LLVM gold plugin How to build your programs with link-time optimization on Linux.

Debugging JIT-ed Code With GDB How to debug JITed code with GDB.

MCJIT Design and Implementation Describes the inner workings of MCJIT execution engine.

LLVM Branch Weight Metadata Provides information about Branch Prediction Information.

LLVM Block Frequency Terminology Provides information about terminology used in the BlockFrequencyInfo analysis pass.

Segmented Stacks in LLVM This document describes segmented stacks and how they are used in LLVM.

LLVM's Optional Rich Disassembly Output This document describes the optional rich disassembly output syntax.

How To Use Attributes Answers some questions about the new Attributes infrastructure.

User Guide for NVPTX Back-end This document describes using the NVPTX backend to compile GPU kernels.

User Guide for AMDGPU Backend This document describes using the AMDGPU backend to compile GPU kernels.

Stack maps and patch points in LLVM LLVM support for mapping instruction addresses to the location of values and allowing code to be patched.

Using ARM NEON instructions in big endian mode LLVM's support for generating NEON instructions on big endian ARM targets is somewhat nonintuitive. This document explains the implementation and rationale.

LLVM Code Coverage Mapping Format This describes the format and encoding used for LLVM's code coverage mapping.

Garbage Collection Safepoints in LLVM This describes a set of experimental extensions for garbage collection support.

MergeFunctions pass, how it works Describes functions merging optimization.

Design and Usage of the InAlloca Attribute Description of the `inalloca` argument attribute.

FaultMaps and implicit checks LLVM support for folding control flow into faulting machine instructions.

Compiling CUDA with clang LLVM support for CUDA.

Coroutines in LLVM LLVM support for coroutines.

Global Instruction Selection This describes the prototype instruction selection replacement, GlobalISel.

XRay Instrumentation High-level documentation of how to use XRay in LLVM.

Debugging with XRay An example of how to debug an application with XRay.

The Microsoft PDB File Format A detailed description of the Microsoft PDB (Program Database) file format.

Development Process Documentation

Information about LLVM's development process.

5.1 LLVM Developer Policy

- *Introduction*
- *Developer Policies*
 - *Stay Informed*
 - *Making and Submitting a Patch*
 - *Code Reviews*
 - *Code Owners*
 - *Test Cases*
 - *Quality*
 - *Commit messages*
 - *Obtaining Commit Access*
 - *Making a Major Change*
 - *Incremental Development*
 - *Attribution of Changes*
 - *IR Backwards Compatibility*
 - *C API Changes*
 - *New Targets*

- *Copyright, License, and Patents*
 - *Copyright*
 - *License*
 - *Patents*

5.1.1 Introduction

This document contains the LLVM Developer Policy which defines the project’s policy towards developers and their contributions. The intent of this policy is to eliminate miscommunication, rework, and confusion that might arise from the distributed nature of LLVM’s development. By stating the policy in clear terms, we hope each developer can know ahead of time what to expect when making LLVM contributions. This policy covers all [llvm.org](#) subprojects, including Clang, LLDB, libc++, etc.

This policy is also designed to accomplish the following objectives:

1. Attract both users and developers to the LLVM project.
2. Make life as simple and easy for contributors as possible.
3. Keep the top of Subversion trees as stable as possible.
4. Establish awareness of the project’s *copyright, license, and patent policies* with contributors to the project.

This policy is aimed at frequent contributors to LLVM. People interested in contributing one-off patches can do so in an informal way by sending them to the [llvm-commits mailing list](#) and engaging another developer to see it through the process.

5.1.2 Developer Policies

This section contains policies that pertain to frequent LLVM developers. We always welcome *one-off patches* from people who do not routinely contribute to LLVM, but we expect more from frequent contributors to keep the system as efficient as possible for everyone. Frequent LLVM contributors are expected to meet the following requirements in order for LLVM to maintain a high standard of quality.

Stay Informed

Developers should stay informed by reading at least the “dev” mailing list for the projects you are interested in, such as [llvm-dev](#) for LLVM, [cfe-dev](#) for Clang, or [lldb-dev](#) for LLDB. If you are doing anything more than just casual work on LLVM, it is suggested that you also subscribe to the “commits” mailing list for the subproject you’re interested in, such as [llvm-commits](#), [cfe-commits](#), or [lldb-commits](#). Reading the “commits” list and paying attention to changes being made by others is a good way to see what other people are interested in and watching the flow of the project as a whole.

We recommend that active developers register an email account with [LLVM Bugzilla](#) and preferably subscribe to the [llvm-bugs](#) email list to keep track of bugs and enhancements occurring in LLVM. We really appreciate people who are proactive at catching incoming bugs in their components and dealing with them promptly.

Please be aware that all public LLVM mailing lists are public and archived, and that notices of confidentiality or non-disclosure cannot be respected.

Making and Submitting a Patch

When making a patch for review, the goal is to make it as easy for the reviewer to read it as possible. As such, we recommend that you:

1. Make your patch against the Subversion trunk, not a branch, and not an old version of LLVM. This makes it easy to apply the patch. For information on how to check out SVN trunk, please see the [Getting Started Guide](#).
2. Similarly, patches should be submitted soon after they are generated. Old patches may not apply correctly if the underlying code changes between the time the patch was created and the time it is applied.
3. Patches should be made with `svn diff`, or similar. If you use a different tool, make sure it uses the `diff -u` format and that it doesn't contain clutter which makes it hard to read.
4. If you are modifying generated files, such as the top-level `configure` script, please separate out those changes into a separate patch from the rest of your changes.

Once your patch is ready, submit it by emailing it to the appropriate project's commit mailing list (or commit it directly if applicable). Alternatively, some patches get sent to the project's development list or component of the LLVM bug tracker, but the commit list is the primary place for reviews and should generally be preferred.

When sending a patch to a mailing list, it is a good idea to send it as an *attachment* to the message, not embedded into the text of the message. This ensures that your mailer will not mangle the patch when it sends it (e.g. by making whitespace changes or by wrapping lines).

For Thunderbird users: Before submitting a patch, please open *Preferences > Advanced > General > Config Editor*, find the key `mail.content_disposition_type`, and set its value to 1. Without this setting, Thunderbird sends your attachment using `Content-Disposition: inline` rather than `Content-Disposition: attachment`. Apple Mail gamely displays such a file inline, making it difficult to work with for reviewers using that program.

When submitting patches, please do not add confidentiality or non-disclosure notices to the patches themselves. These notices conflict with the [LLVM License](#) and may result in your contribution being excluded.

Code Reviews

LLVM has a code review policy. Code review is one way to increase the quality of software. We generally follow these policies:

1. All developers are required to have significant changes reviewed before they are committed to the repository.
2. Code reviews are conducted by email on the relevant project's commit mailing list, or alternatively on the project's development list or bug tracker.
3. Code can be reviewed either before it is committed or after. We expect major changes to be reviewed before being committed, but smaller changes (or changes where the developer owns the component) can be reviewed after commit.
4. The developer responsible for a code change is also responsible for making all necessary review-related changes.
5. Code review can be an iterative process, which continues until the patch is ready to be committed. Specifically, once a patch is sent out for review, it needs an explicit "looks good" before it is submitted. Do not assume silent approval, or request active objections to the patch with a deadline.

Sometimes code reviews will take longer than you would hope for, especially for larger features. Accepted ways to speed up review times for your patches are:

- Review other people's patches. If you help out, everybody will be more willing to do the same for you; goodwill is our currency.

- Ping the patch. If it is urgent, provide reasons why it is important to you to get this patch landed and ping it every couple of days. If it is not urgent, the common courtesy ping rate is one week. Remember that you're asking for valuable time from other professional developers.
- Ask for help on IRC. Developers on IRC will be able to either help you directly, or tell you who might be a good reviewer.
- Split your patch into multiple smaller patches that build on each other. The smaller your patch, the higher the probability that somebody will take a quick look at it.

Developers should participate in code reviews as both reviewers and reviewees. If someone is kind enough to review your code, you should return the favor for someone else. Note that anyone is welcome to review and give feedback on a patch, but only people with Subversion write access can approve it.

There is a web based code review tool that can optionally be used for code reviews. See [Code Reviews with Phabricator](#).

Code Owners

The LLVM Project relies on two features of its process to maintain rapid development in addition to the high quality of its source base: the combination of code review plus post-commit review for trusted maintainers. Having both is a great way for the project to take advantage of the fact that most people do the right thing most of the time, and only commit patches without pre-commit review when they are confident they are right.

The trick to this is that the project has to guarantee that all patches that are committed are reviewed after they go in: you don't want everyone to assume someone else will review it, allowing the patch to go unreviewed. To solve this problem, we have a notion of an 'owner' for a piece of the code. The sole responsibility of a code owner is to ensure that a commit to their area of the code is appropriately reviewed, either by themselves or by someone else. The list of current code owners can be found in the file `CODE_OWNERS.TXT` in the root of the LLVM source tree.

Note that code ownership is completely different than reviewers: anyone can review a piece of code, and we welcome code review from anyone who is interested. Code owners are the "last line of defense" to guarantee that all patches that are committed are actually reviewed.

Being a code owner is a somewhat unglamorous position, but it is incredibly important for the ongoing success of the project. Because people get busy, interests change, and unexpected things happen, code ownership is purely opt-in, and anyone can choose to resign their "title" at any time. For now, we do not have an official policy on how one gets elected to be a code owner.

Test Cases

Developers are required to create test cases for any bugs fixed and any new features added. Some tips for getting your testcase approved:

- All feature and regression test cases are added to the `llvm/test` directory. The appropriate sub-directory should be selected (see the [Testing Guide](#) for details).
- Test cases should be written in *LLVM assembly language*.
- Test cases, especially for regressions, should be reduced as much as possible, by [bugpoint](#) or manually. It is unacceptable to place an entire failing program into `llvm/test` as this creates a *time-to-test* burden on all developers. Please keep them short.

Note that `llvm/test` and `clang/test` are designed for regression and small feature tests only. More extensive test cases (e.g., entire applications, benchmarks, etc) should be added to the `llvm-test` test suite. The `llvm-test` suite is for coverage (correctness, performance, etc) testing, not feature or regression testing.

Quality

The minimum quality standards that any change must satisfy before being committed to the main development branch are:

1. Code must adhere to the [LLVM Coding Standards](#).
2. Code must compile cleanly (no errors, no warnings) on at least one platform.
3. Bug fixes and new features should *include a testcase* so we know if the fix/feature ever regresses in the future.
4. Code must pass the `llvm/test` test suite.
5. The code must not cause regressions on a reasonable subset of `llvm-test`, where “reasonable” depends on the contributor’s judgement and the scope of the change (more invasive changes require more testing). A reasonable subset might be something like “`llvm-test/MultiSource/Benchmarks`”.

Additionally, the committer is responsible for addressing any problems found in the future that the change is responsible for. For example:

- The code should compile cleanly on all supported platforms.
- The changes should not cause any correctness regressions in the `llvm-test` suite and must not cause any major performance regressions.
- The change set should not cause performance or correctness regressions for the LLVM tools.
- The changes should not cause performance or correctness regressions in code compiled by LLVM on all applicable targets.
- You are expected to address any [Bugzilla bugs](#) that result from your change.

We prefer for this to be handled before submission but understand that it isn’t possible to test all of this for every submission. Our build bots and nightly testing infrastructure normally finds these problems. A good rule of thumb is to check the nightly testers for regressions the day after your change. Build bots will directly email you if a group of commits that included yours caused a failure. You are expected to check the build bot messages to see if they are your fault and, if so, fix the breakage.

Commits that violate these quality standards (e.g. are very broken) may be reverted. This is necessary when the change blocks other developers from making progress. The developer is welcome to re-commit the change after the problem has been fixed.

Commit messages

Although we don’t enforce the format of commit messages, we prefer that you follow these guidelines to help review, search in logs, email formatting and so on. These guidelines are very similar to rules used by other open source projects.

Most importantly, the contents of the message should be carefully written to convey the rationale of the change (without delving too much in detail). It also should avoid being vague or overly specific. For example, “bits were not set right” will leave the reviewer wondering about which bits, and why they weren’t right, while “Correctly set overflow bits in `TargetInfo`” conveys almost all there is to the change.

Below are some guidelines about the format of the message itself:

- Separate the commit message into title, body and, if you’re not the original author, a “Patch by” attribution line (see below).
- The title should be concise. Because all commits are emailed to the list with the first line as the subject, long titles are frowned upon. Short titles also look better in *git log*.

- When the changes are restricted to a specific part of the code (e.g. a back-end or optimization pass), it is customary to add a tag to the beginning of the line in square brackets. For example, “[SCEV] ...” or “[OpenMP] ...”. This helps email filters and searches for post-commit reviews.
- The body, if it exists, should be separated from the title by an empty line.
- The body should be concise, but explanatory, including a complete reasoning. Unless it is required to understand the change, examples, code snippets and gory details should be left to bug comments, web review or the mailing list.
- If the patch fixes a bug in bugzilla, please include the PR# in the message.
- *Attribution of Changes* should be in a separate line, after the end of the body, as simple as “Patch by John Doe.”. This is how we officially handle attribution, and there are automated processes that rely on this format.
- Text formatting and spelling should follow the same rules as documentation and in-code comments, ex. capitalization, full stop, etc.
- If the commit is a bug fix on top of another recently committed patch, or a revert or reapply of a patch, include the svn revision number of the prior related commit. This could be as simple as “Revert rNNNN because it caused PR#”.

For minor violations of these recommendations, the community normally favors reminding the contributor of this policy over reverting. Minor corrections and omissions can be handled by sending a reply to the commits mailing list.

Obtaining Commit Access

We grant commit access to contributors with a track record of submitting high quality patches. If you would like commit access, please send an email to [Chris](#) with the following information:

1. The user name you want to commit with, e.g. “hacker”.
2. The full name and email address you want message to llvm-commits to come from, e.g. “J. Random Hacker <hacker@yoyodyne.com>”.
3. A “password hash” of the password you want to use, e.g. “2ACR96qjUqsyM”. Note that you don’t ever tell us what your password is; you just give it to us in an encrypted form. To get this, run “htpasswd” (a utility that comes with apache) in *crypt* mode (often enabled with “-d”), or find a web page that will do it for you. Note that our system does not work with MD5 hashes. These are significantly longer than a crypt hash - e.g. “\$apr1\$vea6bBV2\$Z8IFx.AfeD8Lhq1ZFqJer0”, we only accept the shorter crypt hash.

Once you’ve been granted commit access, you should be able to check out an LLVM tree with an SVN URL of “[https://username@llvm.org/...](https://username@llvm.org/)” instead of the normal anonymous URL of “[http://llvm.org/...](http://llvm.org/)”. The first time you commit you’ll have to type in your password. Note that you may get a warning from SVN about an untrusted key; you can ignore this. To verify that your commit access works, please do a test commit (e.g. change a comment or add a blank line). Your first commit to a repository may require the autogenerated email to be approved by a mailing list. This is normal and will be done when the mailing list owner has time.

If you have recently been granted commit access, these policies apply:

1. You are granted *commit-after-approval* to all parts of LLVM. To get approval, submit a *patch* to [llvm-commits](#). When approved, you may commit it yourself.
2. You are allowed to commit patches without approval which you think are obvious. This is clearly a subjective decision — we simply expect you to use good judgement. Examples include: fixing build breakage, reverting obviously broken patches, documentation/comment changes, any other minor changes.
3. You are allowed to commit patches without approval to those portions of LLVM that you have contributed or maintain (i.e., have been assigned responsibility for), with the proviso that such commits must not break the build. This is a “trust but verify” policy, and commits of this nature are reviewed after they are committed.

4. Multiple violations of these policies or a single egregious violation may cause commit access to be revoked.

In any case, your changes are still subject to *code review* (either before or after they are committed, depending on the nature of the change). You are encouraged to review other peoples' patches as well, but you aren't required to do so.

Making a Major Change

When a developer begins a major new project with the aim of contributing it back to LLVM, they should inform the community with an email to the [llvm-dev](#) email list, to the extent possible. The reason for this is to:

1. keep the community informed about future changes to LLVM,
2. avoid duplication of effort by preventing multiple parties working on the same thing and not knowing about it, and
3. ensure that any technical issues around the proposed work are discussed and resolved before any significant work is done.

The design of LLVM is carefully controlled to ensure that all the pieces fit together well and are as consistent as possible. If you plan to make a major change to the way LLVM works or want to add a major new extension, it is a good idea to get consensus with the development community before you start working on it.

Once the design of the new feature is finalized, the work itself should be done as a series of *incremental changes*, not as a long-term development branch.

Incremental Development

In the LLVM project, we do all significant changes as a series of incremental patches. We have a strong dislike for huge changes or long-term development branches. Long-term development branches have a number of drawbacks:

1. Branches must have mainline merged into them periodically. If the branch development and mainline development occur in the same pieces of code, resolving merge conflicts can take a lot of time.
2. Other people in the community tend to ignore work on branches.
3. Huge changes (produced when a branch is merged back onto mainline) are extremely difficult to *code review*.
4. Branches are not routinely tested by our nightly tester infrastructure.
5. Changes developed as monolithic large changes often don't work until the entire set of changes is done. Breaking it down into a set of smaller changes increases the odds that any of the work will be committed to the main repository.

To address these problems, LLVM uses an incremental development style and we require contributors to follow this practice when making a large/invasive change. Some tips:

- Large/invasive changes usually have a number of secondary changes that are required before the big change can be made (e.g. API cleanup, etc). These sorts of changes can often be done before the major change is done, independently of that work.
- The remaining inter-related work should be decomposed into unrelated sets of changes if possible. Once this is done, define the first increment and get consensus on what the end goal of the change is.
- Each change in the set can be stand alone (e.g. to fix a bug), or part of a planned series of changes that works towards the development goal.
- Each change should be kept as small as possible. This simplifies your work (into a logical progression), simplifies code review and reduces the chance that you will get negative feedback on the change. Small increments also facilitate the maintenance of a high quality code base.

- Often, an independent precursor to a big change is to add a new API and slowly migrate clients to use the new API. Each change to use the new API is often “obvious” and can be committed without review. Once the new API is in place and used, it is much easier to replace the underlying implementation of the API. This implementation change is logically separate from the API change.

If you are interested in making a large change, and this scares you, please make sure to first *discuss the change/gather consensus* then ask about the best way to go about making the change.

Attribution of Changes

When contributors submit a patch to an LLVM project, other developers with commit access may commit it for the author once appropriate (based on the progression of code review, etc.). When doing so, it is important to retain correct attribution of contributions to their contributors. However, we do not want the source code to be littered with random attributions “this code written by J. Random Hacker” (this is noisy and distracting). In practice, the revision control system keeps a perfect history of who changed what, and the CREDITS.txt file describes higher-level contributions. If you commit a patch for someone else, please follow the attribution of changes in the simple manner as outlined by the *commit messages* section. Overall, please do not add contributor names to the source code.

Also, don’t commit patches authored by others unless they have submitted the patch to the project or you have been authorized to submit them on their behalf (you work together and your company authorized you to contribute the patches, etc.). The author should first submit them to the relevant project’s commit list, development list, or LLVM bug tracker component. If someone sends you a patch privately, encourage them to submit it to the appropriate list first.

IR Backwards Compatibility

When the IR format has to be changed, keep in mind that we try to maintain some backwards compatibility. The rules are intended as a balance between convenience for llvm users and not imposing a big burden on llvm developers:

- The textual format is not backwards compatible. We don’t change it too often, but there are no specific promises.
- Additions and changes to the IR should be reflected in `test/Bitcode/compatibility.ll`.
- The current LLVM version supports loading any bitcode since version 3.0.
- After each X.Y release, `compatibility.ll` must be copied to `compatibility-X.Y.ll`. The corresponding bitcode file should be assembled using the X.Y build and committed as `compatibility-X.Y.ll.bc`.
- Newer releases can ignore features from older releases, but they cannot miscompile them. For example, if `nsw` is ever replaced with something else, dropping it would be a valid way to upgrade the IR.
- Debug metadata is special in that it is currently dropped during upgrades.
- Non-debug metadata is defined to be safe to drop, so a valid way to upgrade it is to drop it. That is not very user friendly and a bit more effort is expected, but no promises are made.

C API Changes

- **Stability Guarantees:** The C API is, in general, a “best effort” for stability. This means that we make every attempt to keep the C API stable, but that stability will be limited by the abstractness of the interface and the stability of the C++ API that it wraps. In practice, this means that things like “create debug info” or “create this type of instruction” are likely to be less stable than “take this IR file and JIT it for my current machine”.
- **Release stability:** We won’t break the C API on the release branch with patches that go on that branch, with the exception that we will fix an unintentional C API break that will keep the release consistent with both the previous and next release.

- Testing: Patches to the C API are expected to come with tests just like any other patch.
- Including new things into the API: If an LLVM subcomponent has a C API already included, then expanding that C API is acceptable. Adding C API for subcomponents that don't currently have one needs to be discussed on the mailing list for design and maintainability feedback prior to implementation.
- Documentation: Any changes to the C API are required to be documented in the release notes so that it's clear to external users who do not follow the project how the C API is changing and evolving.

New Targets

LLVM is very receptive to new targets, even experimental ones, but a number of problems can appear when adding new large portions of code, and back-ends are normally added in bulk. We have found that landing large pieces of new code and then trying to fix emergent problems in-tree is problematic for a variety of reasons.

For these reasons, new targets are *always* added as *experimental* until they can be proven stable, and later moved to non-experimental. The difference between both classes is that experimental targets are not built by default (need to be added to `-DLLVM_TARGETS_TO_BUILD` at CMake time).

The basic rules for a back-end to be upstreamed in **experimental** mode are:

- Every target must have a *code owner*. The `CODE_OWNERS.TXT` file has to be updated as part of the first merge. The code owner makes sure that changes to the target get reviewed and steers the overall effort.
- There must be an active community behind the target. This community will help maintain the target by providing buildbots, fixing bugs, answering the LLVM community's questions and making sure the new target doesn't break any of the other targets, or generic code. This behavior is expected to continue throughout the lifetime of the target's code.
- The code must be free of contentious issues, for example, large changes in how the IR behaves or should be formed by the front-ends, unless agreed by the majority of the community via refactoring of the (*IR standard*) **before** the merge of the new target changes, following the *IR Backwards Compatibility*.
- The code conforms to all of the policies laid out in this developer policy document, including license, patent, and coding standards.
- The target should have either reasonable documentation on how it works (ISA, ABI, etc.) or a publicly available simulator/hardware (either free or cheap enough) - preferably both. This allows developers to validate assumptions, understand constraints and review code that can affect the target.

In addition, the rules for a back-end to be promoted to **official** are:

- The target must have addressed every other minimum requirement and have been stable in tree for at least 3 months. This cool down period is to make sure that the back-end and the target community can endure continuous upstream development for the foreseeable future.
- The target's code must have been completely adapted to this policy as well as the *coding standards*. Any exceptions that were made to move into experimental mode must have been fixed **before** becoming official.
- The test coverage needs to be broad and well written (small tests, well documented). The build target `check-all` must pass with the new target built, and where applicable, the `test-suite` must also pass without errors, in at least one configuration (publicly demonstrated, for example, via buildbots).
- Public buildbots need to be created and actively maintained, unless the target requires no additional buildbots (ex. `check-all` covers all tests). The more relevant and public the new target's CI infrastructure is, the more the LLVM community will embrace it.

To **continue** as a supported and official target:

- The maintainer(s) must continue following these rules throughout the lifetime of the target. Continuous violations of aforementioned rules and policies could lead to complete removal of the target from the code base.

- Degradation in support, documentation or test coverage will make the target as nuisance to other targets and be considered a candidate for deprecation and ultimately removed.

In essences, these rules are necessary for targets to gain and retain their status, but also markers to define bit-rot, and will be used to clean up the tree from unmaintained targets.

5.1.3 Copyright, License, and Patents

Note: This section deals with legal matters but does not provide legal advice. We are not lawyers — please seek legal counsel from an attorney.

This section addresses the issues of copyright, license and patents for the LLVM project. The copyright for the code is held by the individual contributors of the code and the terms of its license to LLVM users and developers is the [University of Illinois/NCSA Open Source License](#) (with portions dual licensed under the [MIT License](#), see below). As contributor to the LLVM project, you agree to allow any contributions to the project to licensed under these terms.

Copyright

The LLVM project does not require copyright assignments, which means that the copyright for the code in the project is held by its respective contributors who have each agreed to release their contributed code under the terms of the [LLVM License](#).

An implication of this is that the LLVM license is unlikely to ever change: changing it would require tracking down all the contributors to LLVM and getting them to agree that a license change is acceptable for their contribution. Since there are no plans to change the license, this is not a cause for concern.

As a contributor to the project, this means that you (or your company) retain ownership of the code you contribute, that it cannot be used in a way that contradicts the license (which is a liberal BSD-style license), and that the license for your contributions won't change without your approval in the future.

License

We intend to keep LLVM perpetually open source and to use a liberal open source license. **As a contributor to the project, you agree that any contributions be licensed under the terms of the corresponding subproject.** All of the code in LLVM is available under the [University of Illinois/NCSA Open Source License](#), which boils down to this:

- You can freely distribute LLVM.
- You must retain the copyright notice if you redistribute LLVM.
- Binaries derived from LLVM must reproduce the copyright notice (e.g. in an included readme file).
- You can't use our names to promote your LLVM derived products.
- There's no warranty on LLVM at all.

We believe this fosters the widest adoption of LLVM because it **allows commercial products to be derived from LLVM** with few restrictions and without a requirement for making any derived works also open source (i.e. LLVM's license is not a "copyleft" license like the GPL). We suggest that you read the [License](#) if further clarification is needed.

In addition to the UIUC license, the runtime library components of LLVM (**compiler_rt**, **libc++**, and **libclc**) are also licensed under the [MIT License](#), which does not contain the binary redistribution clause. As a user of these runtime libraries, it means that you can choose to use the code under either license (and thus don't need the binary redistribution clause), and as a contributor to the code that you agree that any contributions to these libraries be licensed under both licenses. We feel that this is important for runtime libraries, because they are implicitly linked into applications and

therefore should not subject those applications to the binary redistribution clause. This also means that it is ok to move code from (e.g.) `libc++` to the LLVM core without concern, but that code cannot be moved from the LLVM core to `libc++` without the copyright owner's permission.

Note that the LLVM Project does distribute `dragonegg`, **which is GPL**. This means that anything “linked” into `dragonegg` must itself be compatible with the GPL, and must be releasable under the terms of the GPL. This implies that **any code linked into `dragonegg` and distributed to others may be subject to the viral aspects of the GPL** (for example, a proprietary code generator linked into `dragonegg` must be made available under the GPL). This is not a problem for code already distributed under a more liberal license (like the UIUC license), and GPL-containing sub-projects are kept in separate SVN repositories whose `LICENSE.txt` files specifically indicate that they contain GPL code.

We have no plans to change the license of LLVM. If you have questions or comments about the license, please contact the [LLVM Developer's Mailing List](#).

Patents

To the best of our knowledge, LLVM does not infringe on any patents (we have actually removed code from LLVM in the past that was found to infringe). Having code in LLVM that infringes on patents would violate an important goal of the project by making it hard or impossible to reuse the code for arbitrary purposes (including commercial use).

When contributing code, we expect contributors to notify us of any potential for patent-related trouble with their changes (including from third parties). If you or your employer own the rights to a patent and would like to contribute code to LLVM that relies on it, we require that the copyright owner sign an agreement that allows any other user of LLVM to freely use your patent. Please contact the [LLVM Foundation Board of Directors](#) for more details.

5.2 Creating an LLVM Project

- *Overview*
- *Source Tree Layout*
- *Writing LLVM Style Makefiles*
 - *Required Variables*
 - *Variables for Building Subdirectories*
 - *Variables for Building Libraries*
 - *Variables for Building Programs*
 - *Miscellaneous Variables*
- *Placement of Object Code*
- *Further Help*

5.2.1 Overview

The LLVM build system is designed to facilitate the building of third party projects that use LLVM header files, libraries, and tools. In order to use these facilities, a `Makefile` from a project must do the following things:

- Set `make` variables. There are several variables that a `Makefile` needs to set to use the LLVM build system:

- `PROJECT_NAME` - The name by which your project is known.
 - `LLVM_SRC_ROOT` - The root of the LLVM source tree.
 - `LLVM_OBJ_ROOT` - The root of the LLVM object tree.
 - `PROJ_SRC_ROOT` - The root of the project's source tree.
 - `PROJ_OBJ_ROOT` - The root of the project's object tree.
 - `PROJ_INSTALL_ROOT` - The root installation directory.
 - `LEVEL` - The relative path from the current directory to the project's root (`$PROJ_OBJ_ROOT`).
- Include `Makefile.config` from `$(LLVM_OBJ_ROOT)`.
 - Include `Makefile.rules` from `$(LLVM_SRC_ROOT)`.

There are two ways that you can set all of these variables:

- You can write your own `Makefiles` which hard-code these values.
- You can use the pre-made LLVM sample project. This sample project includes `Makefiles`, a configure script that can be used to configure the location of LLVM, and the ability to support multiple object directories from a single source directory.

If you want to devise your own build system, studying other projects and LLVM `Makefiles` will probably provide enough information on how to write your own `Makefiles`.

5.2.2 Source Tree Layout

In order to use the LLVM build system, you will want to organize your source code so that it can benefit from the build system's features. Mainly, you want your source tree layout to look similar to the LLVM source tree layout.

Underneath your top level directory, you should have the following directories:

lib

This subdirectory should contain all of your library source code. For each library that you build, you will have one directory in **lib** that will contain that library's source code.

Libraries can be object files, archives, or dynamic libraries. The **lib** directory is just a convenient place for libraries as it places them all in a directory from which they can be linked later.

include

This subdirectory should contain any header files that are global to your project. By global, we mean that they are used by more than one library or executable of your project.

By placing your header files in **include**, they will be found automatically by the LLVM build system. For example, if you have a file **include/jazz/note.h**, then your source files can include it simply with **#include "jazz/note.h"**.

tools

This subdirectory should contain all of your source code for executables. For each program that you build, you will have one directory in **tools** that will contain that program's source code.

test

This subdirectory should contain tests that verify that your code works correctly. Automated tests are especially useful.

Currently, the LLVM build system provides basic support for tests. The LLVM system provides the following:

- LLVM contains regression tests in `llvm/test`. These tests are run by the *Lit* testing tool. This test procedure uses RUN lines in the actual test case to determine how to run the test. See the *LLVM Testing Infrastructure Guide* for more details.
- LLVM contains an optional package called `llvm-test`, which provides benchmarks and programs that are known to compile with the Clang front end. You can use these programs to test your code, gather statistical information, and compare it to the current LLVM performance statistics.

Currently, there is no way to hook your tests directly into the `llvm/test` testing harness. You will simply need to find a way to use the source provided within that directory on your own.

Typically, you will want to build your **lib** directory first followed by your **tools** directory.

5.2.3 Writing LLVM Style Makefiles

The LLVM build system provides a convenient way to build libraries and executables. Most of your project Makefiles will only need to define a few variables. Below is a list of the variables one can set and what they can do:

Required Variables

LEVEL

This variable is the relative path from this Makefile to the top directory of your project's source code. For example, if your source code is in `/tmp/src`, then the Makefile in `/tmp/src/jump/high` would set LEVEL to `"../.."`.

Variables for Building Subdirectories

DIRS

This is a space separated list of subdirectories that should be built. They will be built, one at a time, in the order specified.

PARALLEL_DIRS

This is a list of directories that can be built in parallel. These will be built after the directories in DIRS have been built.

OPTIONAL_DIRS

This is a list of directories that can be built if they exist, but will not cause an error if they do not exist. They are built serially in the order in which they are listed.

Variables for Building Libraries

LIBRARYNAME

This variable contains the base name of the library that will be built. For example, to build a library named `libsample.a`, LIBRARYNAME should be set to `sample`.

BUILD_ARCHIVE

By default, a library is a `.o` file that is linked directly into a program. To build an archive (also known as a static library), set the BUILD_ARCHIVE variable.

SHARED_LIBRARY

If SHARED_LIBRARY is defined in your Makefile, a shared (or dynamic) library will be built.

Variables for Building Programs

TOOLNAME

This variable contains the name of the program that will be built. For example, to build an executable named `sample`, `TOOLNAME` should be set to `sample`.

USEDLIBS

This variable holds a space separated list of libraries that should be linked into the program. These libraries must be libraries that come from your `lib` directory. The libraries must be specified without their `lib` prefix. For example, to link `libsampl.a`, you would set `USEDLIBS` to `sample.a`.

Note that this works only for statically linked libraries.

LLVMLIBS

This variable holds a space separated list of libraries that should be linked into the program. These libraries must be LLVM libraries. The libraries must be specified without their `lib` prefix. For example, to link with a driver that performs an IR transformation you might set `LLVMLIBS` to this minimal set of libraries `LLVMSupport.a LLVMCore.a LLVMBitReader.a LLVMAsmParser.a LLVMAalysis.a LLVMTransformUtils.a LLVMScalarOpts.a LLVMTarget.a`.

Note that this works only for statically linked libraries. LLVM is split into a large number of static libraries, and the list of libraries you require may be much longer than the list above. To see a full list of libraries use: `llvm-config --libs all`. Using `LINK_COMPONENTS` as described below, obviates the need to set `LLVMLIBS`.

LINK_COMPONENTS

This variable holds a space separated list of components that the LLVM Makefiles pass to the `llvm-config` tool to generate a link line for the program. For example, to link with all LLVM libraries use `LINK_COMPONENTS = all`.

LIBS

To link dynamic libraries, add `-l<library base name>` to the `LIBS` variable. The LLVM build system will look in the same places for dynamic libraries as it does for static libraries.

For example, to link `libsampl.so`, you would have the following line in your Makefile:

```
LIBS += -lsample
```

Note that `LIBS` must occur in the Makefile after the inclusion of `Makefile.common`.

Miscellaneous Variables

CFLAGS & CPPFLAGS

This variable can be used to add options to the C and C++ compiler, respectively. It is typically used to add options that tell the compiler the location of additional directories to search for header files.

It is highly suggested that you append to `CFLAGS` and `CPPFLAGS` as opposed to overwriting them. The master Makefiles may already have useful options in them that you may not want to overwrite.

5.2.4 Placement of Object Code

The final location of built libraries and executables will depend upon whether you do a `Debug`, `Release`, or `Profile` build.

Libraries

All libraries (static and dynamic) will be stored in `PROJ_OBJ_ROOT/<type>/lib`, where *type* is `Debug`, `Release`, or `Profile` for a debug, optimized, or profiled build, respectively.

Executables

All executables will be stored in `PROJ_OBJ_ROOT/<type>/bin`, where *type* is `Debug`, `Release`, or `Profile` for a debug, optimized, or profiled build, respectively.

5.2.5 Further Help

If you have any questions or need any help creating an LLVM project, the LLVM team would be more than happy to help. You can always post your questions to the [LLVM Developers Mailing List](#).

5.3 LLVMBuild Guide

- *Introduction*
- *Project Organization*
- *Build Integration*
- *Component Overview*
- *LLVMBuild Format Reference*

5.3.1 Introduction

This document describes the LLVMBuild organization and files which we use to describe parts of the LLVM ecosystem. For description of specific LLVMBuild related tools, please see the command guide.

LLVM is designed to be a modular set of libraries which can be flexibly mixed together in order to build a variety of tools, like compilers, JITs, custom code generators, optimization passes, interpreters, and so on. Related projects in the LLVM system like Clang and LLDB also tend to follow this philosophy.

In order to support this usage style, LLVM has a fairly strict structure as to how the source code and various components are organized. The `LLVMBuild.txt` files are the explicit specification of that structure, and are used by the build systems and other tools in order to develop the LLVM project.

5.3.2 Project Organization

The source code for LLVM projects using the LLVMBuild system (LLVM, Clang, and LLDB) is organized into *components*, which define the separate pieces of functionality that make up the project. These projects may consist of many libraries, associated tools, build tools, or other utility tools (for example, testing tools).

For the most part, the project contents are organized around defining one main component per each subdirectory. Each such directory contains an `LLVMBuild.txt` which contains the component definitions.

The component descriptions for the project as a whole are automatically gathered by the LLVMBuild tools. The tools automatically traverse the source directory structure to find all of the component description files. NOTE: For performance/sanity reasons, we only traverse into subdirectories when the parent itself contains an `LLVMBuild.txt` description file.

5.3.3 Build Integration

The LLVMBuild files themselves are just a declarative way to describe the project structure. The actual building of the LLVM project is handled by another build system (See: *CMake*).

The build system implementation will load the relevant contents of the LLVMBuild files and use that to drive the actual project build. Typically, the build system will only need to load this information at “configure” time, and use it to generate native information. Build systems will also handle automatically reconfiguring their information when the contents of the LLVMBuild.txt files change.

Developers generally are not expected to need to be aware of the details of how the LLVMBuild system is integrated into their build. Ideally, LLVM developers who are not working on the build system would only ever need to modify the contents of the LLVMBuild.txt description files (although we have not reached this goal yet).

For more information on the utility tool we provide to help interfacing with the build system, please see the *llvm-build* documentation.

5.3.4 Component Overview

As mentioned earlier, LLVM projects are organized into logical *components*. Every component is typically grouped into its own subdirectory. Generally, a component is organized around a coherent group of sources which have some kind of clear API separation from other parts of the code.

LLVM primarily uses the following types of components:

- *Libraries* - Library components define a distinct API which can be independently linked into LLVM client applications. Libraries typically have private and public header files, and may specify a link of required libraries that they build on top of.
- *Build Tools* - Build tools are applications which are designed to be run as part of the build process (typically to generate other source files). Currently, LLVM uses one main build tool called *TableGen* to generate a variety of source files.
- *Tools* - Command line applications which are built using the LLVM component libraries. Most LLVM tools are small and are primarily frontends to the library interfaces.

Components are described using LLVMBuild.txt files in the directories that define the component. See the *LLVM-Build Format Reference* section for information on the exact format of these files.

5.3.5 LLVMBuild Format Reference

LLVMBuild files are written in a simple variant of the INI or configuration file format ([Wikipedia entry](#)). The format defines a list of sections each of which may contain some number of properties. A simple example of the file format is below:

```
; Comments start with a semi-colon.

; Sections are declared using square brackets.
[component_0]

; Properties are declared using '=' and are contained in the previous section.
;
; We support simple string and boolean scalar values and list values, where
; items are separated by spaces. There is no support for quoting, and so
; property values may not contain spaces.
property_name = property_value
```

```
list_property_name = value_1 value_2 ... value_n
boolean_property_name = 1 (or 0)
```

LLVMBuild files are expected to define a strict set of sections and properties. A typical component description file for a library component would look like the following example:

```
[component_0]
type = Library
name = Linker
parent = Libraries
required_libraries = Archive BitReader Core Support TransformUtils
```

A full description of the exact sections and properties which are allowed follows.

Each file may define exactly one common component, named `common`. The common component may define the following properties:

- `subdirectories` **[optional]**

If given, a list of the names of the subdirectories from the current subpath to search for additional LLVMBuild files.

Each file may define multiple components. Each component is described by a section whose name starts with `component`. The remainder of the section name is ignored, but each section name must be unique. Typically components are just numbered in order for files with multiple components (`component_0`, `component_1`, and so on).

Warning: Section names not matching this format (or the `common` section) are currently unused and are disallowed.

Every component is defined by the properties in the section. The exact list of properties that are allowed depends on the component type. Components **may not** define any properties other than those expected by the component type.

Every component must define the following properties:

- `type` **[required]**

The type of the component. Supported component types are detailed below. Most components will define additional properties which may be required or optional.

- `name` **[required]**

The name of the component. Names are required to be unique across the entire project.

- `parent` **[required]**

The name of the logical parent of the component. Components are organized into a logical tree to make it easier to navigate and organize groups of components. The parents have no semantics as far as the project build is concerned, however. Typically, the parent will be the main component of the parent directory.

Components may reference the root pseudo component using `$ROOT` to indicate they should logically be grouped at the top-level.

Components may define the following properties:

- `dependencies` **[optional]**

If specified, a list of names of components which *must* be built prior to this one. This should only be exactly those components which produce some tool or source code required for building the component.

Note: `Group` and `LibraryGroup` components have no semantics for the actual build, and are not allowed to specify dependencies.

The following section lists the available component types, as well as the properties which are associated with that component.

- `type = Group`

`Group` components exist purely to allow additional arbitrary structuring of the logical components tree. For example, one might define a `Libraries` group to hold all of the root library components.

`Group` components have no additionally properties.

- `type = Library`

`Library` components define an individual library which should be built from the source code in the component directory.

Components with this type use the following properties:

- `library_name` **[optional]**

If given, the name to use for the actual library file on disk. If not given, the name is derived from the component name itself.

- `required_libraries` **[optional]**

If given, a list of the names of `Library` or `LibraryGroup` components which must also be linked in whenever this library is used. That is, the link time dependencies for this component. When tools are built, the build system will include the transitive closure of all `required_libraries` for the components the tool needs.

- `add_to_library_groups` **[optional]**

If given, a list of the names of `LibraryGroup` components which this component is also part of. This allows nesting groups of components. For example, the `X86` target might define a library group for all of the `X86` components. That library group might then be included in the `all-targets` library group.

- `installed` **[optional]** **[boolean]**

Whether this library is installed. Libraries that are not installed are only reported by `llvm-config` when it is run as part of a development directory.

- `type = LibraryGroup`

`LibraryGroup` components are a mechanism to allow easy definition of useful sets of related components. In particular, we use them to easily specify things like “all targets”, or “all assembly printers”.

Components with this type use the following properties:

- `required_libraries` **[optional]**

See the `Library` type for a description of this property.

- `add_to_library_groups` **[optional]**

See the `Library` type for a description of this property.

- `type = TargetGroup`

`TargetGroup` components are an extension of `LibraryGroups`, specifically for defining LLVM targets (which are handled specially in a few places).

The name of the component should always be the name of the target.

Components with this type use the `LibraryGroup` properties in addition to:

- `has_asmparser` **[optional] [boolean]**
Whether this target defines an assembly parser.
- `has_asmprinter` **[optional] [boolean]**
Whether this target defines an assembly printer.
- `has_disassembler` **[optional] [boolean]**
Whether this target defines a disassembler.
- `has_jit` **[optional] [boolean]**
Whether this target supports JIT compilation.

- `type = Tool`

`Tool` components define standalone command line tools which should be built from the source code in the component directory and linked.

Components with this type use the following properties:

- `required_libraries` **[optional]**
If given, a list of the names of `Library` or `LibraryGroup` components which this tool is required to be linked with.

Note: The values should be the component names, which may not always match up with the actual library names on disk.

Build systems are expected to properly include all of the libraries required by the linked components (i.e., the transitive closure of `required_libraries`).

Build systems are also expected to understand that those library components must be built prior to linking – they do not also need to be listed under `dependencies`.

- `type = BuildTool`

`BuildTool` components are like `Tool` components, except that the tool is supposed to be built for the platform where the build is running (instead of that platform being targeted). Build systems are expected to handle the fact that required libraries may need to be built for multiple platforms in order to be able to link this tool.

`BuildTool` components currently use the exact same properties as `Tool` components, the type distinction is only used to differentiate what the tool is built for.

5.4 How To Release LLVM To The Public

5.4.1 Introduction

This document contains information about successfully releasing LLVM — including sub-projects: e.g., `clang` and `compiler-rt` — to the public. It is the Release Manager's responsibility to ensure that a high quality build of LLVM is released.

If you're looking for the document on how to test the release candidates and create the binary packages, please refer to the [How To Validate a New Release](#) instead.

5.4.2 Release Timeline

LLVM is released on a time based schedule — with major releases roughly every 6 months. In between major releases there may be dot releases. The release manager will determine if and when to make a dot release based on feedback from the community. Typically, dot releases should be made if there are large number of bug-fixes in the stable branch or a critical bug has been discovered that affects a large number of users.

Unless otherwise stated, dot releases will follow the same procedure as major releases.

The release process is roughly as follows:

- Set code freeze and branch creation date for 6 months after last code freeze date. Announce release schedule to the LLVM community and update the website.
- Create release branch and begin release process.
- Send out release candidate sources for first round of testing. Testing lasts 7-10 days. During the first round of testing, any regressions found should be fixed. Patches are merged from mainline into the release branch. Also, all features need to be completed during this time. Any features not completed at the end of the first round of testing will be removed or disabled for the release.
- Generate and send out the second release candidate sources. Only *critical* bugs found during this testing phase will be fixed. Any bugs introduced by merged patches will be fixed. If so a third round of testing is needed.
- The release notes are updated.
- Finally, release!

The release process will be accelerated for dot releases. If the first round of testing finds no critical bugs and no regressions since the last major release, then additional rounds of testing will not be required.

5.4.3 Release Process

- *Release Administrative Tasks*
 - *Create Release Branch*
 - *Update LLVM Version*
 - *Tagging the LLVM Release Candidates*
 - *Build Clang Binary Distribution*
- *Release Qualification Criteria*
- *Official Testing*
- *Community Testing*
- *Reporting Regressions*
- *Release Patch Rules*
 - *Merging Patches*
- *Release Final Tasks*
 - *Update Documentation*
 - *Tag the LLVM Final Release*
- *Update the LLVM Demo Page*

- *Update the LLVM Website*
- *Announce the Release*

Release Administrative Tasks

This section describes a few administrative tasks that need to be done for the release process to begin. Specifically, it involves:

- Creating the release branch,
- Setting version numbers, and
- Tagging release candidates for the release team to begin testing.

Create Release Branch

Branch the Subversion trunk using the following procedure:

1. Remind developers that the release branching is imminent and to refrain from committing patches that might break the build. E.g., new features, large patches for works in progress, an overhaul of the type system, an exciting new TableGen feature, etc.
2. Verify that the current Subversion trunk is in decent shape by examining nightly tester and buildbot results.
3. Create the release branch for `llvm`, `clang`, and other sub-projects, from the last known good revision. The branch's name is `release_XY`, where `X` is the major and `Y` the minor release numbers. Use `utils/release/tag.sh` to tag the release.
4. Advise developers that they may now check their patches into the Subversion tree again.
5. The Release Manager should switch to the release branch, because all changes to the release will now be done in the branch. The easiest way to do this is to grab a working copy using the following commands:

```
$ svn co https://llvm.org/svn/llvm-project/llvm/branches/release_XY llvm-X.Y
$ svn co https://llvm.org/svn/llvm-project/cfe/branches/release_XY clang-X.Y
$ svn co https://llvm.org/svn/llvm-project/test-suite/branches/release_XY test-
↪suite-X.Y
```

Update LLVM Version

After creating the LLVM release branch, update the release branches' `autoconf` and `configure.ac` versions from `'X.Ysvn'` to `'X.Y'`. Update it on mainline as well to be the next version (`'X.Y+1svn'`). Regenerate the configure scripts for both `llvm` and the `test-suite`.

In addition, the version numbers of all the Bugzilla components must be updated for the next release.

Tagging the LLVM Release Candidates

Tag release candidates using the `tag.sh` script in `utils/release`.

```
$ ./tag.sh -release X.Y.Z -rc $RC
```

The Release Manager may supply pre-packaged source tarballs for users. This can be done with the `export.sh` script in `utils/release`.

```
$ ./export.sh -release X.Y.Z -rc $RC
```

This will generate source tarballs for each LLVM project being validated, which can be uploaded to the website for further testing.

Build Clang Binary Distribution

Creating the `clang` binary distribution requires following the instructions [here](#).

That process will perform both Release+Asserts and Release builds but only pack the Release build for upload. You should use the Release+Asserts sysroot, normally under `final/Phase3/Release+Asserts/llvmCore-3.8.1-RCn.install/`, for test-suite and run-time benchmarks, to make sure nothing serious has passed through the net. For compile-time benchmarks, use the Release version.

The minimum required version of the tools you'll need are [here](#)

Release Qualification Criteria

A release is qualified when it has no regressions from the previous release (or baseline). Regressions are related to correctness first and performance second. (We may tolerate some minor performance regressions if they are deemed necessary for the general quality of the compiler.)

More specifically, Clang/LLVM is qualified when it has a clean test with all supported sub-projects included (`make check-all`), per target, and it has no regressions with the `test-suite` in relation to the previous release.

Regressions are new failures in the set of tests that are used to qualify each product and only include things on the list. Every release will have some bugs in it. It is the reality of developing a complex piece of software. We need a very concrete and definitive release criteria that ensures we have monotonically improving quality on some metric. The metric we use is described below. This doesn't mean that we don't care about other criteria, but these are the criteria which we found to be most important and which must be satisfied before a release can go out.

Official Testing

A few developers in the community have dedicated time to validate the release candidates and volunteered to be the official release testers for each architecture.

These will be the ones testing, generating and uploading the official binaries to the server, and will be the minimum tests *necessary* for the release to proceed.

This will obviously not cover all OSs and distributions, so additional community validation is important. However, if community input is not reached before the release is out, all bugs reported will have to go on the next stable release.

The official release managers are:

- Major releases (X.0): Hans Wennborg
- Stable releases (X.n): Tom Stellard

The official release testers are volunteered from the community and have consistently validated and released binaries for their targets/OSs. To contact them, you should email the `release-testers@lists.llvm.org` mailing list.

The official testers list is in the file `RELEASE_TESTERS.TXT`, in the LLVM repository.

Community Testing

Once all testing has been completed and appropriate bugs filed, the release candidate tarballs are put on the website and the LLVM community is notified.

We ask that all LLVM developers test the release in any the following ways:

1. Download `llvm-X.Y`, `llvm-test-X.Y`, and the appropriate `clang` binary. Build LLVM. Run `make check` and the full LLVM test suite (`make TEST=nightly report`).
2. Download `llvm-X.Y`, `llvm-test-X.Y`, and the `clang` sources. Compile everything. Run `make check` and the full LLVM test suite (`make TEST=nightly report`).
3. Download `llvm-X.Y`, `llvm-test-X.Y`, and the appropriate `clang` binary. Build whole programs with it (ex. Chromium, Firefox, Apache) for your platform.
4. Download `llvm-X.Y`, `llvm-test-X.Y`, and the appropriate `clang` binary. Build *your* programs with it and check for conformance and performance regressions.
5. Run the *release process*, if your platform is *different* than that which is officially supported, and report back errors only if they were not reported by the official release tester for that architecture.

We also ask that the OS distribution release managers test their packages with the first candidate of every release, and report any *new* errors in Bugzilla. If the bug can be reproduced with an unpatched upstream version of the release candidate (as opposed to the distribution's own build), the priority should be release blocker.

During the first round of testing, all regressions must be fixed before the second release candidate is tagged.

In the subsequent stages, the testing is only to ensure that bug fixes previously merged in have not created new major problems. *This is not the time to solve additional and unrelated bugs!* If no patches are merged in, the release is determined to be ready and the release manager may move onto the next stage.

Reporting Regressions

Every regression that is found during the tests (as per the criteria above), should be filled in a bug in Bugzilla with the priority *release blocker* and blocking a specific release.

To help manage all the bugs reported and which ones are blockers or not, a new “[meta]” bug should be created and all regressions *blocking* that Meta. Once all blockers are done, the Meta can be closed.

If a bug can't be reproduced, or stops being a blocker, it should be removed from the Meta and its priority decreased to *normal*. Debugging can continue, but on trunk.

Release Patch Rules

Below are the rules regarding patching the release branch:

1. Patches applied to the release branch may only be applied by the release manager, the official release testers or the code owners with approval from the release manager.
2. During the first round of testing, patches that fix regressions or that are small and relatively risk free (verified by the appropriate code owner) are applied to the branch. Code owners are asked to be very conservative in approving patches for the branch. We reserve the right to reject any patch that does not fix a regression as previously defined.
3. During the remaining rounds of testing, only patches that fix critical regressions may be applied.
4. For dot releases all patches must maintain both API and ABI compatibility with the previous major release. Only bug-fixes will be accepted.

Merging Patches

The `utils/release/merge.sh` script can be used to merge individual revisions into any one of the LLVM projects. To merge revision `$N` into project `$PROJ`, do:

1. `svn co https://llvm.org/svn/llvm-project/$PROJ/branches/release_XX $PROJ.src`
2. `$PROJ.src/utils/release/merge.sh --proj $PROJ --rev $N`
3. Run regression tests.
4. `cd $PROJ.src`. Run the `svn commit` command printed out by `merge.sh` in step 2.

Release Final Tasks

The final stages of the release process involves tagging the “final” release branch, updating documentation that refers to the release, and updating the demo page.

Update Documentation

Review the documentation and ensure that it is up to date. The “Release Notes” must be updated to reflect new features, bug fixes, new known issues, and changes in the list of supported platforms. The “Getting Started Guide” should be updated to reflect the new release version number tag available from Subversion and changes in basic system requirements. Merge both changes from mainline into the release branch.

Tag the LLVM Final Release

Tag the final release sources using the `tag.sh` script in `utils/release`.

```
$ ./tag.sh -release X.Y.Z -final
```

Update the LLVM Demo Page

The LLVM demo page must be updated to use the new release. This consists of using the new `clang` binary and building LLVM.

Update the LLVM Website

The website must be updated before the release announcement is sent out. Here is what to do:

1. Check out the `www` module from Subversion.
2. Create a new sub-directory `X.Y` in the `releases` directory.
3. Commit the `llvm`, `test-suite`, `clang` source and binaries in this new directory.
4. Copy and commit the `llvm/docs` and `LICENSE.txt` files into this new directory. The docs should be built with `BUILD_FOR_WEBSITE=1`.
5. Commit the `index.html` to the `release/X.Y` directory to redirect (use from previous release).
6. Update the `releases/download.html` file with the new release.
7. Update the `releases/index.html` with the new release and link to release documentation.

8. Finally, update the main page (`index.html` and sidebar) to point to the new release and release announcement. Make sure this all gets committed back into Subversion.

Announce the Release

Send an email to the list announcing the release, pointing people to all the relevant documentation, download pages and bugs fixed.

5.5 Advice on Packaging LLVM

- *Overview*
- *Compile Flags*
- *C++ Features*
- *Shared Library*
- *Dependencies*

5.5.1 Overview

LLVM sets certain default configure options to make sure our developers don't break things for constrained platforms. These settings are not optimal for most desktop systems, and we hope that packagers (e.g., Redhat, Debian, MacPorts, etc.) will tweak them. This document lists settings we suggest you tweak.

LLVM's API changes with each release, so users are likely to want, for example, both LLVM-2.6 and LLVM-2.7 installed at the same time to support apps developed against each.

5.5.2 Compile Flags

LLVM runs much more quickly when it's optimized and assertions are removed. However, such a build is currently incompatible with users who build without defining `NDEBUG`, and the lack of assertions makes it hard to debug problems in user code. We recommend allowing users to install both optimized and debug versions of LLVM in parallel. The following configure flags are relevant:

- disable-assertions** Builds LLVM with `NDEBUG` defined. Changes the LLVM ABI. Also available by setting `DISABLE_ASSERTIONS=0|1` in make's environment. This defaults to enabled regardless of the optimization setting, but it slows things down.
- enable-debug-symbols** Builds LLVM with `-g`. Also available by setting `DEBUG_SYMBOLS=0|1` in make's environment. This defaults to disabled when optimizing, so you should turn it back on to let users debug their programs.
- enable-optimized** (For svn checkouts) Builds LLVM with `-O2` and, by default, turns off debug symbols. Also available by setting `ENABLE_OPTIMIZED=0|1` in make's environment. This defaults to enabled when not in a checkout.

5.5.3 C++ Features

RTTI LLVM disables RTTI by default. Add `REQUIRES_RTTI=1` to your environment while running `make` to re-enable it. This will allow users to build with RTTI enabled and still inherit from LLVM classes.

5.5.4 Shared Library

Configure with `--enable-shared` to build `libLLVM-<major>.<minor>.(so|dylib)` and link the tools against it. This saves lots of binary size at the cost of some startup time.

5.5.5 Dependencies

`--enable-libffi` Depend on `libffi` to allow the LLVM interpreter to call external functions.

`--with-oprofile`

Depend on `libopagent` (`>=version 0.9.4`) to let the LLVM JIT tell oprofile about function addresses and line numbers.

5.6 How To Validate a New Release

- *Introduction*
- *Scripts*
- *Test Suite*
- *Pre-Release Process*
- *Release Process*
- *Bug Reporting Process*

5.6.1 Introduction

This document contains information about testing the release candidates that will ultimately be the next LLVM release. For more information on how to manage the actual release, please refer to *[How To Release LLVM To The Public](#)*.

Overview of the Release Process

Once the release process starts, the Release Manager will ask for volunteers, and it'll be the role of each volunteer to:

- Test and benchmark the previous release
- Test and benchmark each release candidate, comparing to the previous release and candidates
- Identify, reduce and report every regression found during tests and benchmarks
- Make sure the critical bugs get fixed and merged to the next release candidate

Not all bugs or regressions are show-stoppers and it's a bit of a grey area what should be fixed before the next candidate and what can wait until the next release.

It'll depend on:

- The severity of the bug, how many people it affects and if it's a regression or a known bug. Known bugs are “unsupported features” and some bugs can be disabled if they have been implemented recently.
- The stage in the release. Less critical bugs should be considered to be fixed between RC1 and RC2, but not so much at the end of it.
- If it's a correctness or a performance regression. Performance regression tends to be taken more lightly than correctness.

5.6.2 Scripts

The scripts are in the `utils/release` directory.

test-release.sh

This script will check-out, configure and compile LLVM+Clang (+ most add-ons, like `compiler-rt`, `libcxx`, `libomp` and `clang-extra-tools`) in three stages, and will test the final stage. It'll have installed the final binaries on the `Phase3/Releasei(+Asserts)` directory, and that's the one you should use for the test-suite and other external tests.

To run the script on a specific release candidate run:

```
./test-release.sh \
  -release 3.3 \
  -rc 1 \
  -no-64bit \
  -test-asserts \
  -no-compare-files
```

Each system will require different options. For instance, `x86_64` will obviously not need `-no-64bit` while 32-bit systems will, or the script will fail.

The important flags to get right are:

- On the pre-release, you should change `-rc 1` to `-final`. On RC2, change it to `-rc 2` and so on.
- On non-release testing, you can use `-final` in conjunction with `-no-checkout`, but you'll have to create the `final` directory by hand and link the correct source dir to `final/llvm.src`.
- For release candidates, you need `-test-asserts`, or it won't create a “Release+Asserts” directory, which is needed for release testing and benchmarking. This will take twice as long.
- On the final candidate you just need Release builds, and that's the binary directory you'll have to pack.

This script builds three phases of Clang+LLVM twice each (Release and Release+Asserts), so use `screen` or `nohup` to avoid headaches, since it'll take a long time.

Use the `--help` option to see all the options and chose it according to your needs.

findRegressions-nightly.py

TODO

5.6.3 Test Suite

Follow the [LNT Quick Start Guide](#) link on how to set-up the test-suite

The binary location you'll have to use for testing is inside the `rcN/Phase3/Release+Asserts/llvmCore-REL-RC.install`. Link that directory to an easier location and run the test-suite.

An example on the run command line, assuming you created a link from the correct install directory to `~/devel/llvm/install`:

```
./sandbox/bin/python sandbox/bin/lnt runtest \
  nt \
  -j4 \
  --sandbox sandbox \
  --test-suite ~/devel/llvm/test/test-suite \
  --cc ~/devel/llvm/install/bin/clang \
  --cxx ~/devel/llvm/install/bin/clang++
```

It should have no new regressions, compared to the previous release or release candidate. You don't need to fix all the bugs in the test-suite, since they're not necessarily meant to pass on all architectures all the time. This is due to the nature of the result checking, which relies on direct comparison, and most of the time, the failures are related to bad output checking, rather than bad code generation.

If the errors are in LLVM itself, please report every single regression found as blocker, and all the other bugs as important, but not necessarily blocking the release to proceed. They can be set as “known failures” and to be fix on a future date.

5.6.4 Pre-Release Process

When the release process is announced on the mailing list, you should prepare for the testing, by applying the same testing you'll do on the release candidates, on the previous release.

You should:

- Download the previous release sources from <http://llvm.org/releases/download.html>.
- Run the `test-release.sh` script on `final` mode (change `-rc 1` to `-final`).
- Once all three stages are done, it'll test the final stage.
- Using the `Phase3/Release+Asserts/llvmCore-MAJ.MIN-final.install` base, run the test-suite.

If the final phase's `make check-all` failed, it's a good idea to also test the intermediate stages by going on the `obj` directory and running `make check-all` to find if there's at least one stage that passes (helps when reducing the error for bug report purposes).

5.6.5 Release Process

When the Release Manager sends you the release candidate, download all sources, unzip on the same directory (there will be sym-links from the appropriate places to them), and run the release test as above.

You should:

- Download the current candidate sources from where the release manager points you (ex. <http://llvm.org/pre-releases/3.3/rc1/>).
- Repeat the steps above with `-rc 1`, `-rc 2` etc modes and run the test-suite the same way.

- Compare the results, report all errors on Bugzilla and publish the binary blob where the release manager can grab it.

Once the release manager announces that the latest candidate is the good one, you have to pack the Release (no Asserts) install directory on Phase3 and that will be the official binary.

- Rename (or link) `clang+llvm-REL-ARCH-ENV` to the `.install` directory
- Tar that into the same name with `.tar.gz` extension from outside the directory
- Make it available for the release manager to download

5.6.6 Bug Reporting Process

If you found regressions or failures when comparing a release candidate with the previous release, follow the rules below:

- Critical bugs on compilation should be fixed as soon as possible, possibly before releasing the binary blobs.
- Check-all tests should be fixed before the next release candidate, but can wait until the test-suite run is finished.
- Bugs in the test suite or unimportant check-all tests can be fixed in between release candidates.
- New features or recent big changes, when close to the release, should have done in a way that it's easy to disable. If they misbehave, prefer disabling them than releasing an unstable (but untested) binary package.

LLVM Developer Policy The LLVM project's policy towards developers and their contributions.

Creating an LLVM Project How-to guide and templates for new projects that *use* the LLVM infrastructure. The templates (directory organization, Makefiles, and test tree) allow the project code to be located outside (or inside) the `llvm/` tree, while using LLVM header files and libraries.

LLVMBuild Guide Describes the LLVMBuild organization and files used by LLVM to specify component descriptions.

How To Release LLVM To The Public This is a guide to preparing LLVM releases. Most developers can ignore it.

How To Validate a New Release This is a guide to validate a new release, during the release process. Most developers can ignore it.

Advice on Packaging LLVM Advice on packaging LLVM into a distribution.

Code Reviews with Phabricator Describes how to use the Phabricator code review tool hosted on <http://reviews.llvm.org/> and its command line interface, Arcanist.

LLVM has a thriving community of friendly and helpful developers. The two primary communication mechanisms in the LLVM community are mailing lists and IRC.

6.1 Mailing Lists

If you can't find what you need in these docs, try consulting the mailing lists.

Developer's List ([llvm-dev](#)) This list is for people who want to be included in technical discussions of LLVM. People post to this list when they have questions about writing code for or using the LLVM tools. It is relatively low volume.

Commits Archive ([llvm-commits](#)) This list contains all commit messages that are made when LLVM developers commit code changes to the repository. It also serves as a forum for patch review (i.e. send patches here). It is useful for those who want to stay on the bleeding edge of LLVM development. This list is very high volume.

Bugs & Patches Archive ([llvm-bugs](#)) This list gets emailed every time a bug is opened and closed. It is higher volume than the LLVM-dev list.

Test Results Archive ([llvm-testresults](#)) A message is automatically sent to this list by every active nightly tester when it completes. As such, this list gets email several times each day, making it a high volume list.

LLVM Announcements List ([llvm-announce](#)) This is a low volume list that provides important announcements regarding LLVM. It gets email about once a month.

6.2 IRC

Users and developers of the LLVM project (including subprojects such as Clang) can be found in #llvm on [irc.oftc.net](#). This channel has several bots.

- Buildbot reporters
 - llvmbb - Bot for the main LLVM buildbot master. <http://lab.llvm.org:8011/console>

- bb-chapuni - An individually run buildbot master. <http://bb.pgr.jp/console>
- smooshlab - Apple's internal buildbot master.
- robot - Bugzilla linker. %bug <number>
- clang-bot - A [geordi](#) instance running near-trunk clang instead of gcc.

6.3 Community wide proposals

Proposals for massive changes in how the community behaves and how the work flow can be better.

6.3.1 Moving LLVM Projects to GitHub

Table of Contents

- *Introduction*
- *What This Proposal is Not About*
- *Why Git, and Why GitHub?*
 - *Why Move At All?*
 - *Why Git?*
 - *Why GitHub?*
 - *On Managing Revision Numbers with Git*
 - *What About Branches and Merges?*
 - *What About Commit Emails?*
- *Straw Man Migration Plan*
 - *Step #1 : Before The Move*
 - *Step #2 : Git Move*
 - *Step #3: Write Access Move*
 - *Step #4 : Post Move*
- *One or Multiple Repositories?*
 - *Multirepo Variant*
 - * *Umbrella Repository*
 - * *Living Downstream*
 - * *Multirepo Preview*
 - * *Concerns*
 - * *Workflows*
 - *Monorepo Variant*
 - * *Building a single sub-project*
 - * *Read/write sub-project mirrors*

- * *Living Downstream*
- * *Monorepo Preview*
- * *Concerns*
- * *Workflows*
- *Multi/Mono Hybrid Variant*
 - * *Concerns*
- *Workflow Before/After*
 - *Checkout/Clone a Single Project, without Commit Access*
 - *Checkout/Clone a Single Project, with Commit Access*
 - * *Currently*
 - * *Multirepo Variant*
 - * *Monorepo Variant*
 - *Checkout/Clone Multiple Projects, with Commit Access*
 - * *Currently*
 - * *Multirepo Variant*
 - * *Monorepo Variant*
 - *Commit an API Change in LLVM and Update the Sub-projects*
 - *Branching/Stashing/Updating for Local Development or Experiments*
 - * *Currently*
 - * *Multirepo Variant*
 - * *Monorepo Variant*
 - *Bisecting*
 - * *Currently*
 - * *Multirepo Variant*
 - * *Monorepo Variant*
- *References*

Introduction

This is a proposal to move our current revision control system from our own hosted Subversion to GitHub. Below are the financial and technical arguments as to why we are proposing such a move and how people (and validation infrastructure) will continue to work with a Git-based LLVM.

There will be a survey pointing at this document which we'll use to gauge the community's reaction and, if we collectively decide to move, the time-frame. Be sure to make your view count.

Additionally, we will discuss this during a BoF at the next US LLVM Developer meeting (<http://llvm.org/devmtg/2016-11/>).

What This Proposal is *Not* About

Changing the development policy.

This proposal relates only to moving the hosting of our source-code repository from SVN hosted on our own servers to Git hosted on GitHub. We are not proposing using GitHub's issue tracker, pull-requests, or code-review.

Contributors will continue to earn commit access on demand under the Developer Policy, except that that a GitHub account will be required instead of SVN username/password-hash.

Why Git, and Why GitHub?

Why Move At All?

This discussion began because we currently host our own Subversion server and Git mirror on a voluntary basis. The LLVM Foundation sponsors the server and provides limited support, but there is only so much it can do.

Volunteers are not sysadmins themselves, but compiler engineers that happen to know a thing or two about hosting servers. We also don't have 24/7 support, and we sometimes wake up to see that continuous integration is broken because the SVN server is either down or unresponsive.

We should take advantage of one of the services out there (GitHub, GitLab, and BitBucket, among others) that offer better service (24/7 stability, disk space, Git server, code browsing, forking facilities, etc) for free.

Why Git?

Many new coders nowadays start with Git, and a lot of people have never used SVN, CVS, or anything else. Websites like GitHub have changed the landscape of open source contributions, reducing the cost of first contribution and fostering collaboration.

Git is also the version control many LLVM developers use. Despite the sources being stored in a SVN server, these developers are already using Git through the Git-SVN integration.

Git allows you to:

- Commit, squash, merge, and fork locally without touching the remote server.
- Maintain local branches, enabling multiple threads of development.
- Collaborate on these branches (e.g. through your own fork of llvm on GitHub).
- Inspect the repository history (blame, log, bisect) without Internet access.
- Maintain remote forks and branches on Git hosting services and integrate back to the main repository.

In addition, because Git seems to be replacing many OSS projects' version control systems, there are many tools that are built over Git. Future tooling may support Git first (if not only).

Why GitHub?

GitHub, like GitLab and BitBucket, provides free code hosting for open source projects. Any of these could replace the code-hosting infrastructure that we have today.

These services also have a dedicated team to monitor, migrate, improve and distribute the contents of the repositories depending on region and load.

GitHub has one important advantage over GitLab and BitBucket: it offers read-write **SVN** access to the repository (<https://github.com/blog/626-announcing-svn-support>). This would enable people to continue working post-migration as though our code were still canonically in an SVN repository.

In addition, there are already multiple LLVM mirrors on GitHub, indicating that part of our community has already settled there.

On Managing Revision Numbers with Git

The current SVN repository hosts all the LLVM sub-projects alongside each other. A single revision number (e.g. r123456) thus identifies a consistent version of all LLVM sub-projects.

Git does not use sequential integer revision number but instead uses a hash to identify each commit. (Linus mentioned that the lack of such revision number is “the only real design mistake” in Git [[TorvaldRevNum](#)].)

The loss of a sequential integer revision number has been a sticking point in past discussions about Git:

- “The ‘branch’ I most care about is mainline, and losing the ability to say ‘fixed in r1234’ (with some sort of monotonically increasing number) would be a tragic loss.” [[LattnerRevNum](#)]
- “I like those results sorted by time and the chronology should be obvious, but timestamps are incredibly cumbersome and make it difficult to verify that a given checkout matches a given set of results.” [[TrickRevNum](#)]
- “There is still the major regression with unreadable version numbers. Given the amount of Bugzilla traffic with ‘Fixed in...’, that’s a non-trivial issue.” [[JSonnRevNum](#)]
- “Sequential IDs are important for LNT and llvmlab bisection tool.” [[MatthewsRevNum](#)].

However, Git can emulate this increasing revision number: `git rev-list --count <commit-hash>`. This identifier is unique only within a single branch, but this means the tuple (*num*, *branch-name*) uniquely identifies a commit.

We can thus use this revision number to ensure that e.g. `clang -v` reports a user-friendly revision number (e.g. *master-12345* or *4.0-5321*), addressing the objections raised above with respect to this aspect of Git.

What About Branches and Merges?

In contrast to SVN, Git makes branching easy. Git’s commit history is represented as a DAG, a departure from SVN’s linear history. However, we propose to mandate making merge commits illegal in our canonical Git repository.

Unfortunately, GitHub does not support server side hooks to enforce such a policy. We must rely on the community to avoid pushing merge commits.

GitHub offers a feature called *Status Checks*: a branch protected by *status checks* requires commits to be whitelisted before the push can happen. We could supply a pre-push hook on the client side that would run and check the history, before whitelisting the commit being pushed [[statuschecks](#)]. However this solution would be somewhat fragile (how do you update a script installed on every developer machine?) and prevents SVN access to the repository.

What About Commit Emails?

We will need a new bot to send emails for each commit. This proposal leaves the email format unchanged besides the commit URL.

Straw Man Migration Plan

Step #1 : Before The Move

1. Update docs to mention the move, so people are aware of what is going on.
2. Set up a read-only version of the GitHub project, mirroring our current SVN repository.
3. Add the required bots to implement the commit emails, as well as the umbrella repository update (if the multi-repo is selected) or the read-only Git views for the sub-projects (if the monorepo is selected).

Step #2 : Git Move

4. Update the buildbots to pick up updates and commits from the GitHub repository. Not all bots have to migrate at this point, but it'll help provide infrastructure testing.
5. Update Phabricator to pick up commits from the GitHub repository.
6. LNT and llvmlab have to be updated: they rely on unique monotonically increasing integer across branch *[MatthewsRevNum]*.
7. Instruct downstream integrators to pick up commits from the GitHub repository.
8. Review and prepare an update for the LLVM documentation.

Until this point nothing has changed for developers, it will just boil down to a lot of work for buildbot and other infrastructure owners.

The migration will pause here until all dependencies have cleared, and all problems have been solved.

Step #3: Write Access Move

9. Collect developers' GitHub account information, and add them to the project.
10. Switch the SVN repository to read-only and allow pushes to the GitHub repository.
11. Update the documentation.
12. Mirror Git to SVN.

Step #4 : Post Move

13. Archive the SVN repository.
14. Update links on the LLVM website pointing to viewvc/klaus/phab etc. to point to GitHub instead.

One or Multiple Repositories?

There are two major variants for how to structure our Git repository: The “multirepo” and the “monorepo”.

Multirepo Variant

This variant recommends moving each LLVM sub-project to a separate Git repository. This mimics the existing official read-only Git repositories (e.g., <http://llvm.org/git/compiler-rt.git>), and creates new canonical repositories for each sub-project.

This will allow the individual sub-projects to remain distinct: a developer interested only in compiler-rt can checkout only this repository, build it, and work in isolation of the other sub-projects.

A key need is to be able to check out multiple projects (i.e. lldb+clang+llvm or clang+llvm+libcxx for example) at a specific revision.

A tuple of revisions (one entry per repository) accurately describes the state across the sub-projects. For example, a given version of clang would be `<LLVM-12345, clang-5432, libcxx-123, etc.>`.

Umbrella Repository

To make this more convenient, a separate *umbrella* repository will be provided. This repository will be used for the sole purpose of understanding the sequence in which commits were pushed to the different repositories and to provide a single revision number.

This umbrella repository will be read-only and continuously updated to record the above tuple. The proposed form to record this is to use Git [\[submodules\]](#), possibly along with a set of scripts to help check out a specific revision of the LLVM distribution.

A regular LLVM developer does not need to interact with the umbrella repository – the individual repositories can be checked out independently – but you would need to use the umbrella repository to bisect multiple sub-projects at the same time, or to check-out old revisions of LLVM with another sub-project at a consistent state.

This umbrella repository will be updated automatically by a bot (running on notice from a webhook on every push, and periodically) on a per commit basis: a single commit in the umbrella repository would match a single commit in a sub-project.

Living Downstream

Downstream SVN users can use the read/write SVN bridges with the following caveats:

- Be prepared for a one-time change to the upstream revision numbers.
- The upstream sub-project revision numbers will no longer be in sync.

Downstream Git users can continue without any major changes, with the minor change of upstreaming using *git push* instead of *git svn dcommit*.

Git users also have the option of adopting an umbrella repository downstream. The tooling for the upstream umbrella can easily be reused for downstream needs, incorporating extra sub-projects and branching in parallel with sub-project branches.

Multirepo Preview

As a preview (disclaimer: this rough prototype, not polished and not representative of the final solution), you can look at the following:

- Repository: <https://github.com/llvm-beanz/llvm-submodules>
- Update bot: <http://beanz-bot.com:8180/jenkins/job/submodule-update/>

Concerns

- Because GitHub does not allow server-side hooks, and because there is no “push timestamp” in Git, the umbrella repository sequence isn’t totally exact: commits from different repositories pushed around the same time can appear in different orders. However, we don’t expect it to be the common case or to cause serious issues in practice.
- You can’t have a single cross-projects commit that would update both LLVM and other sub-projects (something that can be achieved now). It would be possible to establish a protocol whereby users add a special token to their commit messages that causes the umbrella repo’s updater bot to group all of them into a single revision.
- Another option is to group commits that were pushed closely enough together in the umbrella repository. This has the advantage of allowing cross-project commits, and is less sensitive to mis-ordering commits. However, this has the potential to group unrelated commits together, especially if the bot goes down and needs to catch up.
- This variant relies on heavier tooling. But the current prototype shows that it is not out-of-reach.
- Submodules don’t have a good reputation / are complicating the command line. However, in the proposed setup, a regular developer will seldom interact with submodules directly, and certainly never update them.
- Refactoring across projects is not friendly: taking some functions from clang to make it part of a utility in libSupport wouldn’t carry the history of the code in the llvm repo, preventing recursively applying *git blame* for instance. However, this is not very different than how most people are interacting with the repository today, by splitting such change in multiple commits.

Workflows

- *Checkout/Clone a Single Project, without Commit Access.*
- *Checkout/Clone a Single Project, with Commit Access.*
- *Checkout/Clone Multiple Projects, with Commit Access.*
- *Commit an API Change in LLVM and Update the Sub-projects.*
- *Branching/Stashing/Updating for Local Development or Experiments.*
- *Bisecting.*

Monorepo Variant

This variant recommends moving all LLVM sub-projects to a single Git repository, similar to <https://github.com/llvm-project/llvm-project>. This would mimic an export of the current SVN repository, with each sub-project having its own top-level directory. Not all sub-projects are used for building toolchains. In practice, `www/` and `test-suite/` will probably stay out of the monorepo.

Putting all sub-projects in a single checkout makes cross-project refactoring naturally simple:

- New sub-projects can be trivially split out for better reuse and/or layering (e.g., to allow libSupport and/or LIT to be used by runtimes without adding a dependency on LLVM).
- Changing an API in LLVM and upgrading the sub-projects will always be done in a single commit, designing away a common source of temporary build breakage.
- Moving code across sub-project (during refactoring for instance) in a single commit enables accurate *git blame* when tracking code change history.

- Tooling based on *git grep* works natively across sub-projects, allowing to easier find refactoring opportunities across projects (for example reusing a datastructure initially in LLDB by moving it into libSupport).
- Having all the sources present encourages maintaining the other sub-projects when changing API.

Finally, the monorepo maintains the property of the existing SVN repository that the sub-projects move synchronously, and a single revision number (or commit hash) identifies the state of the development across all projects.

Building a single sub-project

Nobody will be forced to build unnecessary projects. The exact structure is TBD, but making it trivial to configure builds for a single sub-project (or a subset of sub-projects) is a hard requirement.

As an example, it could look like the following:

```
mkdir build && cd build
# Configure only LLVM (default)
cmake path/to/monorepo
# Configure LLVM and lld
cmake path/to/monorepo -DLLVM_ENABLE_PROJECTS=lld
# Configure LLVM and clang
cmake path/to/monorepo -DLLVM_ENABLE_PROJECTS=clang
```

Read/write sub-project mirrors

With the Monorepo, the existing single-subproject mirrors (e.g. <http://llvm.org/git/compiler-rt.git>) with git-svn read-write access would continue to be maintained: developers would continue to be able to use the existing single-subproject git repositories as they do today, with *no changes to workflow*. Everything (git fetch, git svn dcommit, etc.) could continue to work identically to how it works today. The monorepo can be set-up such that the SVN revision number matches the SVN revision in the GitHub SVN-bridge.

Living Downstream

Downstream SVN users can use the read/write SVN bridge. The SVN revision number can be preserved in the monorepo, minimizing the impact.

Downstream Git users can continue without any major changes, by using the git-svn mirrors on top of the SVN bridge.

Git users can also work upstream with monorepo even if their downstream fork has split repositories. They can apply patches in the appropriate subdirectories of the monorepo using, e.g., *git am -directory=...*, or plain *diff* and *patch*.

Alternatively, Git users can migrate their own fork to the monorepo. As a demonstration, we've migrated the "CHERI" fork to the monorepo in two ways:

- Using a script that rewrites history (including merges) so that it looks like the fork always lived in the monorepo [[LebarCHERI](#)]. The upside of this is when you check out an old revision, you get a copy of all llvm sub-projects at a consistent revision. (For instance, if it's a clang fork, when you check out an old revision you'll get a consistent version of llvm proper.) The downside is that this changes the fork's commit hashes.
- Merging the fork into the monorepo [[AminiCHERI](#)]. This preserves the fork's commit hashes, but when you check out an old commit you only get the one sub-project.

Monorepo Preview

As a preview (disclaimer: this rough prototype, not polished and not representative of the final solution), you can look at the following:

- Full Repository: <https://github.com/joker-eph/llvm-project>
- Single sub-project view with *SVN write access* to the full repo: <https://github.com/joker-eph/compiler-rt>

Concerns

- Using the monolithic repository may add overhead for those contributing to a standalone sub-project, particularly on runtimes like `libcxx` and `compiler-rt` that don't rely on LLVM; currently, a fresh clone of `libcxx` is only 15MB (vs. 1GB for the monorepo), and the commit rate of LLVM may cause more frequent *git push* collisions when upstreaming. Affected contributors can continue to use the SVN bridge or the single-subproject Git mirrors with `git-svn` for read-write.
- Using the monolithic repository may add overhead for those *integrating* a standalone sub-project, even if they aren't contributing to it, due to the same disk space concern as the point above. The availability of the sub-project Git mirror addresses this, even without SVN access.
- Preservation of the existing read/write SVN-based workflows relies on the GitHub SVN bridge, which is an extra dependency. Maintaining this locks us into GitHub and could restrict future workflow changes.

Workflows

- *Checkout/Clone a Single Project, without Commit Access.*
- *Checkout/Clone a Single Project, with Commit Access.*
- *Checkout/Clone Multiple Projects, with Commit Access.*
- *Commit an API Change in LLVM and Update the Sub-projects.*
- *Branching/Stashing/Updating for Local Development or Experiments.*
- *Bisecting.*

Multi/Mono Hybrid Variant

This variant recommends moving only the LLVM sub-projects that are *rev-locked* to LLVM into a monorepo (`clang`, `lld`, `lldb`, ...), following the multirepo proposal for the rest. While neither variant recommends combining sub-projects like `www/` and `test-suite/` (which are completely standalone), this goes further and keeps sub-projects like `libcxx` and `compiler-rt` in their own distinct repositories.

Concerns

- This has most disadvantages of multirepo and monorepo, without bringing many of the advantages.
- Downstream have to upgrade to the monorepo structure, but only partially. So they will keep the infrastructure to integrate the other separate sub-projects.
- All projects that use LIT for testing are effectively rev-locked to LLVM. Furthermore, some runtimes (like `compiler-rt`) are rev-locked with Clang. It's not clear where to draw the lines.

Workflow Before/After

This section goes through a few examples of workflows, intended to illustrate how end-users or developers would interact with the repository for various use-cases.

Checkout/Clone a Single Project, without Commit Access

Except the URL, nothing changes. The possibilities today are:

```
svn co http://llvm.org/svn/llvm-project/llvm/trunk llvm
# or with Git
git clone http://llvm.org/git/llvm.git
```

After the move to GitHub, you would do either:

```
git clone https://github.com/llvm-project/llvm.git
# or using the GitHub svn native bridge
svn co https://github.com/llvm-project/llvm/trunk
```

The above works for both the monorepo and the multirepo, as we'll maintain the existing read-only views of the individual sub-projects.

Checkout/Clone a Single Project, with Commit Access

Currently

```
# direct SVN checkout
svn co https://user@llvm.org/svn/llvm-project/llvm/trunk llvm
# or using the read-only Git view, with git-svn
git clone http://llvm.org/git/llvm.git
cd llvm
git svn init https://llvm.org/svn/llvm-project/llvm/trunk --username=<username>
git config svn-remote.svn.fetch :refs/remotes/origin/master
git svn rebase -l # -l avoids fetching ahead of the git mirror.
```

Commits are performed using *svn commit* or with the sequence *git commit* and *git svn dcommit*.

Multirepo Variant

With the multirepo variant, nothing changes but the URL, and commits can be performed using *svn commit* or *git commit* and *git push*:

```
git clone https://github.com/llvm/llvm.git llvm
# or using the GitHub svn native bridge
svn co https://github.com/llvm/llvm/trunk/ llvm
```

Monorepo Variant

With the monorepo variant, there are a few options, depending on your constraints. First, you could just clone the full repository:

```
git clone https://github.com/llvm/llvm-projects.git llvm
# or using the GitHub svn native bridge
svn co https://github.com/llvm/llvm-projects/trunk/ llvm
```

At this point you have every sub-project (llvm, clang, lld, lldb, ...), which *doesn't imply you have to build all of them*. You can still build only compiler-rt for instance. In this way it's not different from someone who would check out all the projects with SVN today.

You can commit as normal using *git commit* and *git push* or *svn commit*, and read the history for a single project (*git log libcxx* for example).

Secondly, there are a few options to avoid checking out all the sources.

Using the GitHub SVN bridge

The GitHub SVN native bridge allows to checkout a subdirectory directly:

```
svn co https://github.com/llvm/llvm-projects/trunk/compiler-rt compiler-rt --username=...
```

This checks out only compiler-rt and provides commit access using “svn commit”, in the same way as it would do today.

Using a Subproject Git Mirror

You can use *git-svn* and one of the sub-project mirrors:

```
# Clone from the single read-only Git repo
git clone http://llvm.org/git/llvm.git
cd llvm
# Configure the SVN remote and initialize the svn metadata
$ git svn init https://github.com/joker-eph/llvm-project/trunk/llvm --username=...
git config svn-remote.svn.fetch :refs/remotes/origin/master
git svn rebase -l
```

In this case the repository contains only a single sub-project, and commits can be made using *git svn dcommit*, again exactly as we do today.

Using a Sparse Checkouts

You can hide the other directories using a Git sparse checkout:

```
git config core.sparseCheckout true
echo /compiler-rt > .git/info/sparse-checkout
git read-tree -mu HEAD
```

The data for all sub-projects is still in your *.git* directory, but in your checkout, you only see *compiler-rt*. Before you push, you'll need to fetch and rebase (*git pull --rebase*) as usual.

Note that when you fetch you'll likely pull in changes to sub-projects you don't care about. If you are using sparse checkout, the files from other projects won't appear on your disk. The only effect is that your commit hash changes.

You can check whether the changes in the last fetch are relevant to your commit by running:

```
git log origin/master@{1}..origin/master -- libcxx
```

This command can be hidden in a script so that *git llvmpush* would perform all these steps, fail only if such a dependent change exists, and show immediately the change that prevented the push. An immediate repeat of the command would (almost) certainly result in a successful push. Note that today with SVN or git-svn, this step is not possible since the “rebase” implicitly happens while committing (unless a conflict occurs).

Checkout/Clone Multiple Projects, with Commit Access

Let's look how to assemble llvm+clang+libcxx at a given revision.

Currently

```
svn co http://llvm.org/svn/llvm-project/llvm/trunk llvm -r $REVISION
cd llvm/tools
svn co http://llvm.org/svn/llvm-project/clang/trunk clang -r $REVISION
cd ../projects
svn co http://llvm.org/svn/llvm-project/libcxx/trunk libcxx -r $REVISION
```

Or using git-svn:

```
git clone http://llvm.org/git/llvm.git
cd llvm/
git svn init https://llvm.org/svn/llvm-project/llvm/trunk --username=<username>
git config svn-remote.svn.fetch :refs/remotes/origin/master
git svn rebase -l
git checkout `git svn find-rev -B r258109`
cd tools
git clone http://llvm.org/git/clang.git
cd clang/
git svn init https://llvm.org/svn/llvm-project/clang/trunk --username=<username>
git config svn-remote.svn.fetch :refs/remotes/origin/master
git svn rebase -l
git checkout `git svn find-rev -B r258109`
cd ../../projects/
git clone http://llvm.org/git/libcxx.git
cd libcxx
git svn init https://llvm.org/svn/llvm-project/libcxx/trunk --username=<username>
git config svn-remote.svn.fetch :refs/remotes/origin/master
git svn rebase -l
git checkout `git svn find-rev -B r258109`
```

Note that the list would be longer with more sub-projects.

Multirepo Variant

With the multirepo variant, the umbrella repository will be used. This is where the mapping from a single revision number to the individual repositories revisions is stored.:

```
git clone https://github.com/llvm-beanz/llvm-submodules
cd llvm-submodules
git checkout $REVISION
git submodule init
git submodule update clang llvm libcxx
# the list of sub-project is optional, `git submodule update` would get them all.
```

At this point the clang, llvm, and libcxx individual repositories are cloned and stored alongside each other. There are CMake flags to describe the directory structure; alternatively, you can just symlink *clang* to *llvm/tools/clang*, etc.

Another option is to checkout repositories based on the commit timestamp:

```
git checkout `git rev-list -n 1 --before="2009-07-27 13:37" master`
```

Monorepo Variant

The repository contains natively the source for every sub-projects at the right revision, which makes this straightforward:

```
git clone https://github.com/llvm/llvm-projects.git llvm-projects
cd llvm-projects
git checkout $REVISION
```

As before, at this point clang, llvm, and libcxx are stored in directories alongside each other.

Commit an API Change in LLVM and Update the Sub-projects

Today this is possible, even though not common (at least not documented) for subversion users and for git-svn users. For example, few Git users try to update LLD or Clang in the same commit as they change an LLVM API.

The multirepo variant does not address this: one would have to commit and push separately in every individual repository. It would be possible to establish a protocol whereby users add a special token to their commit messages that causes the umbrella repo's updater bot to group all of them into a single revision.

The monorepo variant handles this natively.

Branching/Stashing/Updating for Local Development or Experiments

Currently

SVN does not allow this use case, but developers that are currently using git-svn can do it. Let's look in practice what it means when dealing with multiple sub-projects.

To update the repository to tip of trunk:

```
git pull
cd tools/clang
git pull
cd ../../projects/libcxx
git pull
```

To create a new branch:

```
git checkout -b MyBranch
cd tools/clang
git checkout -b MyBranch
cd ../../projects/libcxx
git checkout -b MyBranch
```

To switch branches:

```
git checkout AnotherBranch
cd tools/clang
git checkout AnotherBranch
```

```
cd ../../projects/libcxx
git checkout AnotherBranch
```

Multirepo Variant

The multirepo works the same as the current Git workflow: every command needs to be applied to each of the individual repositories. However, the umbrella repository makes this easy using *git submodule foreach* to replicate a command on all the individual repositories (or submodules in this case):

To create a new branch:

```
git submodule foreach git checkout -b MyBranch
```

To switch branches:

```
git submodule foreach git checkout AnotherBranch
```

Monorepo Variant

Regular Git commands are sufficient, because everything is in a single repository:

To update the repository to tip of trunk:

```
git pull
```

To create a new branch:

```
git checkout -b MyBranch
```

To switch branches:

```
git checkout AnotherBranch
```

Bisecting

Assuming a developer is looking for a bug in clang (or lld, or lldb, ...).

Currently

SVN does not have builtin bisection support, but the single revision across sub-projects makes it possible to script around.

Using the existing Git read-only view of the repositories, it is possible to use the native Git bisection script over the llvm repository, and use some scripting to synchronize the clang repository to match the llvm revision.

Multirepo Variant

With the multi-repositories variant, the cross-repository synchronization is achieved using the umbrella repository. This repository contains only submodules for the other sub-projects. The native Git bisection can be used on the

umbrella repository directly. A subtlety is that the bisect script itself needs to make sure the submodules are updated accordingly.

For example, to find which commit introduces a regression where clang-3.9 crashes but not clang-3.8 passes, one should be able to simply do:

```
git bisect start release_39 release_38
git bisect run ./bisect_script.sh
```

With the *bisect_script.sh* script being:

```
#!/bin/sh
cd $UMBRELLA_DIRECTORY
git submodule update llvm clang libcxx #....
cd $BUILD_DIR

ninja clang || exit 125    # an exit code of 125 asks "git bisect"
                          # to "skip" the current commit

./bin/clang some_crash_test.cpp
```

When the *git bisect run* command returns, the umbrella repository is set to the state where the regression is introduced. The commit diff in the umbrella indicate which submodule was updated, and the last commit in this sub-projects is the one that the bisect found.

Monorepo Variant

Bisecting on the monorepo is straightforward, and very similar to the above, except that the bisection script does not need to include the *git submodule update* step.

The same example, finding which commit introduces a regression where clang-3.9 crashes but not clang-3.8 passes, will look like:

```
git bisect start release_39 release_38
git bisect run ./bisect_script.sh
```

With the *bisect_script.sh* script being:

```
#!/bin/sh
cd $BUILD_DIR

ninja clang || exit 125    # an exit code of 125 asks "git bisect"
                          # to "skip" the current commit

./bin/clang some_crash_test.cpp
```

Also, since the monorepo handles commits update across multiple projects, you're less like to encounter a build failure where a commit change an API in LLVM and another later one "fixes" the build in clang.

References

LLVM Community Code of Conduct Proposal to adopt a code of conduct on the LLVM social spaces (lists, events, IRC, etc).

Moving LLVM Projects to GitHub Proposal to move from SVN/Git to GitHub.

Vectorization Plan Proposal to model the process and upgrade the infrastructure of LLVM's Loop Vectorizer.

CHAPTER 7

Indices and tables

- `genindex`
- `search`

Bibliography

- [AMD-R6xx] AMD R6xx shader ISA
- [AMD-R7xx] AMD R7xx shader ISA
- [AMD-Evergreen] AMD Evergreen shader ISA
- [AMD-Cayman-Trinity] AMD Cayman/Trinity shader ISA
- [AMD-Souther-Islands] AMD Southern Islands Series ISA
- [AMD-Sea-Islands] AMD Sea Islands Series ISA
- [AMD-Volcanic-Islands] AMD GCN3 Instruction Set Architecture
- [AMD-OpenCL_Programming-Guide] AMD Accelerated Parallel Processing OpenCL Programming Guide
- [AMD-APP-SDK] AMD Accelerated Parallel Processing APP SDK Documentation
- [AMD-ROCm] ROCm: Open Platform for Development, Discovery and Education Around GPU Computing
- [AMD-ROCm-github] ROCm github
- [HSA] Heterogeneous System Architecture (HSA) Foundation
- [ELF] Executable and Linkable Format (ELF)
- [DWARF] DWARF Debugging Information Format
- [YAML] YAML Ain't Markup Language (YAML™) Version 1.2
- [OpenCL] The OpenCL Specification Version 2.0
- [HRF] Heterogeneous-race-free Memory Models
- [AMD-AMDGPU-Compute-Application-Binary-Interface] AMDGPU Compute Application Binary Interface
- [LattnerRevNum] Chris Lattner, <http://lists.llvm.org/pipermail/llvm-dev/2011-July/041739.html>
- [TrickRevNum] Andrew Trick, <http://lists.llvm.org/pipermail/llvm-dev/2011-July/041721.html>
- [JSonnRevNum] Joerg Sonnenberg, <http://lists.llvm.org/pipermail/llvm-dev/2011-July/041688.html>
- [TorvaldRevNum] Linus Torvald, <http://git.661346.n2.nabble.com/Git-commit-generation-numbers-td6584414.html>
- [MatthewsRevNum] Chris Matthews, <http://lists.llvm.org/pipermail/cfe-dev/2016-July/049886.html>
- [submodules] Git submodules, <https://git-scm.com/book/en/v2/Git-Tools-Submodules>

[statuschecks] GitHub status-checks, <https://help.github.com/articles/about-required-status-checks/>

[LebarCHERI] Port *CHERI* to a single repository rewriting history, <http://lists.llvm.org/pipermail/llvm-dev/2016-July/102787.html>

[AminiCHERI] Port *CHERI* to a single repository preserving history, <http://lists.llvm.org/pipermail/llvm-dev/2016-July/102804.html>

Symbols

- check-prefix prefix
 - command line option, 301
- check-prefixes prefix1,prefix2,...
 - command line option, 301
- config-prefix=NAME
 - command line option, 310
- debug
 - command line option, 312
- debug-syms, -a
 - llvm-nm command line option, 279
- defined-only
 - llvm-nm command line option, 279
- disable-excess-fp-precision
 - command line option, 269
- disable-fp-elim
 - command line option, 269
- dynamic, -D
 - llvm-nm command line option, 279
- enable-no-infs-fp-math
 - command line option, 269
- enable-no-nans-fp-math
 - command line option, 269
- enable-unsafe-fp-math
 - command line option, 269
- enable-var-scope
 - command line option, 302
- extern-only, -g
 - llvm-nm command line option, 279
- filter=REGEXP
 - command line option, 311
- format=format, -f format
 - llvm-nm command line option, 279
- help
 - llvm-cov-gcov command line option, 284
- implicit-check-not check-pattern
 - command line option, 302
- input-file filename
 - command line option, 301
- load=<dso_path>
 - command line option, 269
- match-full-lines
 - command line option, 302
- max-tests=N
 - command line option, 311
- max-time=N
 - command line option, 311
- no-progress-bar
 - command line option, 310
- no-sort, -p
 - llvm-nm command line option, 279
- num-shards=M
 - command line option, 311
- numeric-sort, -n, -v
 - llvm-nm command line option, 279
- path=PATH
 - command line option, 311
- print-file-name, -A, -o
 - llvm-nm command line option, 279
- print-machineinstrs
 - command line option, 269
- print-size, -S
 - llvm-nm command line option, 280
- radix=RADIX, -t
 - llvm-nm command line option, 280
- regalloc=<allocator>
 - command line option, 269
- run-shard=N
 - command line option, 311
- show-suites
 - command line option, 312
- show-tests
 - command line option, 312
- show-unsupported
 - command line option, 310
- show-xfail
 - command line option, 310
- shuffle
 - command line option, 311

- size-sort
 - llvm-nm command line option, 280
- spiller=<spiller>
 - command line option, 270
- stats
 - command line option, 269
- strict-whitespace
 - command line option, 302
- time-passes
 - command line option, 269
- time-tests
 - command line option, 311
- undefined-only, -u
 - llvm-nm command line option, 280
- vg
 - command line option, 311
- vg-arg=ARG
 - command line option, 311
- vg-leak
 - command line option, 311
- x86-asm-syntax=[attintel]
 - command line option, 270
- B (default)
 - llvm-nm command line option, 279
- D NAME[=VALUE], -param NAME[=VALUE]
 - command line option, 310
- I directory
 - tblgen command line option, 308
- O=uint
 - command line option, 268
- P
 - llvm-nm command line option, 279
- S
 - command line option, 267, 273
- Xdemangler=<TOOL>|<TOOL-OPTION>
 - llvm-cov-show command line option, 286
- {passname}
 - command line option, 267
- a, -all-blocks
 - llvm-cov-gcov command line option, 284
- a, -show-all
 - command line option, 310
- all-functions
 - llvm-profdata-show command line option, 289
- arch=<name>
 - llvm-cov-export command line option, 287
 - llvm-cov-report command line option, 287
 - llvm-cov-show command line option, 285
- asmparsernum N
 - tblgen command line option, 308
- asmwriternum N
 - tblgen command line option, 308
- b, -branch-probabilities
 - llvm-cov-gcov command line option, 284
- binary (default)
 - llvm-profdata-merge command line option, 288
- c, -branch-counts
 - llvm-cov-gcov command line option, 284
- class className
 - tblgen command line option, 308
- code-model=model
 - command line option, 272
- counts
 - llvm-profdata-show command line option, 289
- d
 - command line option, 273
- debug
 - command line option, 267
- debug-dump=section
 - command line option, 293
- default-arch
 - command line option, 292
- demangle
 - command line option, 292
- disable-excess-fp-precision
 - command line option, 271
- disable-inlining
 - command line option, 267
- disable-opt
 - command line option, 267
- disable-post-RA-scheduler
 - command line option, 272
- disable-spill-fusing
 - command line option, 272
- dsym-hint=<path/to/file.dSYM>
 - command line option, 292
- dump
 - llvm-bcanalyzer command line option, 297
- dyn-symbols
 - command line option, 317
- dynamic-table
 - command line option, 317
- elf-section-groups, -g
 - command line option, 318
- enable-no-infs-fp-math
 - command line option, 271
- enable-no-nans-fp-math
 - command line option, 271
- enable-unsafe-fp-math
 - command line option, 271
- expand-relocs
 - command line option, 317
- f
 - command line option, 267, 273
- f, -function-summaries
 - llvm-cov-gcov command line option, 284
- fake-argv0=executable
 - command line option, 271

- file-headers, -h
 - command line option, 317
- filetype=<output file type>
 - command line option, 268
- force-interpret={ false,true}
 - command line option, 271
- format=<FORMAT>
 - llvm-cov-show command line option, 286
- function=string
 - llvm-profdata-show command line option, 289
- functions=[none|short|linkage]
 - command line option, 292
- gcc
 - llvm-profdata-merge command line option, 288
- gen-asm-matcher
 - tblgen command line option, 309
- gen-asm-writer
 - tblgen command line option, 309
- gen-dag-isel
 - tblgen command line option, 309
- gen-dfa-packetizer
 - tblgen command line option, 309
- gen-disassembler
 - tblgen command line option, 309
- gen-emitter
 - tblgen command line option, 308
- gen-enhanced-disassembly-info
 - tblgen command line option, 309
- gen-fast-isel
 - tblgen command line option, 309
- gen-instr-info
 - tblgen command line option, 309
- gen-intrinsic
 - tblgen command line option, 309
- gen-pseudo-lowering
 - tblgen command line option, 309
- gen-register-info
 - tblgen command line option, 308
- gen-subtarget
 - tblgen command line option, 309
- gen-tgt-intrinsic
 - tblgen command line option, 309
- h, -help
 - command line option, 310
- help
 - command line option, 267, 268, 271, 273, 301, 317
 - llvm-bcanalyzer command line option, 297
 - llvm-nm command line option, 279
 - llvm-profdata-merge command line option, 288
 - llvm-profdata-show command line option, 290
 - tblgen command line option, 308
- inlining
 - command line option, 292
- input-files=path, -f=path
 - llvm-profdata-merge command line option, 288
- instr (default)
 - llvm-profdata-merge command line option, 288
 - llvm-profdata-show command line option, 290
- j N, -threads=N
 - command line option, 310
- jit-enable-eh
 - command line option, 272
- join-liveintervals
 - command line option, 272
- l, -long-file-names
 - llvm-cov-gcov command line option, 284
- line-coverage-gt=<N>
 - llvm-cov-show command line option, 286
- line-coverage-lt=<N>
 - llvm-cov-show command line option, 286
- load=<plugin>
 - command line option, 267
- load=pluginfilename
 - command line option, 271
- march=<arch>
 - command line option, 268
- march=arch
 - command line option, 271
- mattr=a1,+a2,-a3,...
 - command line option, 269, 271
- mcpu=<cpu name>
 - command line option, 268
- mcpu=cpu name
 - command line option, 271
- meabi=[default|gnul4|5]
 - command line option, 269
- memop-sizes
 - llvm-profdata-show command line option, 290
- mtriple=<target triple>
 - command line option, 268
- mtriple=target triple
 - command line option, 271
- n, -no-output
 - llvm-cov-gcov command line option, 284
- name-regex=<PATTERN>
 - llvm-cov-show command line option, 286
- name=<NAME>
 - llvm-cov-show command line option, 286
- needed-libs
 - command line option, 317
- nodetails
 - llvm-bcanalyzer command line option, 297
- nozero-initialized-in-bss
 - command line option, 272
- num-threads=N, -j=N
 - llvm-cov-show command line option, 286
 - llvm-profdata-merge command line option, 289
- o <filename>

- command line option, [267](#)
- o filename
 - command line option, [273](#), [290](#)
 - tblgen command line option, [308](#)
- o=<DIR|FILE>, --object-directory=<DIR>, --object-file=<FILE>
 - llvm-cov-gcov command line option, [284](#)
- obj
 - command line option, [292](#)
- output-dir=PATH
 - llvm-cov-show command line option, [286](#)
- output=output, -o=output
 - llvm-profdata-merge command line option, [288](#)
 - llvm-profdata-show command line option, [290](#)
- p
 - command line option, [267](#)
- p, --preserve-paths
 - llvm-cov-gcov command line option, [284](#)
- pre-RA-sched=scheduler
 - command line option, [272](#)
- pretty-print
 - command line option, [292](#)
- print-address
 - command line option, [292](#)
- print-enums
 - tblgen command line option, [308](#)
- print-records
 - tblgen command line option, [308](#)
- print-sets
 - tblgen command line option, [308](#)
- program-headers
 - command line option, [317](#)
- q, --quiet
 - command line option, [310](#)
- regalloc=allocator
 - command line option, [272](#)
- region-coverage-gt=<N>
 - llvm-cov-show command line option, [286](#)
- region-coverage-lt=<N>
 - llvm-cov-show command line option, [286](#)
- relocation-model=model
 - command line option, [272](#)
- relocations, -r
 - command line option, [317](#)
- s, --succinct
 - command line option, [310](#)
- sample
 - llvm-profdata-merge command line option, [288](#)
 - llvm-profdata-show command line option, [290](#)
- section-data, -sd
 - command line option, [317](#)
- section-relocations, -sr
 - command line option, [317](#)
- section-symbols, -st
 - command line option, [317](#)
- sections, -s
 - command line option, [317](#)
- seed seed
 - command line option, [290](#)
- show-expansions
 - llvm-cov-show command line option, [285](#)
- show-functions
 - llvm-cov-report command line option, [287](#)
- show-instantiations
 - llvm-cov-show command line option, [285](#)
- show-line-counts
 - llvm-cov-show command line option, [285](#)
- show-line-counts-or-regions
 - llvm-cov-show command line option, [285](#)
- show-regions
 - llvm-cov-show command line option, [285](#)
- size size
 - command line option, [290](#)
- soft-float
 - command line option, [271](#)
- sparse[=true|false]
 - llvm-profdata-merge command line option, [289](#)
- spiller
 - command line option, [272](#)
- stats
 - command line option, [267](#), [271](#)
- strip-debug
 - command line option, [267](#)
- symbols, -t
 - command line option, [317](#)
- tab-size=<TABSIZ>
 - llvm-cov-show command line option, [286](#)
- text
 - llvm-profdata-merge command line option, [288](#)
 - llvm-profdata-show command line option, [290](#)
- time-passes
 - command line option, [267](#), [271](#)
- topn=n
 - llvm-profdata-show command line option, [290](#)
- u, --unconditional-branches
 - llvm-cov-gcov command line option, [284](#)
- unwind, -u
 - command line option, [317](#)
- use-color[=VALUE]
 - llvm-cov-report command line option, [287](#)
 - llvm-cov-show command line option, [285](#)
- use-symbol-table
 - command line option, [292](#)
- v
 - command line option, [273](#)
- v, --verbose
 - command line option, [310](#)
- verify

llvm-bcanalyzer command line option, 297

-verify-each
command line option, 267

-version
command line option, 271, 302, 317
llvm-cov-gcov command line option, 284
tblgen command line option, 309

-vv, -echo-all-commands
command line option, 310

-weighted-input=weight,filename
llvm-profdata-merge command line option, 288

-x86-asm-syntax=syntax
command line option, 273

C

command line option

-check-prefix prefix, 301

-check-prefixes prefix1,prefix2,..., 301

-config-prefix=NAME, 310

-debug, 312

-disable-excess-fp-precision, 269

-disable-fp-elim, 269

-enable-no-infs-fp-math, 269

-enable-no-nans-fp-math, 269

-enable-unsafe-fp-math, 269

-enable-var-scope, 302

-filter=REGEXP, 311

-implicit-check-not check-pattern, 302

-input-file filename, 301

-load=<dso_path>, 269

-match-full-lines, 302

-max-tests=N, 311

-max-time=N, 311

-no-progress-bar, 310

-num-shards=M, 311

-path=PATH, 311

-print-machineinstrs, 269

-regalloc=<allocator>, 269

-run-shard=N, 311

-show-suites, 312

-show-tests, 312

-show-unsupported, 310

-show-xfail, 310

-shuffle, 311

-spiller=<spiller>, 270

-stats, 269

-strict-whitespace, 302

-time-passes, 269

-time-tests, 311

-vg, 311

-vg-arg=ARG, 311

-vg-leak, 311

-x86-asm-syntax=[attlntel], 270

-D NAME[=VALUE], -param NAME[=VALUE],
310

-O=uint, 268

-S, 267, 273

-{passname}, 267

-a, -show-all, 310

-code-model=model, 272

-d, 273

-debug, 267

-debug-dump=section, 293

-default-arch, 292

-demangle, 292

-disable-excess-fp-precision, 271

-disable-inlining, 267

-disable-opt, 267

-disable-post-RA-scheduler, 272

-disable-spill-fusing, 272

-dsym-hint=<path/to/file.dSYM>, 292

-dyn-symbols, 317

-dynamic-table, 317

-elf-section-groups, -g, 318

-enable-no-infs-fp-math, 271

-enable-no-nans-fp-math, 271

-enable-unsafe-fp-math, 271

-expand-relocs, 317

-f, 267, 273

-fake-argv0=executable, 271

-file-headers, -h, 317

-filetype=<output file type>, 268

-force-interpreter={ false,true }, 271

-functions=[none|short|linkage], 292

-h, -help, 310

-help, 267, 268, 271, 273, 301, 317

-inlining, 292

-j N, -threads=N, 310

-jit-enable-eh, 272

-join-liveintervals, 272

-load=<plugin>, 267

-load=pluginfilename, 271

-march=<arch>, 268

-march=arch, 271

-mattr=a1,+a2,-a3,..., 269, 271

-mcpu=<cpu name>, 268

-mcpu=cpu name, 271

-meabi=[default|gnul4l5], 269

-mtriple=<target triple>, 268

-mtriple=target triple, 271

-needed-libs, 317

-nozero-initialized-in-bss, 272

-o <filename>, 267

-o filename, 273, 290

-obj, 292

-p, 267

-pre-RA-sched=scheduler, 272

- pretty-print, [292](#)
- print-address, [292](#)
- program-headers, [317](#)
- q, -quiet, [310](#)
- regalloc=allocator, [272](#)
- relocation-model=model, [272](#)
- relocations, -r, [317](#)
- s, -succinct, [310](#)
- section-data, -sd, [317](#)
- section-relocations, -sr, [317](#)
- section-symbols, -st, [317](#)
- sections, -s, [317](#)
- seed seed, [290](#)
- size size, [290](#)
- soft-float, [271](#)
- spiller, [272](#)
- stats, [267](#), [271](#)
- strip-debug, [267](#)
- symbols, -t, [317](#)
- time-passes, [267](#), [271](#)
- unwind, -u, [317](#)
- use-symbol-table, [292](#)
- v, [273](#)
- v, -verbose, [310](#)
- verify-each, [267](#)
- version, [271](#), [302](#), [317](#)
- vv, -echo-all-commands, [310](#)
- x86-asm-syntax=syntax, [273](#)

L

llvm-bcanalyzer command line option

- dump, [297](#)
- help, [297](#)
- nodetails, [297](#)
- verify, [297](#)

llvm-cov-export command line option

- arch=<name>, [287](#)

llvm-cov-gcov command line option

- help, [284](#)
- a, -all-blocks, [284](#)
- b, -branch-probabilities, [284](#)
- c, -branch-counts, [284](#)
- f, -function-summaries, [284](#)
- l, -long-file-names, [284](#)
- n, -no-output, [284](#)
- o=<DIR|FILE>, -object-directory=<DIR>, -
object-file=<FILE>, [284](#)
- p, -preserve-paths, [284](#)
- u, -unconditional-branches, [284](#)
- version, [284](#)

llvm-cov-report command line option

- arch=<name>, [287](#)
- show-functions, [287](#)
- use-color[=VALUE], [287](#)

llvm-cov-show command line option

- Xdemangler=<TOOL>|<TOOL-OPTION>, [286](#)
- arch=<name>, [285](#)
- format=<FORMAT>, [286](#)
- line-coverage-gt=<N>, [286](#)
- line-coverage-lt=<N>, [286](#)
- name-regex=<PATTERN>, [286](#)
- name=<NAME>, [286](#)
- num-threads=N, -j=N, [286](#)
- output-dir=PATH, [286](#)
- region-coverage-gt=<N>, [286](#)
- region-coverage-lt=<N>, [286](#)
- show-expansions, [285](#)
- show-instantiations, [285](#)
- show-line-counts, [285](#)
- show-line-counts-or-regions, [285](#)
- show-regions, [285](#)
- tab-size=<TABSIZ>, [286](#)
- use-color[=VALUE], [285](#)

llvm-nm command line option

- debug-syms, -a, [279](#)
- defined-only, [279](#)
- dynamic, -D, [279](#)
- extern-only, -g, [279](#)
- format=format, -f format, [279](#)
- no-sort, -p, [279](#)
- numeric-sort, -n, -v, [279](#)
- print-file-name, -A, -o, [279](#)
- print-size, -S, [280](#)
- radix=RADIX, -t, [280](#)
- size-sort, [280](#)
- undefined-only, -u, [280](#)
- B (default), [279](#)
- P, [279](#)
- help, [279](#)

llvm-profdata-merge command line option

- binary (default), [288](#)
- gcc, [288](#)
- help, [288](#)
- input-files=path, -f=path, [288](#)
- instr (default), [288](#)
- num-threads=N, -j=N, [289](#)
- output=output, -o=output, [288](#)
- sample, [288](#)
- sparse[=true|false], [289](#)
- text, [288](#)
- weighted-input=weight,filename, [288](#)

llvm-profdata-show command line option

- all-functions, [289](#)
- counts, [289](#)
- function=string, [289](#)
- help, [290](#)
- instr (default), [290](#)
- memop-sizes, [290](#)

- output=output, -o=output, 290
- sample, 290
- text, 290
- topn=n, 290

T

tblgen command line option

- I directory, 308
- asmparsernum N, 308
- asmwriternum N, 308
- class className, 308
- gen-asm-matcher, 309
- gen-asm-writer, 309
- gen-dag-isel, 309
- gen-dfa-packetizer, 309
- gen-disassembler, 309
- gen-emitter, 308
- gen-enhanced-disassembly-info, 309
- gen-fast-isel, 309
- gen-instr-info, 309
- gen-intrinsic, 309
- gen-pseudo-lowering, 309
- gen-register-info, 308
- gen-subtarget, 309
- gen-tgt-intrinsic, 309
- help, 308
- o filename, 308
- print-enums, 308
- print-records, 308
- print-sets, 308
- version, 309