Code Generation for Data Processing

Alexis Engelke

Winter 2025/26

Contents

1. Introduction and Interpretation

1.1. Organization

[Slide 2] Module "Code Generation for Data Processing"

Learning Goals

- Getting from an intermediate code representation to machine code
- Designing and implementing IRs and machine code generators
- Apply for: JIT compilation, query compilation, ISA emulation

Prerequisites

- Computer Architecture, Assembly — ERA, GRA/ASP
- Databases, Relational Algebra — GDB
- Beneficial: Compiler Construction, Modern DBs

[Slide 3] Topic Overview

Introduction

- Introduction and Interpretation
- Compiler Front-end

Intermediate Representations

- IR Concepts and Design
- LLVM-IR
- Analyses and Optimizations

Compiler Back-end

- Instruction Selection
- Register Allocation
- Linker, Loader, Debuginfo

Applications

- JIT-compilation + Sandboxing
- Query Compilation
- Binary Translation

[Slide 4] Lecture Organization

- Lecturer: Dr. Alexis Engelke `engelke@in.tum.de`
- Time slot: Thu 10-14, 02.11.018
- Material: `https://db.in.tum.de/teaching/ws2526/codegen/`

Exam

- Written exam, 90 minutes, **no retake**, 2026-02-27 14:00
- (Might change to oral on very low registration count)

[Slide 5] Exercises

- Regular homework, often with programming exercise
- Submission via POST request (see assignments)
  - Grading with $\{*, +, \sim, -\}$, feedback on best effort

> * Grade $*$: excellent submission, goes beyond expectations
> * Grade $+$: good submission, meets expectations
> * Grade $\sim$: acceptable submission, but is below expectations
> * Grade $-$: unacceptable submission, too far below expectations

- Exercises integrated into lecture
- Hands-on programming or analysis of systems (needs laptop)
- Occasionally: present and discuss homework solutions

Grade Bonus

- Requirement: $N - 2$ "sufficiently working" homework submissions and one presentations of homework in class (depends on submission count)
- Bonus: grades in $[1.3; 4.0]$ improved by $0.3/0.4$

[Slide 6] Why study compilers?

- Critical component of every system, functionality and performance
  - Compiler mostly *alone* responsible for using hardware well
- Brings together many aspects of CS:
  - Theory, algorithms, systems, architecture, software engineering, (ML)
- New developments/requirements pose new challenges
  - New architectures, environments, language concepts, ...
- High complexity!

[Slide 7] Compiler Lectures @ TUM

Compiler Construction IN2227, SS, THEO	Program Optimization IN2053, WS, THEO	Virtual Machines IN2040, SS, THEO
Front-end, parsing, semantic analyses, types	Analyses, transformations, abstract interpretation	Mapping programming paradigms to IR/bytecode

Programming Languages CIT3230000, WS	Code Generation CIT3230001, WS
Implementation of advanced language features	Back-end, machine code generation, JIT comp.

[Slide 8] Why study code generation?

- Frameworks (LLVM, . . .) exist and are comparably good, but often not good enough (performance, features)
  - Many systems with code gen. have their own back-end
  - E.g.: V8, WebKit FTL, .NET RyuJIT, GHC, Zig, QEMU, Umbra, . . .
- Machine code is not the only target: bytecode
  - Often used for code execution
  - E.g.: V8, Java, .NET MSIL, BEAM (Erlang), Python, MonetDB, eBPF, . . .
  - Allows for flexible design
  - But: efficient execution needs machine code generation

[Slide 9] Proebsting's Law

> "Compiler advances double computing power every *18* years."
>
> – Todd Proebsting, 1998[a]

[a] `http://proebsting.cs.arizona.edu/law.html`

- Still optimistic; depends on number of abstractions

The performance increases compilers can make on existing code are typically low. However, optimizing compilers gain more abilities in simplifying needlessly complex code, enabling the use of more abstractions and therefore higher level code. These abstractions are removed/optimized during compilation, enabling languages to promote these as *zero-cost abstractions*. They do, however, have a cost: compile times.

Also note that some of these "zero-cost" abstractions actually *do* have some run-time cost. For example, the mere possibility of C++ exceptions can cause less efficient

> machine code and might prevents optimizations due to the more complex control flow possibilities.

1.2. Overview

[Slide 10] Motivational Example: Brainfuck

- Turing-complete esoteric programming language, 8 operations
  - Input/output: `. ,`
  - Moving pointer over infinite array: `< >`
  - Increment/decrement: `+ -`
  - Jump to matching bracket if (not) zero: `[ ]`

```
++++++[->+++++++<]>.
```

- Execution with pen/paper? ☹

[Slide 11] Program Execution

Programs

- High flexibility (possibly)
- Many abstractions (typically)
- Several paradigms

Hardware/ISA

- Low-level interface
- Few operations, imperative
- "Not easy" to write

[Slide 12] Motivational Example: Brainfuck – Interpretation

- Write an interpreter!

```
unsigned char state[10000];
unsigned ptr = 0, pc = 0;
while (prog[pc])
  switch (prog[pc++]) {
  case '.': putchar(state[ptr]); break;
  case ',': state[ptr] = getchar(); break;
  case '>': ptr++; break;
  case '<': ptr--; break;
  case '+': state[ptr]++; break;
  case '-': state[ptr]--; break;
  case '[': state[ptr] || (pc = matchParen(pc, prog)); break;
  case ']': state[ptr] && (pc = matchParen(pc, prog)); break;
  }
```

[Slide 13] Program Execution

Compiler

- Translate program to other lang.
- Might optimize/improve program

- C, C++, Rust → machine code
- Python, Java → bytecode

Interpreter

- Directly execute program
- Computes program result

- Shell scripts, Python bytecode, machine code (conceptually)

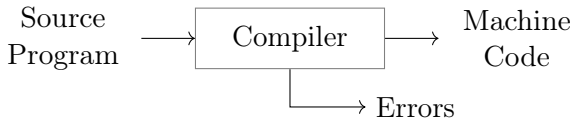
Multiple compilation steps can precede the "final interpretation"

1.3. High-Level Structure of Compilers

[Slide 14] Compilers

- Targets: machine code, bytecode, or other source language
- Typical goals: better language usability and performance
  - Make lang. usable at all, faster, use less resources, etc.
- Constraints: specs, resources (comp.-time, etc.), requirements (perf., etc.)
- Examples:
  - "Classic" compilers source → machine code
  - JIT compilation of JavaScript, WebAssembly, Java bytecode, ...
  - Database query compilation
  - ISA emulation/binary translation

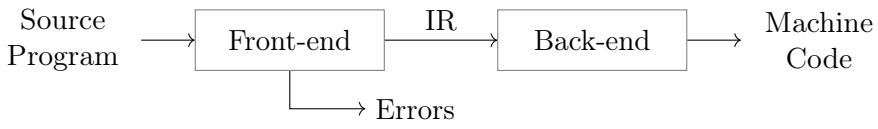
[Slide 15] Compiler Structure: Monolithic



- Inflexible architecture, hard to retarget

> Some languages like C are designed to be compilable in a single pass without building any intermediate representation of the code between source and assembly. Single-pass compilers exist, but often have very limited possibilities to transform the code. They might not even know basic code properties, e.g., the size of the stack frame, during compilation of a function.

[Slide 16] Compiler Structure: Two-phase architecture



Front-end

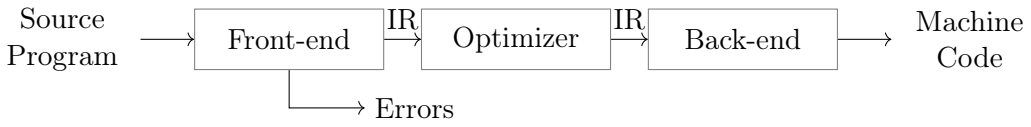
- Parses source code
- Detect syntax/semantical errors
- Emit *intermediate representation* encode semantics/knowledge
- Typically: $\mathcal{O}(n)$ or $\mathcal{O}(n \log n)$

Back-end

- Translate IR to target architecture
- Can assume valid IR ($\rightsquigarrow$ no errors)
- Possibly one back-end per arch.
- Contains $\mathcal{NP}$-complete problems

> After parsing, all information is encoded in the IR, including references to source code constructs for debugging support. The input source code is (at least conceptually) no longer needed.
>
> In practice, there are very rare cases where the back-end can also raise errors. This can happen, for example, when some very architecture-specific constraints might be hard to verify during parsing (e.g., inline assembly constraints in combination with available registers).

[Slide 17] Compiler Structure: Three-phase architecture



- Optimizer: analyze/transform/rewrite program inside IR

- Conceptual architecture: real compilers typically much more complex
  - Several IRs in front-end and back-end, optimizations on different IRs
  - Multiple front-ends for different languages
  - Multiple back-ends for different architectures

> Example Clang/LLVM (will be covered in more detail later): Clang parses the input into an abstract syntax tree (IR 1), uses this for semantic analyses; then Clang transforms the code into LLVM-IR (IR 2), which is primarily used for optimization; then the LLVM back-end transforms the code further into LLVM's Machine IR (IR 3), executes some low-level optimizations and register allocation there; the assembly

printer of the back-end then lowers the code further to LLVM's machine code representation (IR 4), before finally emitting machine code. Some optimizations inside this pipeline, e.g. vectorization, might even build further representation of the code.

Why are compilers using so many different code representations? Different transformations work best at different abstraction levels. Diagnosing unused variables, for example, requires information about the source code. Optimization of arithmetic computations is easier in a data-flow-focused representation, where no explicit variables exist. Low-level modifications, like folding operations into complex addressing modes of the ISA, need a code representation where ISA instructions are already present.

[Slide 18] Compiler Front-end

1. Tokenizer: recognize words, numbers, operators, etc. $\mathcal{R}e$
   - Example: `a+b*c` → `ID(a) PLUS ID(b) TIMES ID(c)`
2. Parser: build (abstract) syntax tree, check for syntax errors $\mathcal{CFG}$
   - Syntax Tree: describe grammatical structure of complete program Example: `expr("a", op("+"), expr("b", op("*"), expr("c"))`
   - Abstract Syntax Tree: only relevant information, more concise Example: `plus("a", times("b", "c"))`
3. Semantic Analysis: check types, variable existence, etc.
4. IR Generator: produce IR for next stage
   - This might be the AST itself

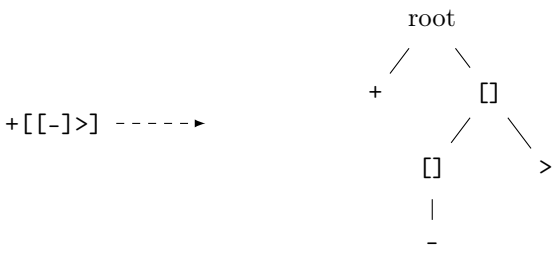
[Slide 19] Compiler Back-end

1. Instruction Selection: map IR operations to target instructions
   - Use target features: special insts., addressing modes, ...
   - Still using virtual/unlimited registers
2. Instruction Scheduling: optimize order for target arch.
   - Start memory/high-latency earlier, etc.
   - Requires knowledge about micro-architecture
3. Register Allocation: map values to fixed register set/stack
   - Use available registers effectively, minimize stack usage

1.4. Interpretation

[Slide 20] Motivational Example: Brainfuck – Front-end

- Need to skip comments
- Bracket searching is expensive/redundant
- Idea: "parse" program!
- Tokenizer: yield next operation, skipping comments
- Parser: find matching brackets, construct AST



[Slide 21] Motivational Example: Brainfuck – AST Interpretation

- AST can be interpreted recursively

```
struct node { char kind; unsigned cldCnt; struct node* cld; };
struct state { unsigned char* arr; size_t ptr; };
void donode(struct node* n, struct state* s) {
  switch (n->kind) {
  case '+': s->arr[s->ptr]++; break;
  // ...
  case '[': while (s->arr[s->ptr]) children(n, s); break;
  case 0: children(n, s); break; // root
  }
}
void children(struct node* n, struct state* s) {
  for (unsigned i = 0; i < n->cldCnt; i++) donode(n->cld + i, s);
}
```

[Slide 22] Motivational Example: Brainfuck – Optimization

- Inefficient sequences of +/-/</> can be combined
  - Trivially done when generating IR
- Fold patterns into more high-level operations

In-Class Exercise:

Look at some Brainfuck programs. Which patterns are beneficial to fold?

Solution on page 83.

[Slide 23] Motivational Example: Brainfuck – Optimization

- Fold offset into operation
  - `right(2) add(1)` = `addoff(2, 1) right(2)`
  - Also possible with loops
- Analysis: does loop move pointer?
  - Loops that keep position intact allow more optimizations
  - Maybe distinguish "regular loops" from arbitrary loops?
- Get rid of all "effect-less" pointer movements
- Combine arithmetic operations, disambiguate addresses, etc.

[Slide 24] Motivational Example: Brainfuck – Bytecode

- Tree is nice, but rather inefficient ⇝ flat and compact bytecode
- Avoid pointer dereferences/indirections; keep code size small
- Maybe dispatch two instructions at once?
  - `switch (ops[pc] | ops[pc+1] << 8)`
- Superinstructions: combine common sequences to one instruction

> Dispatching multiple instructions at once can be problematic due to the explosion of cases that need to be implemented (often results in large jump tables and lots of code with resulting cache misses and branch mispredictions). Often, it is advisable to not always switch over multiple neighbored instructions, but instead combine common sequences into superinstructions.

[Slide 25] Threaded Interpretation[1]

- Simple `switch`–`case` dispatch has lots of branch misses
- Threaded interpretation: at end of a handler, jump to next op

```
struct op { char op; char data; };
struct state { unsigned char* arr; size_t ptr; };
void threadedInterp(struct op* ops, struct state* s) {
  static const void* table[] = { &&CASE_ADD, &&CASE_RIGHT, };
#define DISPATCH do { goto *table[(++pc)->op]; } while (0)

  struct op* pc = ops;
  DISPATCH;

CASE_ADD: s->arr[s->ptr] += pc->data; DISPATCH;
CASE_RIGHT: s->arr += pc->data; DISPATCH;
}
```

> With threaded interpretation there is not a single indirect jump instruction inside the dispatcher, but one indirect jump instruction per operation. Each of these indirect jumps then occupies a different branch prediction slot in the CPU. If an operation of type X is typically followed by an operation of type Y, with threaded interpretation the CPU has a much better chance of correctly predicting the dispatch branch to the next operation, because the indirect jump at the end of operation X typically jumps to operation Y. Without threaded interpretation, there would be only a single indirect branch, which is much harder to predict.
>
> Threaded interpretation is especially useful on older and less powerful CPUs. Recent CPUs (e.g., Intel since Skylake, AMD since Zen 3, Apple Silicon) store the history of branches and use this for better prediction. On such processors, threaded interpretation might not improve performance (or gains might be lower).

[1] MA Ertl and D Gregg. "The structure and performance of efficient interpreters". In: *JILP* 5 (2003), pp. 1–25. URL: `http://www.jilp.org/vol5/v5paper12.pdf`.

[Slide 26] Threaded Interpretation with Tail Calls

- Threaded interpretation can also be implemented with tails calls

```
struct op { char op; char data; };
struct state { unsigned char* arr; size_t ptr; };
void tcInterp(struct op* pc, struct state* s);
static void fn_add(struct op* pc, struct state* s) {
  s->arr[s->ptr] += pc->data; tcInterp(pc + 1, s); }
static void fn_right(struct op* pc, struct state* s) {
  s->arr += pc->data; tcInterp(pc + 1, s); }
void tcInterp(struct op* pc, struct state* s) {
  typedef void (* Fn)(struct op* pc, struct state* s);
  static const Fn fns[] = { fn_add, fn_right };
  fns[pc->op](pc, s);
}
```

> Computed `goto` is a GNU extension and similar features are not available in many other languages. However, it is possible to achieve a similar effect through indirect tail calls[a].
>
> In this example[b], the resulting assembly is almost identical for x86-64, but for AArch64, some more engineering effort would be required.
>
> The code as written has another problem: At `-O0`, no tail call optimization will be performed. Some compilers like Clang provide a way to force tail calls in C/C++ code:
>
> ```
> void tcInterp(struct op* pc, struct state* s) {
>   // ... (and adapt individual functions likewise)
>   [[clang::musttail]] return fns[pc->op](pc, s);
> }
> ```
>
> In the context of functional programming, this style is also referred to as *continuation-passing style*.
>
> [a] A tail call is a call at the end of a function without adding a new entry to the stack frame.
>
> [b] `https://godbolt.org/z/E5drG61aK`

[Slide 27] Direct Threading[2]

- Use function pointer as operation to avoid indirection

```
struct op; struct state;
typedef void (* Fn)(struct op* pc, struct state* s);
struct op { Fn op; char data; };
struct state { unsigned char* arr; size_t ptr; };

void fn_add(struct op* pc, struct state* s) {
  s->arr[s->ptr] += pc->data; pc[1].op(pc + 1, s); }
void fn_right(struct op* pc, struct state* s) {
  s->arr += pc->data; pc[1].op(pc + 1, s); }
void dtInterp(struct op* ops, struct state* s) {
  ops[0].op(ops, s);
}
```

[2] JR Bell. "Threaded Code". In: *CACM* 16.6 (1973), pp. 370–372. URL: `https://dl.acm.org/doi/pdf/10.1145/362248.362270`.

> While there might be a benefit of reducing the instruction count, the effect is typically insignificant on out-of-order CPUs, as the extra load instruction is typically a cache hit and the latency is hidden by other operations.

[Slide 28] Threaded Interpretation — Comparison

In-Class Exercise:

What are benefits/drawbacks of the three threading approaches[a]?

- Indirect threading with computed `goto`
- Indirect threading with tail calls
- Direct threading with tail calls

Solution on page 83.

[a] `https://godbolt.org/z/4rcEv4sqj`

- Some differences on code size, readability, and maintainability
- Performance: depends on hardware, context – always measure!

[Slide 29] Fast Interpretation

- Key technique to "avoid" compilation to machine code
- Preprocess program into efficiently executable bytecode
  - Easily identifiable opcode, homogeneous structure
  - Can be linear (fast to execute), but trees also work
  - Match bytecode ops with needed operations $\rightsquigarrow$ fewer instructions
  - Larger operations preferable, but not too many (prediction)
- Perhaps optimize – if it's worth the benefit
  - Fold constants, combine instructions, ...
  - Consider superinstructions for common sequences
- For very cold code: avoid transformations at all

[Slide 30] Interpretation vs. Compilation

Fundamental benefits of compilation:

- Elimination of interpreter dispatch
  - Can be significant for bytecodes with many small operations
- Use of CPU registers for values across bytecode instructions
  - Interpreter: most values must be stored in memory
  - Register allocation can bring large improvements

> This is a major driver for performance even on modern CPUs, which have two related optimizations: Store-to-load forwarding, where the value from a store will be forwarded to a later load from the same address; and memory renaming,

> where the physical register of a stored value is kept and a subsequent load with a similar address operand speculatively reuses that register. However, CPU resources for these optimizations are limited and using registers directly is much more efficient.

- No fundamental benefit: optimizations
  - Many optimizations can also be applied on bytecode

Fundamental benefits of interpretation: simple, portable

1.5. Context of Compilation

[Slide 31] Compiler: Surrounding – Compile-time

- Typical environment for a C/C++ compiler:

- Calling Convention: interface with other objects/libraries
- Build systems, dependencies, debuggers, etc.
- Compilation target machine (hardware, VM, etc.)

[Slide 32] Compiler: Surrounding – Run-time

- OS interface (I/O, ...)
- Memory management (allocation, GC, ...)
- Parallelization, threads, ...
- VM for execution of virtual assembly (JVM, ...)
- Run-time type checking
- Error handling: exception unwinding, assertions, ...
- Reflection, RTTI

[Slide 33] Motivational Example: Brainfuck – Runtime Environment

- Needs I/O for `.` and `,`
- Error handling: unmatched brackets
- Memory management: infinitely-sized array

In-Class Exercise:

How to efficiently emulate an infinitely sized array?

Solution on page 83.

[Slide 34] Compilation point: AoT vs. JIT

Ahead-of-Time (AoT)

- All code has to be compiled

- No dynamic optimizations
- Compilation-time secondary concern

Just-in-Time (JIT)

- Compilation-time is critical
- Code can be compiled on-demand
  - Incremental optimization, too
- Handle cold code fast
- Dynamic specializations possible
- Allows for `eval()`

Various hybrid combinations possible

[Slide 35] Introduction and Interpretation – Summary

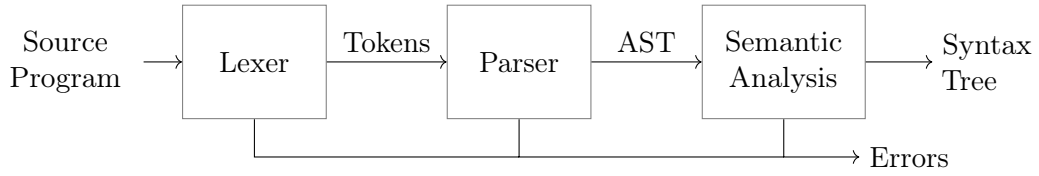
- Compilation vs. interpretation and combinations
- Compilers are key to usable/performant languages
- Target language typically machine code or bytecode
- Three-phase architecture widely used
- Interpretation techniques: bytecode, threaded interpretation, ...
- JIT compilation imposes different constraints

[Slide 36] Introduction and Interpretation – Questions

- What is typically compiled and what is interpreted? Why?
  - PostScript, C, JavaScript, HTML, SQL
- What are typical types of output languages of compilers?
- How does a compiler IR differ from the source input?
- What is the impact of the language paradigm on optimizations?
- What are important factors for an efficient interpreter?
- What are inherent benefits of compilation over interpretation?
- What are key differences between AoT and JIT compilation?

2. Compiler Front-end

[Slide 38] Compiler Front-end



- Typical architecture: separate lexer, parser, and context analysis
  - Allows for more efficient lexical analysis
  - Smaller components, easier to understand, etc.
- Some languages: preprocessor and macro expansion

2.1. Lexing

[Slide 39] Lexer

- Convert stream of chars to stream of words (*tokens*)
- Detect/classify identifiers, numbers, operators, ...
- Strip whitespace, comments, etc.

`a+b*c` → `ID(a) PLUS ID(b) TIMES ID(c)`

- Typically representable as regular expressions

[Slide 40] Typical Token Kinds

- Punctuators `( ) [ ] { } ; = + += | ||`
- Identifiers `abc123 main`
- Keywords `void int __asm__`
- Numeric constants `123 0xab1 5.7e3 0x1.8p1 09.1f`
- Char constants `'a' u'œ'`
- String literals `"abc\x12\n"`
- Internal `EOF COMMENT UNKNOWN INDENT DEDENT`
  - Comments might be useful for annotations, e.g. `// fallthrough`

> Indentation-based languages like Python need separate tokens for indent/dedent, the indentation level is tracked in the lexer. Parsing numbers may need special care to correctly handle all possible cases of integer and floating-point numbers.

[Slide 41] Lexer Implementation

```
struct Token { enum Kind { IDENT, EOF, PLUS, PLUSEQ, /*...*/ };
  std::string_view v; Kind kind; };
Token next(std::string_view v) {
  if (v.empty()) return Token{v, Token::EOF};
  if (v.starts_with("+=")) return Token{"+="sv, Token::PLUSEQ};
  if (v.starts_with("+")) return Token{"+"sv, Token::PLUS};
  switch (v[0]) {
  case ' ', '\n', '\t': return next(v.substr(1)); // skip whitespace
  case 'a' ... 'z', 'A' ... 'Z', '_': {
    Token t = // ... parse identifer, e.g. using regex
    if (auto kind = isKeyword(t.v)) return Token{*kind, t.v};
    return t;
  }
  case '0' ... '9': // ... parse number
  default: return Token{v.substr(0, 1), Token::ERROR};
  }
}
```

> This is just a minimal and non-optimized implementation to illustrate the concept. Performance-focused implementations do not use explicit regular expressions but write the state machine into code.
>
> The struct `Token` has room for improvement. First, a `string_view` is unnecessarily large with 16 bytes, most tokens are smaller than 2^{16} bytes. Some tracking of the source locations is advisable for attaching diagnostics to their origin inside the code, for example by storing a file ID and the byte offset into the file. By tracking the byte offsets of line breaks, the line number can be reconstructed in $\mathcal{O}(\log n)$ from the byte offset.
>
> Another optimization strategy is string interning, where identifiers are converted into unique integers (or pointers) during parsing. During later phases, comparing interned strings is much more efficient, as it is just an integer/pointer comparison. Another benefit is that the entire input file does not need to be kept in memory during parsing.

[Slide 42] Lexing C??=

```
main() <%
  // yay, this is C99??/
  puts("hi world!");
  puts("what's up??!");
%>
```

Output: `what's up|`

- Trigraphs for systems with more limited encodings/char sets
- Digraphs to provide a more readable alternative...

> Besides digraphs, trigraphs, and the preprocessor, C has another weird property: identifier names can be split by \, which concatenates two lines. It is necessary to construct the "real" identifier first. To simplify memory management in such cases, a bump pointer allocator (allocate large chunks of memory from the OS, then simply bump the end pointer for every allocation) can be useful to store such constructed names.

[Slide 43] Lexer Implementation

- Essentially a DFA (for most languages)
  - Set of regexes → NFA → DFA
- Respect whitespace/separators for operators, e.g. `+` and `+=`
- Automatic tools (e.g., flex) exist; most compilers do their own
- Keywords typically parsed as identifiers first
  - Check identifier if it is a keyword; can use perfect hashing
- Other practical problems
  - UTF-8 homoglyphs; trigraphs; pre-processing directives

> A tool to generate perfect hash tables from a set of keywords is gperf. Example, compile with `gperf -L C++ -C -E -t <input>`:
>
> ```
> struct keyword {char* name; int val; }
> %%
> int, 1
> char, 2
> void, 3
> if, 4
> else, 5
> while, 6
> return, 7
> ```

2.2. Parsing

[Slide 44] Parsing

- Convert stream of tokens into (abstract) syntax tree
- Most programming languages are context-sensitive
  - Variable declarations, argument count, type match, etc. ⇝ separated into semantic analysis

  Syntactically valid: `void foo = doesntExist / "abc";`
- Grammar usually specified as CFG

[Slide 45] Context-Free Grammar (CFG)

- Terminals: basic symbols/tokens
- Non-terminals: syntactic variables

- Start symbol: non-terminal defining language
- Productions: non-terminal → series of (non-)terminals

$$
\begin{aligned}
\mathit{stmt} &\rightarrow \mathit{whileStmt} \mid \mathit{breakStmt} \mid \mathit{exprStmt} \\
\mathit{whileStmt} &\rightarrow \textbf{while (}\ \mathit{expr}\ \textbf{)}\ \mathit{stmt} \\
\mathit{breakStmt} &\rightarrow \textbf{break ;} \\
\mathit{exprStmt} &\rightarrow \mathit{expr}\ \textbf{;} \\
\mathit{expr} &\rightarrow \mathit{expr} + \mathit{expr} \mid \mathit{expr}\ \textbf{*}\ \mathit{expr} \mid \mathit{expr} = \mathit{expr} \mid \textbf{(}\ \mathit{expr}\ \textbf{)} \mid \textbf{number}
\end{aligned}
$$

[Slide 46] Hand-written Parsing – First Try

- One function per non-terminal
- Check expected structure
- Return AST node
- Need look-ahead!

```
NodePtr parseBreakStmt() {
  consume(Token::BREAK);
  consume(Token::SEMICOLON);
  return newNode(Node::BreakStmt);
}
NodePtr parseWhileStmt() {
  consume(Token::WHILE);
  consume(Token::LPAREN);
  NodePtr expr = parseExpr();
  consume(Token::RPAREN);
  NodePtr body = parseStmt();
  return newNode(Node::WhileStmt,
    {expr, body});
}
NodePtr parseStmt() {
  // whoops!
}
```

[Slide 47] Hand-written Parsing – Second Try

- Need look-ahead to distinguish production rules
- Consequences for grammar:
  - No left-recursion
  - First n terminals must allow distinguishing rules
  - $LL(n)$ grammar; n typically 1

  ⇒ Not all CFGs (easily) parseable (but most programming langs. are)
- Now... expressions

```
NodePtr parseBreakStmt() { /*...*/ }
NodePtr parseWhileStmt() { /*...*/ }

NodePtr parseStmt() {
  Token t = peekToken();
  if (t.kind == Token::BREAK)
    return parseBreakStmt();
```

```
  if (t.kind == Token::WHILE)
    return parseWhileStmt();
  // ...
  NodePtr expr = parseExpr();
  consume(Token::SEMICOLON);
  return newNode(Node::ExprStmt,
    {expr});
}
```

[Slide 48] Ambiguity

$$expr \rightarrow expr + expr \mid expr \ \mathbf{*} \ expr \mid expr = expr \mid (\ expr\ ) \mid \mathbf{number}$$

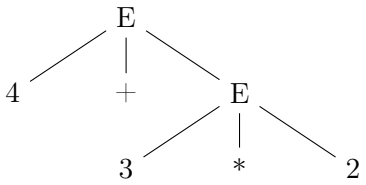
Input: $4 + 3 * 2$

> The grammar, as specified, is ambiguous, there are two possible ways to parse the input.

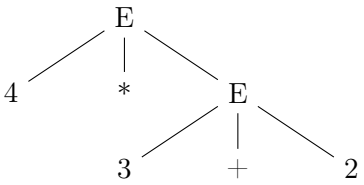
[Slide 49] Ambiguity – Rewrite Grammar?

$$primary \rightarrow (\ expr\ ) \mid \mathbf{number}$$
$$expr \rightarrow primary + expr \mid primary \ \mathbf{*} \ expr \mid primary = expr \mid primary$$

Input: $4 + 3 * 2$



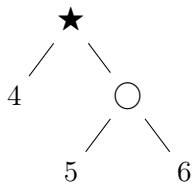
Input: $4 * 3 + 2$



> The grammar is no longer ambiguous, but the result might not be expected, conventionally, multiplication has a stronger binding than addition.

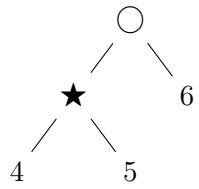
[Slide 50] Ambiguity – Precedence

Input: $4 \bigstar 5 \bigcirc 6$



If $prec(◯) > prec(★)$ or equal prec. and ★ is right-assoc.

> Examples:
> - $4 + 5 \cdot 6$ ($prec(\cdot) > prec(+)$)
> - $a = b = c$ ($=$ is right-assoc.)
>
> $b = c$ should be executed first



If $prec(◯) < prec(★)$ or equal prec. and ★ is left-assoc.

> Examples:
> - $4 + 5 < 6$ ($prec(<) < prec(+)$)
> - $a + b - c$ ($+$ is left-assoc.)
>
> $a + b$ should be executed first

[Slide 51] Hand-written Parsing – Expression Parsing

- Start with basic expr.:
- Number, variable, etc.
- Parenthesized expr.
  - Parse full expression
  - Next token must be `)`
- Unary expr: followed by expr. with higher prec.
  - `-` < unary `-` < `[]`/`->`

```
NodePtr parseExpr(unsigned minPrec=0);
NodePtr parsePrimaryExpr() {
  switch (Token t = next(); t.kind) {
  case Token::IDENT:
    return makeNode(Node::IDENT, t.v);
  case Token::NUMBER: // ...
  case Token::MINUS:
    // Only exprs with high precedence
    return makeNode(Node::UMINUS,
      {parseExpr(UNARY_PREC)});
  case Token::LPAREN: // ...
  // ...
  }
}
```

[Slide 52] Hand-written Parsing – Expression Parsing

- Only allow ops. with higher prec. on the right child
  - Right-assoc.: allow same
- Lower prec.: return + insert higher up in the tree

```
OpDesc OPS[] = { // {prec, rassoc}
  [Token::MUL] = {12, false},
  [Token::ADD] = {11, false},
```

```
  [Token::EQ] = {2, true},
  [Token::QUEST] = {3, true}, // ?:
}
NodePtr parseExpr(unsigned minPrec=1) {
  auto lhs = parsePrimaryExpr();
  while (auto op = OPS[next().kind];
         op.prec >= minPrec) {
    // ... handle (, [, ?: ...
    auto newPrec = op.rassoc ?
      op.prec : op.prec + 1;
    auto rhs = parseExpr(newPrec);
    lhs = makeNode(op.nodeKind,
      {lhs, rhs});
  }
  return lhs;
}
```

In-Class Exercise:

`a = 3 * 2 + 1;` `a = b + c + d = 1;` `a ? 1 : b ? 2 : 3;`

Solution on page 84.

[Slide 53] Top-down vs. Bottom-up Parsing

Top-down Parsing

- Start with top rule
- Every step: choose expansion
- LL(1) parser
  - Left-to-right, Leftmost Derivation
- "Easily" writable by hand
- Error handling rather simple
- Covers many prog. languages

Bottom-up Parsing

- Start with text
- Reduce to non-terminal
- LR(1) parser
  - Left-to-right, Rightmost Derivation
  - Strict super-set of LL(1)
- Often: uses parser generator
- Error handling more complex
- Covers nearly all prog. languages

[Slide 54] Parser Generators

- Writing parsers by hand can be large effort
- Parser generators can simplify parser writing a lot
  - Yacc/Bison, PLY, ANTLR, ...

- Automatic generation of parser/parsing tables from CFG
  - Finds ambiguities in the grammar
  - Lexer often written by hand
- Used heavily in practice, unless error handling is important

[Slide 55] Bison Example – part 1

```
%define api.pure full
%define api.value.type {ASTNode*}
%param { Lexer* lexer }
%code{
static int yylex(ASTNode** lvalp, Lexer* lexer);
}
%token NUMBER
%token WHILE "while"
%token BREAK "break"

// precedence and associativity
%right '='
%left '+'
%left '*'
```

[Slide 56] Bison Example – part 2

```
%%
stmt : WHILE '(' expr ')' stmt { $$ = mkNode(WHILE, $1, $2); }
     | BREAK ';'               { $$ = mkNode(BREAK, NULL, NULL); }
     | expr ';'                { $$ = $1; }
     ;
expr : expr '+' expr           { $$ = mkNode('+', $1, $2); }
     | expr '*' expr           { $$ = mkNode('*', $1, $2); }
     | expr '=' expr           { $$ = mkNode('=', $1, $2); }
     | '(' expr ')'            { $$ = $1; }
     | NUMBER
     ;
%%
static int yylex(ASTNode** lvalp, Lexer* lexer) {
    /* return next token, or YYEOF/... */ }
```

> Compile with `bison -dg input.ypp`, it will emit a C++ header, the implementation file, and also a graph showing the state machine of he parser.

[Slide 57] Parsing in Practice

- Some use parser generators, e.g. Python some use hand-written parsers, e.g. GCC, Clang, Swift, Go
- Optimization of grammar for performance
  - Rewrite rules to reduce states, etc.
- Useful error-handling: complex!
  - Try skipping to next separator, e.g. `;` or `,`

- Programming languages are not always context-*free*
  - C: `foo* bar;`
  - May need to break separation between lexer and parser

> In fact, many compilers[a] use hand-written parsers, because they allow for better error messages a more graceful handling of syntax errors, leading to more reported errors during a single (failing) ompilation.
>
> [a] `https://notes.eatonphil.com/parser-generators-vs-handwritten-parsers-survey-2021.html`

[Slide 58] Parsing C++

- C++ is not context-free (inherited from C): `T * a;`
- C++ is ambiguous: `Type (a), b;`
  - Can be a declaration or a comma expression
- C++ templates are Turing-complete[1]
- C++ *parsing* is hence *undecidable*[2]
  - Template instantiation combined with C `T * a` ambiguity

2.3. Semantic Analysis

[Slide 59] Semantic Analysis

- Syntactical correctness $\not\Rightarrow$ correct program `void foo = doesntExist / ++"abc";`
- Needs context-sensitive analysis:
  - Variable existence, storage, accessibility, ...
  - Function existence, arguments, ...
  - Operator type compatibility
  - Attribute allowance
- Additional type complexity: inference, polymorphism, ...

[Slide 60] Semantic Analysis: Scope Checking with AST Walking

- Idea: walk through AST (in DFS-order) and validate on the way
- Keep track of scope with declared variables
  - Might need to keep track of defined types separately

In-Class Exercise:

How to implement the scope data structure?

[1] TL Veldhuizen. *C++ templates are Turing complete*. 2003. URL: `http://port70.net/~nsz/c/c%2B%2B/turing.pdf`.

[2] J Haberman. *Parsing C++ is literally undecidable*. 2013. URL: `https://blog.reverberate.org/2013/08/parsing-c-is-literally-undecidable.html`.

Solution on page 84.

- For identifiers: check existence and get type
- For expressions: check types and derive result type
- For assignment: check lvalue-ness of left side
- *Might* be possible during AST creation
- Needs care with built-ins and other special constructs

[Slide 61] Semantic Analysis and Post-Parsing Transformations

- Check for error-prone code patterns
  - Completeness of `switch`, out-of-range constants, unused variables, ...
- Check method calls, parameter types
- Duplicate code for templates
- Make implicit value conversions explicit
- Handle attributes: visibility, warnings, etc.
- Mangle names, split functions (OpenMP), ABI-specific setup, ...
- Last step: generate IR code

2.4. Miscellaneous

[Slide 62] Parsing Performance

Is parsing/front-end performance important?

- Not necessarily: normal compilers
  - Some languages (e.g., Rust) need unbounded time *for parsing*
- Somewhat: JIT compilers
  - Start-up time is generally noticable
- Somewhat more: Developer tools
  - Imagine: waiting for seconds just for updated syntax highlighting
  - Often uses tricks like incremental updates to parse tree

[Slide 63] Data Types

- Important part of programming languages
- Might have large variety and compatibility
  - Numbers, Strings, Arrays, Compound Types (struct/union), Enum, Templates, Functions, Pointers, ...
  - Class hierarchy, Interfaces, Abstract Classes, ...
  - Integer/float compatibility, promotion, ...
- Might have implicit conversions

[Slide 64] Data Types: Implementing Classes

- Simple `class`/`struct`: trivial, just bunch of fields
  - Methods take (pointer to) `this` as implicit parameter
- Single inheritance: also trivial – extend struct at end
- Virtual methods: store vtable in object representation
  - vtable = table of function pointers for virtual methods
  - Each sub-class has their own vtable
- Multiple inheritance is much more involved
- Dynamic casts: needs run-time type information (RTTI)

[Slide 65] Recommended Lectures

AD IN2227 "Compiler Constructions" covers parsing/analysis in depth

AD CIT3230000 "Programming Languages" covers dispatching/mixins/...

[Slide 66] Compiler Front-end – Summary

- Lexer splits input into tokens
  - Essentially Regex-Matching + Keywords; rather simple
- Parser constructs (abstract) syntax tree from tokens
  - Top-down vs. bottom-up parsing
  - Typical: top-down for control flow; bottom-up for expressions
  - Respect precedence and associativity for operators
- Semantic analysis ensures meaningful program
- Some data structures are complex to implement
- Some programming languages are more difficult to parse

[Slide 67] Compiler Front-end – Questions

- What are typical components of a compiler front-end?
- What output does the lexer produce?
- How does a parser disambiguate rules?
- What is the typical way to handle operator precedence?
- Why are not all programming languages describable using CFGs?
- How to implement classes with virtual functions?

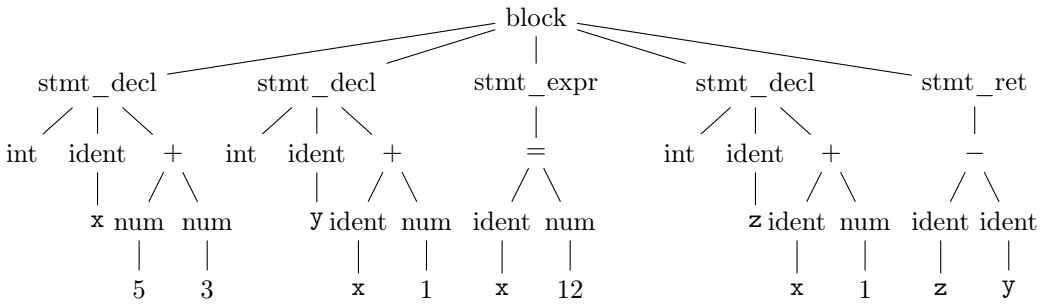
3. Intermediate Representations

3.1. Motivation

[Slide 69] Intermediate Representations: Motivation

- • So far: program parsed into AST
- \+ Great for language-related checks
- \+ Easy to correlate with original source code (e.g., errors)
- – Hard for analyses/optimizations due to high complexity
  - – variable names, control flow constructs, etc.
  - – Data and control flow implicit
- – Highly language-specific

[Slide 70] Intermediate Representations: Motivation



Question: how to optimize? Is `x+1` redundant? ⇝ hard to tell ☹

> In this representation, it is very easy to see that the two `+1` operations have different operands on the left side and are therefore not trivially redundant.

[Slide 71] Intermediate Representations: Motivation

$$
\begin{array}{llllll}
x_1 & \leftarrow & 5 & + & 3 \\
y_1 & \leftarrow & x_1 & + & 1 \\
x_2 & \leftarrow & 12 & & \\
z_1 & \leftarrow & x_2 & + & 1 \\
tmp_1 & \leftarrow & z_1 & - & y_1 \\
\text{return} & & tmp_1 & &
\end{array}
$$

Question: how to optimize? Is `x+1` redundant? ⇝ No! ☺

[Slide 72] Intermediate Representations

- Definitive program representation inside compiler
  - During compilation, only the (current) IR is considered

> In practice, there are, of course, exceptions to the general rule; sometimes an IR contains references to a previous/higher-level IR. An example is LLVM's low-level Machine IR, which only represents single functions and therefore references to global variables use the higher-level LLVM IR.

- Goal: simplify analyses/transformations
  - *Technically*, single-step compilation is possible for, e.g., C ... but optimizations are hard without proper IRs
- Compilers *design* IRs to support frequent operations
  - IR design can vary strongly between compilers
- Typically based on **graphs** or **linear instructions** (or both)

[Slide 73] Compiler Design: Effect of Languages – Imperative

- Step-by-step execution of program ⇝ modification of state
- Close to hardware execution model
- Direct influence of result

- Tracking of state is complex
- Dynamic typing: more complexity
- Limits optimization possibilities

```
void addvec(int* a, const int* b) {
  for (unsigned i = 0; i < 4; i++)
    a[i] += b[i]; // vectorizable?
}
func:
  mov [rdi], rsi
  mov [rdi+8], rdx
  mov [rdi], 0 // redundant?
  ret
```

> Tracking state, especially when memory is involved, is one of the main challenges during optimization. In the first example, the loop is not easily vectorizable, because `a` and `b` could point to the same underlying array (e.g., with `addvec(buf + 1, buf)`).

[Slide 74] Compiler Design: Effect of Languages – Declarative

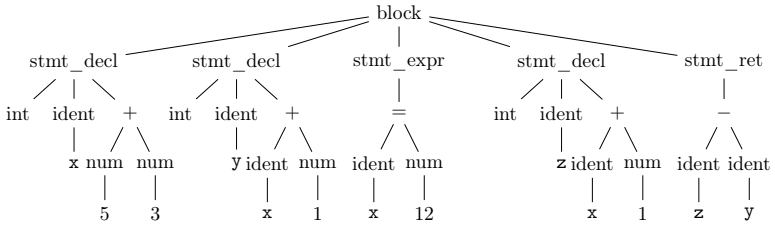
- Describes execution target
- Compiler has to derive good mapping to imperative hardware

- Allows for more optimizations
- Mapping to hardware non-trivial
  - Might need more stages
  - Preserve semantic info for opt!
- Programmer has less "control"

```
select s.name
from studenten s
where exists (select 1
              from hoeren h
              where h.matrno=s.matrno)
let rec fac = function
    | 0 | 1 -> 1
    | n -> n * fac (n - 1)
```

[Slide 75] Graph IRs: Abstract Syntax Tree (AST)

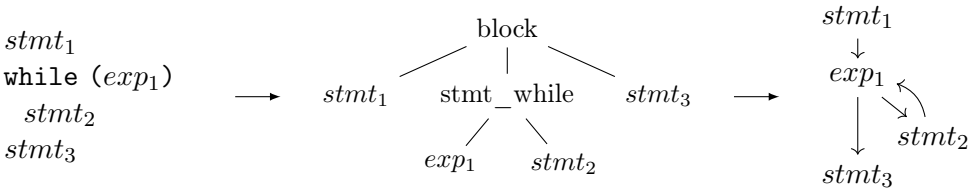
- Code representation close to the source
- Representation of types, constants, etc. might differ
- Storage might be problematic for large inputs



3.2. Control Flow Graph

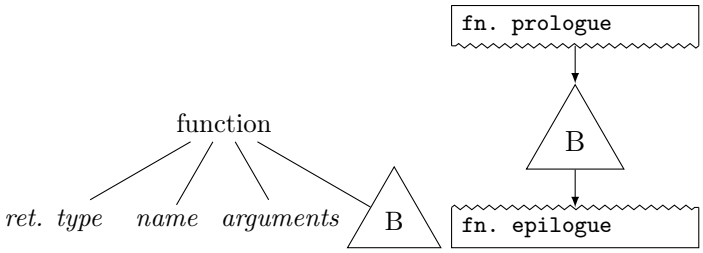
[Slide 76] Graph IRs: Control Flow Graph (CFG)

- Motivation: model control flow between different code sections
- Graph nodes represent **basic blocks**
  - Basic block: sequence of branch-free code (modulo exceptions)
  - Typically represented using a linear IR

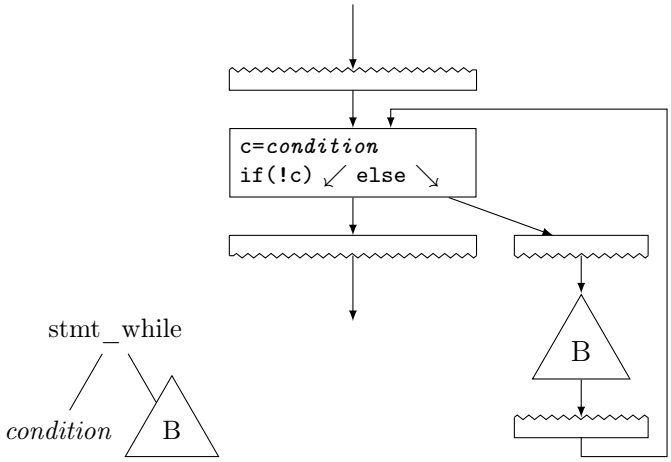


[Slide 77] Build CFG from AST – Function

- Idea: Keep track of current insert block while walking through AST



[Slide 78] Build CFG from AST – While Loop

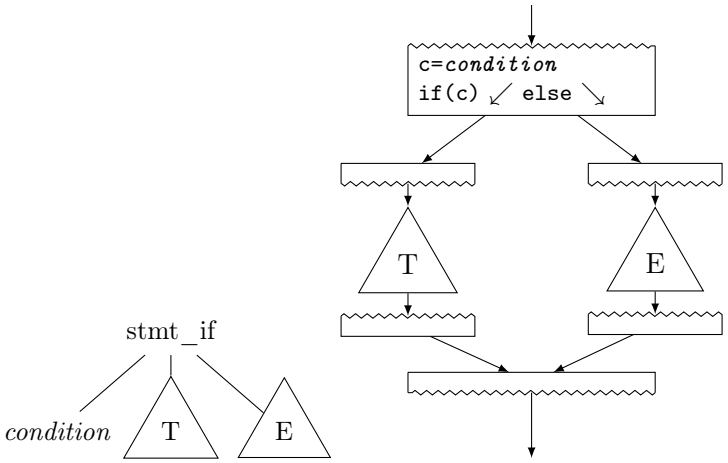


Written in pseudo-code:

```
IRValue generateCFG(ASTNode* node, BasicBlock*& insPos) {
  switch (node->kind()) {
  case ASTNode::Function:
    insPos = generatePrologue(node);
    generateCFG(node->child(0), insPos);
    generateEpilogue(insPos);
    return nullptr;
  case ASTNode::Block:
    for (ASTNode* child : node->children())
      generateCFG(child, insPos);
    return nullptr;
  case ASTNode::While: {
    BasicBlock* cond = newBlock();
    BasicBlock* body = newBlock();
    BasicBlock* end = newBlock();
    branchTo(insPos, cond);
    insPos = cond;
    IRValue brcond = generateCFG(node->child(0), insPos);
    // NB: generateCFG can modify insPos
    branchToCond(insPos, brcond, body, end);
    insPos = body;
```

```
    generateCFG(node->child(1), insPos);
    branchTo(insPos, cond);
    insPos = end;
    return nullptr;
  }
  // ...
  }
}
```

[Slide 79] Build CFG from AST – If Condition



[Slide 80] Build CFG from AST: Switch

Linear search

```
t ← exp
if t == 3: goto B3
if t == 4: goto B4
if t == 7: goto B7
if t == 9: goto B9
goto BD
```

+ Trivial
− Slow, lot of code

Binary search

```
t ← exp
if t == 7: goto B7
elif t > 7:
  if t == 9: goto B9
else:
  if t == 3: goto B3
  if t == 4: goto B4
goto BD
```

+ Good: sparse values
− Even more code

Jump table

```
t ← exp
if 0 ≤ t < 10:
  goto table[t]
goto BD

table = {
  BD, BD, BD, B3,
  B4, BD, ... }
```

+ Fastest
− Table can be large, needs ind. jump

> Typically, it is beneficial to keep switches in a high-level form during optimizations, as they are more expressive than, e.g., a corresponding search tree. The corresponding node in the CFG has then one outgoing edge per case plus the edge for the default target.

[Slide 81] Build CFG from AST: Break, Continue, Goto, Computed Goto

- `break`/`continue`: trivial

  - Keep track of target block, insert branch
- `goto`: also trivial
  - Split block at target label, if needed
  - But: may lead to irreducible control flow graph (see later)
- Computed `goto`: trivial is bad
  - Split block at target label, if needed
  - Every computed `goto` is a branch to every address-taken label
  - CFG can grow extremely dense for dispatch code!

> Very dense CFGs are problematic for the runtime of many graph algorithms — many algorithms need at the very least to consider every edge, so the runtime of such algorithms grows by at least $\Omega(|N|^2)$, often more. Typically, therefore, only a single computed `goto` is emitted into the IR as a separate basic block; all computed `goto`s in the code are re-routed to go through this new block. Only very late in the code generation pipeline, this block is duplicated again into all its predecessors to produce the expected behavior. (Note that, due to compiler bugs, this might not always happen.)

[Slide 82] CFG: Formal Definition

- **Flow graph:** $G = (N, E, s)$ with a digraph (N, E) and entry $s \in N$
  - Each node is a basic block, s is the entry block
  - $(n_1, n_2) \in E$ iff n_2 might be executed immediately after n_1
  - All $n \in N$ shall be reachable from s (unreachable nodes can be discarded)
  - Nodes without successors are end points

[Slide 83] CFG from C – Example

In-Class Exercise:

Derive the CFG for the these functions. Assume a `switch` instruction exists.

```
int fn1() {
  if (a()) {
    while (b()) {
      c();
      if (d())
        continue;
      e();
    }
  } else {
    f();
  }
  return g();
}
```

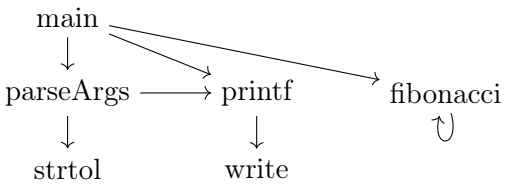
```
int fn2() {
  a();
  do switch (c()) {
  case 1:
    while (d()) {
      e();
    case 2:
      f();
    }
  default:
    g();
  } while (h());
  return b();
}
```

Solution on page 85.

3.3. Other Graph IRs

[Slide 84] Graph IRs: Call Graph

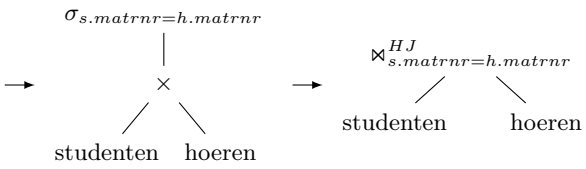
- Graph showing (possible) call relations between functions
- Useful for interprocedural optimizations
  - Function ordering
  - Stack depth estimation
  - ...



[Slide 85] Graph IRs: Relational Algebra

- Higher-level representation of query plans
  - Explicit data flow
- Allow for optimization and selection actual implementations
  - Elimination of common sub-trees
  - Joins: ordering, implementation, etc.

```
SELECT s.name, h.vorlnr
FROM studenten s, hoeren h
WHERE s.matrnr = h.matrnr
```



3.4. Linear IRs

[Slide 86] Linear IRs: Stack Machines

- Operands stored on a stack
- Operations pop arguments from top and push result
- Typically accompanied with variable storage
- Generating IR from AST: trivial
- Often used for bytecode, e.g. Java, Python

\+ Compact code, easy to generate and implement
\− Performance, hard to analyze

```
push 5
push 3
add
pop x
```

```
push x
push 1
add
pop y
push 12
pop x
push x
push 1
add
pop z
```

[Slide 87] Linear IRs: Register Machines

- Operands stored in registers
- Operations read and write registers
- Typically: infinite number of registers
- Typically: three-address form
  - *dst = src1 op src2*
- Generating IR from AST: trivial
- E.g., GIMPLE, eBPF, Assembly

$$
\begin{array}{llllll}
x & \leftarrow & 5 & + & 3 \\
y & \leftarrow & x & + & 1 \\
x & \leftarrow & 12 & & \\
z & \leftarrow & x & + & 1 \\
tmp_1 & \leftarrow & z & - & y \\
\text{return} & & tmp_1 & &
\end{array}
$$

[Slide 88] Example: High GIMPLE

```
int foo(int n) {
  int res = 1;
  while (n) {
    res *= n * n;
    n -= 1;
  }
  return res;
}
int fac (int n)
gimple_bind < // <-- still has lexical scopes
  int D.1950;
  int res;

  gimple_assign <integer_cst, res, 1, NULL, NULL>
  gimple_goto <<D.1947>>
  gimple_label <<D.1948>>
  gimple_assign <mult_expr, _1, n, n, NULL>
  gimple_assign <mult_expr, res, res, _1, NULL>
  gimple_assign <plus_expr, n, n, -1, NULL>
  gimple_label <<D.1947>>
  gimple_cond <ne_expr, n, 0, <D.1948>, <D.1946>>
```

```
  gimple_label <<D.1946>>
  gimple_assign <var_decl, D.1950, res, NULL, NULL>
  gimple_return <D.1950>
>
```

```
$ gcc -fdump-tree-gimple-raw -c foo.c
```

[Slide 89] Example: Low GIMPLE

```
int foo(int n) {
  int res = 1;
  while (n) {
    res *= n * n;
    n -= 1;
  }
  return res;
}
int fac (int n)
{
  int res;
  int D.1950;

  gimple_assign <integer_cst, res, 1, NULL, NULL>
  gimple_goto <<D.1947>>
  gimple_label <<D.1948>>
  gimple_assign <mult_expr, _1, n, n, NULL>
  gimple_assign <mult_expr, res, res, _1, NULL>
  gimple_assign <plus_expr, n, n, -1, NULL>
  gimple_label <<D.1947>>
  gimple_cond <ne_expr, n, 0, <D.1948>, <D.1946>>
  gimple_label <<D.1946>>
  gimple_assign <var_decl, D.1950, res, NULL, NULL>
  gimple_goto <<D.1951>>
  gimple_label <<D.1951>>
  gimple_return <D.1950>
}
```

```
$ gcc -fdump-tree-lower-raw -c foo.c
```

[Slide 90] Example: Low GIMPLE with CFG

```
int foo(int n) {
  int res = 1;
  while (n) {
    res *= n * n;
    n -= 1;
  }
  return res;
}
int fac (int n) {
  int res;
  int D.1950;
  <bb 2> :
  gimple_assign <integer_cst, res, 1, NULL, NULL>
  goto <bb 4>; [INV]
  <bb 3> :
  gimple_assign <mult_expr, _1, n, n, NULL>
  gimple_assign <mult_expr, res, res, _1, NULL>
  gimple_assign <plus_expr, n, n, -1, NULL>
  <bb 4> :
```

```
  gimple_cond <ne_expr, n, 0, NULL, NULL>
    goto <bb 3>; [INV]
  else
    goto <bb 5>; [INV]
  <bb 5> :
  gimple_assign <var_decl, D.1950, res, NULL, NULL>
  <bb 6> :
gimple_label <<L3>>
  gimple_return <D.1950>
}
$ gcc -fdump-tree-cfg-raw -c foo.c
```

[Slide 91] Linear IRs: Register Machines

- Problem: no clear def–use information
  - Is $x + 1$ the same?
  - Hard to track actual values!
- **How to optimize?**

⇒ Disallow mutations of variables

$$
\begin{array}{llllll}
x & \leftarrow & 5 & + & 3 \\
y & \leftarrow & x & + & 1 \\
x & \leftarrow & 12 & & \\
z & \leftarrow & x & + & 1 \\
tmp_1 & \leftarrow & z & - & y \\
\text{return} & & tmp_1 & &
\end{array}
$$

3.5. Single Static Assignment

[Slide 92] Single Static Assignment: Introduction

- Idea: disallow mutations of variables, value set in declaration
- Instead: create new variable for updated value
- SSA form: every computed value has a unique definition
  - Equivalent formulation: each name describes result of one operation

$$
\begin{array}{llllll}
x & \leftarrow & 5 & + & 3 \\
y & \leftarrow & x & + & 1 \\
x & \leftarrow & 12 & & \\
z & \leftarrow & x & + & 1 \\
tmp_1 & \leftarrow & z & - & y \\
\text{return} & & tmp_1 & &
\end{array}
\quad \longrightarrow \quad
\begin{array}{llllll}
v_1 & \leftarrow & 5 & + & 3 \\
v_2 & \leftarrow & v_1 & + & 1 \\
v_3 & \leftarrow & 12 & & \\
v_4 & \leftarrow & v_3 & + & 1 \\
v_5 & \leftarrow & v_4 & - & v_2 \\
\text{return} & & v_5 & &
\end{array}
$$

[Slide 93] Single Static Assignment: Control Flow

- How to handle diverging values in control flow?

- Solution: Φ-nodes to merge values depending on predecessor
  - Value depends on edge used to enter the block
  - All Φ-nodes of a block execute concurrently (ordering irrelevant)

$$
\begin{array}{rl}
\mathit{entry}: & x \leftarrow \ldots \\
 & \text{if } (x > 2) \text{ goto } \mathit{cont} \\
\mathit{then}: & x \leftarrow x * 2 \\
\mathit{cont}: & \text{return } x
\end{array}
$$

$$\rightarrow$$

$$
\begin{array}{rl}
\mathit{entry}: & v_1 \leftarrow \ldots \\
 & \text{if } (v_1 > 2) \text{ goto } \mathit{cont} \\
\mathit{then}: & v_2 \leftarrow v_1 * 2 \\
\mathit{cont}: & v_3 \leftarrow \Phi(\mathit{entry}: v_1, \mathit{then}: v_2) \\
 & \text{return } v_3
\end{array}
$$

[Slide 94] Example: GIMPLE in SSA form

```
int foo(int n) {
  int res = 1;
  while (n) {
    res *= n * n;
    n -= 1;
  }
  return res;
}
int fac (int n) { int res, D.1950, _1, _6;
  <bb 2> :
  gimple_assign <integer_cst, res_4, 1, NULL, NULL>
  goto <bb 4>; [INV]
  <bb 3> :
  gimple_assign <mult_expr, _1, n_2, n_2, NULL>
  gimple_assign <mult_expr, res_8, res_3, _1, NULL>
  gimple_assign <plus_expr, n_9, n_2, -1, NULL>
  <bb 4> :
  # gimple_phi <n_2, n_5(D)(2), n_9(3)>
  # gimple_phi <res_3, res_4(2), res_8(3)>
  gimple_cond <ne_expr, n_2, 0, NULL, NULL>
    goto <bb 3>; [INV]
  else
    goto <bb 5>; [INV]
  <bb 5> :
  gimple_assign <ssa_name, _6, res_3, NULL, NULL>
  <bb 6> :
gimple_label <<L3>>
  gimple_return <_6>
}
```

```
$ gcc -fdump-tree-ssa-raw -c foo.c
```

[Slide 95] SSA Construction – Local Value Numbering

- Simple case: inside block – keep mapping of variable to value

Code

$$
\begin{array}{llllll}
x & \leftarrow & 5 & + & 3 \\
y & \leftarrow & x & + & 1 \\
x & \leftarrow & 12 & & \\
z & \leftarrow & x & + & 1 \\
tmp_1 & \leftarrow & z & - & y \\
\text{return} & & tmp_1 &
\end{array}
$$

SSA IR

$$
\begin{array}{lll}
v_1 & \leftarrow & \texttt{add } 5, 3 \\
v_2 & \leftarrow & \texttt{add } v_1, 1 \\
v_3 & \leftarrow & \texttt{const } 12 \\
v_4 & \leftarrow & \texttt{add } v_3, 1 \\
v_5 & \leftarrow & \texttt{sub } v_4, v_2 \\
 & & \texttt{ret } v_5
\end{array}
$$

Variable Mapping

$$
\begin{array}{lll}
x & \rightarrow & v_3 \\
y & \rightarrow & v_2 \\
z & \rightarrow & v_4 \\
tmp_1 & \rightarrow & v_5
\end{array}
$$

[Slide 96] SSA Construction – Across Blocks

- SSA construction with control flow is non-trivial
- Key problem: find value for variable in predecessor
- Naive approach: Φ-nodes for all variables everywhere
  - Create empty Φ-nodes for variables, populate variable mapping
  - Fill blocks (as on last slide)
  - Fill Φ-nodes with last value of variable in predecessor
- Why is this a bad idea? ⇒ *don't do this!*
  - *Extremely inefficient, code size explosion, many dead* Φ

[Slide 97] SSA Construction – Across Blocks ("simple"[1])

- Key problem: find value in predecessor
- Idea: *seal* block once all direct predecessors are known
  - For acyclic constructs: trivial
  - For loops: seal header once loop block is generated
- Current block not sealed: add Φ-node, fill on sealing
- Single predecessor: recursively query that
- Multiple preds.: add Φ-node, fill now

> Confer the (very readable) paper for a more formal specification of the algorithm. The removal of trivial and redundant Φ-nodes is not strictly required.

[Slide 98] SSA Construction – Example

```
int foo(int n) {
  int res = 1;
  while (n) {
    res *= n * n;
```

[1] M Braun et al. "Simple and efficient construction of static single assignment form". In: *CC*. 2013, pp. 102–122. URL: https://link.springer.com/content/pdf/10.1007/978-3-642-37051-9_6.pdf.

```
    n -= 1;
  }
  return res;
}
```

```
            func foo(v1)
  entry:    sealed; varmap: n→ v1, res→ v2
            v2 ← 1
 header:    sealed; varmap: n→ φ1, res→ φ2
            φ1 ← φ(entry: v1, body: v6)
            φ2 ← φ(entry: v2, body: v5)
            v3 ← equal φ1, 0
            br v3, cont, body
   body:    sealed; varmap: n→v6, res→ v5
            v4 ← mul φ1, φ1
            v5 ← mul φ2, v4
            v6 ← sub φ1, 1
            br header
   cont:    sealed; varmap: res→ φ2
            ret φ2
```

Note: the *sealed* state and the variable map are only used during SSA construction. They can be discarded afterwards and are *not* part of the IR. The information is also not sufficient for debug information, for that purpose as it doesn't sufficiently cover all program points; tracking variable states for debug information typically requires extra instructions or annotations in the IR.

[Slide 99] SSA Construction – Example

In-Class Exercise:

Construct an IR in SSA form for the following C functions.

```
int phis(int a, int b){
  a = a * b;
  if (a > b * b) {
    int c = 1;
    while (a > 0)
      a = a - c;
  } else {
    a = b * b;
  }
  return a;
}
```

```
int swap(int a, int b, int c) {
  while (c > 0) {
    int tmp = a;
    a = b;
    b = tmp;
    c = c - 1;
  }
  return a;
}
```

Solution on page 85.

[Slide 100] SSA Construction – Pruned/Minimal Form

- Resulting SSA is *pruned* – all ϕ are used
- But not *minimal* – ϕ nodes might have single, unique value
- When filling ϕ, check that multiple real values exist
  - Otherwise: replace ϕ with the single value

  - On replacement, update all ϕ using this value, they might be trivial now, too
- Sufficient? Not for irreducible CFG
  - Needs more complex algorithms[2] or different construction method[3]

AD IN2053 "Program Optimization" covers this more formally

[Slide 101] SSA: Implementation

- Value is often just a reference to the instruction
- ϕ nodes placed at beginning of block
  - They execute "concurrently" and on the edges, after all
- Variable number of operands required for ϕ nodes
- Storage format for instructions and basic blocks
  - Consecutive in memory: hard to modify/traverse
  - Array of pointers: $\mathcal{O}(n)$ for a single insertion...
  - Linked List: easy to insert, but pointer overhead

Is SSA a graph IR?

Only if instructions have no side effects,consider `load`, `store`, `call`, ...
These *can* be solved using explicit dependencies as SSA values, e.g. for memory

3.6. IR Design

[Slide 103] IR Design: High-level Considerations

- Define purpose!
- Structure: SSA vs. something else; control flow
  - Control flow: basic blocks/CFG vs. structured control flow
  - Remember: SSA can be considered as a DAG, too
  - SSA is easy to analyse, but non-trivial to construct/leave
- Broader integration: keep multiple stages in single IR?
  - Example: create IR with high-level operations, then incrementally lower
  - Model machine instructions in same IR?
  - Can avoid costly transformations, but adds complexity

[2] M Braun et al. "Simple and efficient construction of static single assignment form". In: *CC*. 2013, pp. 102–122. URL: https://link.springer.com/content/pdf/10.1007/978-3-642-37051-9_6.pdf.
[3] R Cytron et al. "Efficiently computing static single assignment form and the control dependence graph". In: *TOPLAS* 13.4 (1991), pp. 451–490. URL: https://dl.acm.org/doi/pdf/10.1145/115372.115320.

[Slide 104] IR Design: Operations

- Data types
  - Simple type structure vs. complex/aggregate types?
  - Keep relation to high-level types vs. low-level only?
  - Virtual data types, e.g. for flags/memory?
- Instruction format
  - Single vs. multiple results?
  - Strongly typed vs. more generic result/operand types?
  - Operand number – fixed vs. dynamic?

[Slide 105] IR Design: Operations

- Allow instruction side effects?
  - E.g.: memory, floating-point arithmetic, implicit control flow
- Operation complexity and abstraction
  - E.g.: `CheckBounds`, `GetStackPtr`, `HashInt128`
  - E.g.: `load` vs. `MOVQconstidx4`
- Extensibility for new operations (e.g., new targets, high-level ops)

[Slide 106] IR Design: Desired Operations on IR

- Replacing all uses of an instruction with a different value
- Inserting an instruction or phi node/block argument
- Removing an instruction or phi node/block argument
- Changing the operand of an instruction
- Finding predecessors and successors of a basic block
- Finding all users of an instruction result
- For optimization-focused IRs, all these operations should be fast

[Slide 107] IR Design: Implementation

- Maintain user lists?
  - Simplifies optimizations, but adds considerable overhead
  - Replacement can use `copy` and lazy canonicalization
  - User *count* might be sufficient alternative
- Storage layout: operation size and locations
  - For performance: reduce heap allocations, small data structures
- Special handling for arguments vs. all-instructions?
- Metadata for source location, register allocation, etc.
- SSA: ϕ nodes vs. block arguments?

[Slide 108] IR Example: Go SSA

- Strongly typed
  - Structured types decomposed
- Explicit memory side-effects
- Also High-level operations
  - `IsInBounds`, `VarDef`
- Only one type of value/instruction
  - `Const64`, `Arg`, `Phi`
- No user list, but user count
- Also used for arch-specific repr.

```
env GOSSAFUNC=fac go build test.go
b1:
    v1 (?) = InitMem <mem>
    v2 (?) = SP <uintptr>
    v5 (?) = LocalAddr <*int> {~r1} v2 v1
    v6 (7) = Arg <int> {n} (n[int])
    v8 (?) = Const64 <int> [1] (res[int])
    v9 (?) = Const64 <int> [2] (i[int])
Plain -> b2 (+9)
b2: <- b1 b4
    v10 (9) = Phi <int> v9 v17 (i[int])
    v23 (12) = Phi <int> v8 v15 (res[int])
    v12 (+9) = Less64 <bool> v10 v6
If v12 -> b4 b5 (likely) (9)
b4: <- b2
    v15 (+10) = Mul64 <int> v23 v10 (res[int])
    v17 (+9) = Add64 <int> v10 v8 (i[int])
Plain -> b2 (9)
b5: <- b2
    v20 (12) = VarDef <mem> {~r1} v1
    v21 (+12) = Store <mem> {int} v5 v23 v20
Ret v21 (+12)
```

[Slide 109] Intermediate Representations – Summary

- An IR is an internal representation of a program
- Main goal: simplify analyses and transformations
- IRs typically based on graphs or linear instructions
- Graph IRs: AST, Control Flow Graph, Relational Algebra
- Linear IRs: stack machines, register machines, SSA
- Single Static Assignment makes data flow explicit
- SSA is extremely popular, although non-trivial to construct
- IR design depends on purpose and integration constraints

[Slide 110] Intermediate Representations – Questions

- Who designs an IR? What are design criteria?
- Why is an AST not suited for program optimization?
- How to convert an AST to another IR?

- What are the benefits/drawbacks of stack/register machines?
- What benefits does SSA offer over a normal register machine?
- How do ϕ-instructions differ from normal instructions?

4. LLVM-IR

4.1. Overview

[Slide 112] LLVM[1]

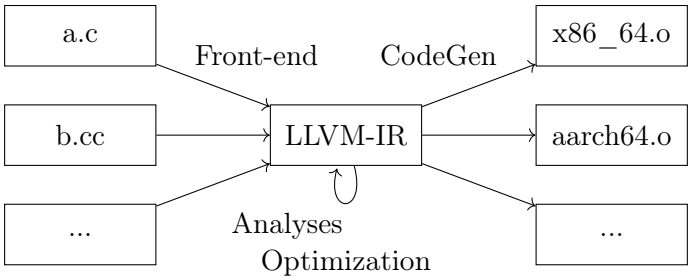
LLVM "Core" Library

- Optimizer and compiler back-end
- "Set of compiler components"
  - IRs: LLVM-IR, SelDag, MIR
  - Analyses and Optimizations
  - Code generation back-ends
- Started from Chris Lattner's master's thesis
- Used for C, C++, Swift, D, Julia, Rust, Haskell, ...

LLVM Project

- Umbrella for several projects related to compilers/toolchain
  - LLVM Core
  - Clang: C/C++ front-end for LLVM
  - libc++, compiler-rt: runtime support
  - LLDB: debugger
  - LLD: linker
  - MLIR: experimental IR framework

[Slide 113] LLVM: Overview



- Independent front-end derives LLVM-IR, LLVM does opt. and code gen.
- LTO: dump LLVM-IR into object file, optimize at link-time

[1] C Lattner and V Adve. "LLVM: A compilation framework for lifelong program analysis & transformation". In: *CGO*. 2004, pp. 75–86. URL: http://www.llvm.org/pubs/2004-01-30-CGO-LLVM.pdf.

The single IR allows multiple front-ends to reuse the same back-end infrastructure. Thus, generating LLVM-IR provides an easy way to target a wide range of architectures.

For link-time optimization, the LLVM-IR is stored in the object files instead of the machine code. At link-time, a linker plugin detects these files, merges the LLVM-IR from all object files, and then runs the actual compilation as part of the linking step. We will look at LTO again later when discussing object file generation and linking.

4.2. LLVM-IR

[Slide 114] LLVM-IR: Overview

- SSA-based IR, representations textual, bitcode, in-memory
- Hierarchical structure
  - Module
  - Functions, global variables
  - Basic blocks
  - Instructions
- Strongly/strictly typed

```
define dso_local i32 @foo(i32 %0) {
  %2 = icmp eq i32 %0, 0
  br i1 %2, label %10, label %3

3: ; preds = %1, %3
  %4 = phi i32 [ %7, %3 ], [ 1, %1 ]
  %5 = phi i32 [ %8, %3 ], [ %0, %1 ]
  %6 = mul nsw i32 %5, %5
  %7 = mul nsw i32 %6, %4
  %8 = add nsw i32 %5, -1
  %9 = icmp eq i32 %8, 0
  br i1 %9, label %10, label %3

10: ; preds = %3, %1
  %11 = phi i32 [ 1, %1 ], [ %7, %3 ]
  ret i32 %11
}
```

[Slide 115] LLVM-IR: Data types

- First class types:
  - `i<N>` – arbitrary bit width integer, e.g. `i1`, `i25`, `i1942652`
  - `ptr`/`ptr addrspace(1)` – pointer with optional address space
  - `float`/`double`/`half`/`bfloat`/`fp128`/...
  - `<N x ty>` – vector type, e.g. `<4 x i32>`
- Aggregate types:
  - `[N x ty]` – constant-size array type, e.g. `[32 x float]`
  - `{ ty, ... }` – struct (can be packed/opaque), e.g. `{i32, float}`

- Other types:
  - `ty (ty, ...)` – function type, e.g. `{i32, i32} (ptr, ...)`
  - `void`
  - `label`/`token`/`metadata`

> Although structure types can be used in various places in the IR, e.g., a single instruction to load a large structure from memory, this is strongly discouraged: LLVM is not optimized for this and both code quality and compile times get considerably worse. Only use struct types for globals and to implement multiple return values.

[Slide 116] LLVM-IR: Modules

- Top-level entity, one compilation unit – akin to C/C++
- Contains global values, specified with linkage type
- Global variable declarations/definitions

```
@externInt = external global i32, align 4
@globVar = global i32 4, align 4
@staticPtr = internal global ptr null, align 8
```

- Function declarations/definitions

```
declare i32 @readPtr(ptr)
define i32 @return1() {
  ret i32 1
}
```

- Global named metadata (discarded during compilation)

[Slide 117] LLVM-IR: Functions

- Functions definitions contain all code, not nestable
- Single return type (or `void`), multiple parameters, list of basic blocks
  - No basic blocks ⇒ function declaration
- Specifiers for callconv, section name, other attributes
  - E.g.: `noinline`/`alwaysinline`, `noreturn`, `readonly`
- Parameter and return can also have attributes
  - E.g.: `noalias`, `nonnull`, `sret(<ty>)`

[Slide 118] LLVM-IR: Basic Block

- Sequence of instructions
  - ϕ nodes come first
  - Regular instructions come next
  - Must end with a terminator
- First block in function is entry block Entry block cannot be branch target

[Slide 119] LLVM-IR: Instructions – Control Flow and Terminators

- Terminators end a block/modify control flow
- `ret <ty> <val>`/`ret void`
- `br label <dest>`/`br i1 <cond>, label <then>, label <else>`
- `switch`/`indirectbr`
- `unreachable`
- Few others for exception handling
- Not a terminator: `call`

> Although `call` does modify control flow in some sense, the assumption is that every function call returns ordinarily. When special control flow for exceptions is needed, the `invoke` instruction is used, which specifies one basic block as successor for the ordinary case and one basic block for the exceptional case.

[Slide 120] LLVM-IR: Instructions – Arithmetic-Logical

- `add`/`sub`/`mul`/`udiv`/`sdiv`/`urem`/`srem`
  - Arithmetic uses two's complement
  - Division corner cases are *undefined behavior*
- `fneg`/`fadd`/`fsub`/`fmul`/`fdiv`/`frem`
- `shl`/`lshr`/`ashr`/`and`/`or`/`xor`
  - Out-of-range shifts have an undefined result
- `icmp <pred>`/`fcmp <pred>`/`select <cond>, <then>, <else>`
- `trunc`/`zext`/`sext`/`fptrunc`/`fpext`/`fptoui`/`fptosi`/`uitofp`/`sitofp`
- `bitcast`
  - Cast between equi-sized datatypes by reinterpreting bits

> Technically, out-of-range shifts return `poison`, see below.

[Slide 121] LLVM-IR: Instructions – Memory and Pointer

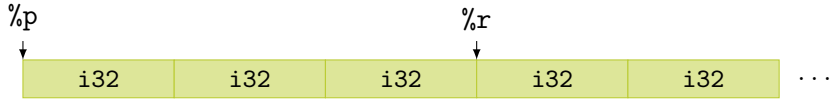
- `alloca <ty>` – allocate addressable stack slot
- `load <ty>, ptr <ptr>`/`store <ty> <val>, ptr <ptr>`
  - May be `volatile` (e.g., MMIO) and/or `atomic`
- `cmpxchg`/`atomicrmw` – similar to hardware operations
- `ptrtoint`/`inttoptr`
- `getelementptr` – address computation on `ptr`/structs/arrays

> `alloca` allocates stack memory whenever it is executed. There is a difference between statically-sized `alloca`s that are in the entry block versus in dynamically-sized `alloca`s or `alloca`s in other blocks:

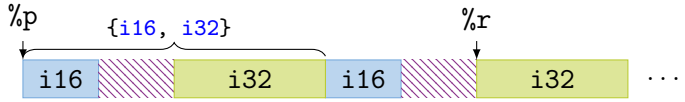
- Static `allocas` become part of the fixed-size stack frame.
- Dynamic `allocas` require a dynamically-sized stack frame, which is less efficient (use of a frame pointer is required) and can cause unbounded stack growth.

[Slide 122] LLVM-IR: `getelementptr` Examples

- `%r = getelementptr i32, ptr %p, i64 3`



  Equivalent in C: `&((int*) p)[3]`
- `%r = getelementptr {i16, i32}, ptr %p, i64 1, i32 1`



  Equivalent in C: `&((struct {short _0; int _1;}*) p)[1]._1`
- Also works with nested structs and arrays (and vectors)
- Pointer and array can be dynamic, struct indices must be constant

[Slide 123] LLVM-IR: `undef` and `poison`

- `undef` – unspecified value, compiler may choose any value
  - `%b = add i32 %a, i32 undef` → `i32 undef`
  - `%c = and i32 %a, i32 undef` → `i32 %a`
  - `%d = xor i32 %b, i32 %b` → `i32 undef`
  - `br i1 undef, label %p, label %q` → *undefined behavior*
- `poison` – result of erroneous operations
  - Delay *undefined behavior* on illegal operation until actually relevant
  - Allows to speculatively "execute" instructions in IR
  - `%d = shl i32 %b, i32 34` → `i32 poison`

[Slide 124] LLVM-IR: Intrinsics

- Not all operations provided as instructions
- Intrinsic functions: special functions with defined semantics
  - Replaced during compilation, e.g., with instruction or lib call
- Benefit: no changes needed for parser/bitcode/... on addition
- Examples:
  - `declare iN @llvm.ctpop.iN(iN <src>)`
  - `declare {iN, i1} @llvm.sadd.with.overflow.iN(iN %a, iN %b)`
  - `memcpy`, `memset`, `sqrt`, `returnaddress`, ...

[Slide 125] LLVM-IR: Tools

- `clang` can emit LLVM-IR bitcode `clang -O -emit-llvm -c test.c -o test.bc`

- `llvm-dis` disassembles bitcode to textual LLVM-IR `clang -O -emit-llvm -c test.c -o - | llvm-dis`
- `llc` compiles LLVM-IR (textual or bitcode) to assembly `clang -O -emit-llvm -c test.c -o - | llc clang -O -emit-llvm -c test.c -o - | llvm-dis | llc`

Example Listings omitted – they would span several slides

> Note: `-emit-llvm` emits the LLVM IR after optimizations. To get the IR before optimizations, use `-Xclang -disable-llvm-passes`. The LLVM IR generated when compiling with `-O0` is different.

[Slide 126] LLVM-IR: Example

```
define <4 x float> @foo2(<4 x float> %0, <4 x float> %1) {
  %3 = alloca <4 x float>, align 16
  %4 = alloca <4 x float>, align 16
  store <4 x float> %0, ptr %3, align 16
  store <4 x float> %1, ptr %4, align 16
  %5 = load <4 x float>, ptr %3, align 16
  %6 = load <4 x float>, ptr %4, align 16
  %7 = fadd <4 x float> %5, %6
  ret <4 x float> %7
}
```

[Slide 127] LLVM-IR: Example

```
define i32 @foo3(i32 %0, i32 %1) {
  %3 = tail call { i32, i1 } @llvm.smul.with.overflow.i32(i32 %0, i32 %1)
  %4 = extractvalue { i32, i1 } %3, 1
  %5 = extractvalue { i32, i1 } %3, 0
  %6 = select i1 %4, i32 -2147483648, i32 %5
  ret i32 %6
}
```

[Slide 128] LLVM-IR: Example

In-Class Exercise:

```
define i32 @sw(i32 %0) {
  switch i32 %0, label %4 [
    i32 4, label %5
    i32 5, label %2
    i32 8, label %3
    i32 100, label %5
  ]
2: ; preds = %1
  br label %5
3: ; preds = %1
  br label %5
4: ; preds = %1
  br label %5
5: ; preds = %1, %1, %4, %3, %2
```

```
  %6 = phi i32 [ %0, %4 ], [ 9, %3 ], [ 32, %2 ], [ 12, %1 ], [ 12, %1 ]
  ret i32 %6
}
```

Solution on page 86.

[Slide 129] LLVM-IR: Example

In-Class Exercise:

```
@a = private unnamed_addr constant [7 x i32] [i32 12, i32 32, i32 12,
                                              i32 12, i32 9, i32 12, i32 12], align 4
define dso_local i32 @f(i32 %0) {
  %2 = add i32 %0, -4
  %3 = icmp ult i32 %2, 7
  br i1 %3, label %4, label %13
4: ; preds = %1
  %5 = trunc i32 %2 to i8
  %6 = lshr i8 83, %5
  %7 = and i8 %6, 1
  %8 = icmp eq i8 %7, 0
  br i1 %8, label %13, label %9
9: ; preds = %4
  %10 = sext i32 %2 to i64
  %11 = getelementptr inbounds [7 x i32], ptr @a, i64 0, i64 %10
  %12 = load i32, ptr %11, align 4
  br label %13
13: ; preds = %1, %4, %9
  %14 = phi i32 [ %12, %9 ], [ %0, %4 ], [ %0, %1 ]
  ret i32 %14
}
```

Solution on page 86.

4.3. API

[Slide 130] LLVM-IR API

- LLVM offers two APIs: C++ and C
  - C++ is the full API, exposing nearly all internals
  - C API is more limited, but more stable
- Nearly all major versions have breaking changes
- Some support for multi-threading:
  - All modules/types/... associated with an `LLVMContext`
  - Different contexts may be used in different threads

[Slide 131] LLVM-IR C++ API: Basic Example

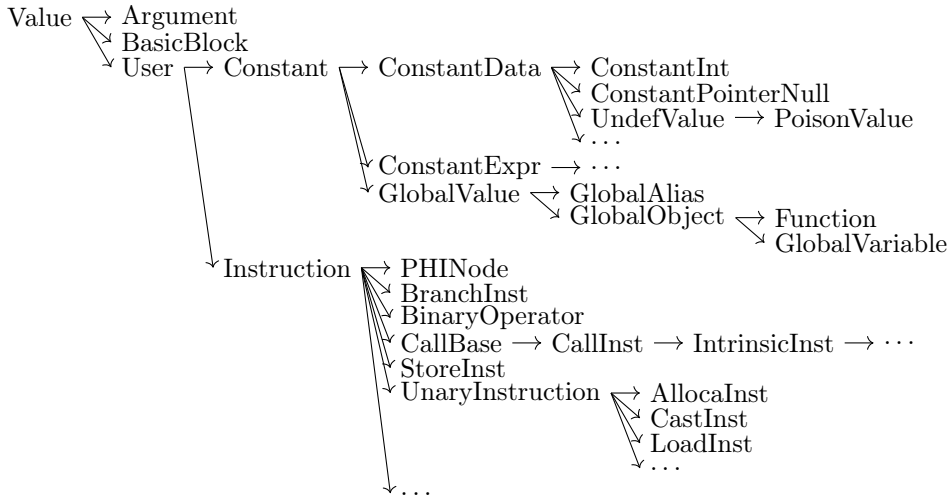
```
#include <llvm/IR/IRBuilder.h>
int main(void) {
  llvm::LLVMContext ctx;
  auto modUP = std::make_unique<llvm::Module>("mod", ctx);

  llvm::Type* i64 = llvm::Type::getInt64Ty(ctx);
```

```
  llvm::FunctionType* fnTy = llvm::FunctionType::get(i64, {i64}, false);
  llvm::Function* fn = llvm::Function::Create(fnTy,
            llvm::GlobalValue::ExternalLinkage, "addOne", modUP.get());
  llvm::BasicBlock* entryBB = llvm::BasicBlock::Create(ctx, "entry", fn);

  llvm::IRBuilder<> irb(entryBB);
  llvm::Value* add = irb.CreateAdd(fn->getArg(0), irb.getInt64(1));
  irb.CreateRet(add);
  modUP->print(llvm::outs(), nullptr);
  return 0;
}
```

[Slide 132] LLVM-IR API: Almost Everything is a Value... (excerpt)



> See LLVM Doxygen[a] for a full graph.
>
> [a] https://llvm.org/doxygen/classllvm_1_1Value.html

[Slide 133] LLVM-IR API: Programming Environment

- LLVM implements custom RTTI
  - `isa<>`, `cast<>`, `dyn_cast<>`
- LLVM implements a multitude of specialized data structures
  - E.g.: `SmallVector<T, N>` to keep N elements stack-allocated
  - Custom vectors, sets, maps; see manual[2]
- Preferably uses `ArrayRef`, `StringRef`, `Twine` for references
- LLVM implements custom streams instead of std streams
  - `outs()`, `errs()`, `dbgs()`

[2] https://www.llvm.org/docs/ProgrammersManual.html

Many of these data types are used for efficiency. Standard C++ RTTI is inefficient while LLVM's implementation is very flexible, fast, and has a low memory usage in data structures. The sub-class type of the value is stored in an 8-bit integer.

`SmallVector` is preferred over `std::vector` not just because of the inline storage, but also because (for non-`char` types) it only uses 32-bit integers for length/capacity (lower memory usage, often sufficient) and grows more efficiently for trivially movable data structures (`realloc`).

`Twine` is a lazily evaluated string. For example, when specifying `Twine("foo") + 5`, on-stack data structures are constructed to represent this sequence, but the resulting string is constructed only when and if it is actually used. This also allows constructing strings directly into target buffers.

Standard C++ streams are not just inefficient, implementations also tend to inject global constructors in all files. Therefore, LLVM has its own stream implementation. With `raw_svector_ostream` and `raw_string_ostream`, a `raw_ostream` can be used to write into a `SmallVector` or `std::string`.

[Slide 134] LLVM and IR Design

- LLVM provides a decent general-purpose IR for compilers
- But: not ideal for all purposes
  - High-level optimizations difficult, e.g. due to lost semantics
  - Several low-level operations only exposed as intrinsics
  - IR rather complex, high code complexity
  - High compilation times, not very efficient data structures
- Thus: heavy trend towards custom IRs

[Slide 135] LLVM-IR – Summary

- LLVM is a modular compiler framework
- Extremely popular and high-quality compiler back-end
- Primarily provides optimizations and a code generator
- Main interface is the SSA-based LLVM-IR
  - Easy to generate, friendly for writing front-ends/optimizations

[Slide 136] LLVM-IR – Questions

- What is the structure of an LLVM-IR module/function?
- Which LLVM-IR data types exist? How do they relate to the target architecture?
- How do semantically invalid operations in LLVM-IR behave?
- What is special about intrinsic functions?
- How to derive LLVM-IR from C code using Clang?

5. Analyses and Transformations

5.1. Motivation

[Slide 138] Program Transformation: Motivation

- "User code" is often not very efficient
- Also: no need to, compiler can (often?) optimize better
  - More knowledge: e.g., data layout, constants after inlining, etc.
- Allows for more pragmatic/simple code
- Generating "better" IR code on first attempt is expensive
  - What parts are actually used? How to find out?
- Transformation to "better" code must be done *somewhere*
- Optimization is a misnomer: we don't know whether it improves code!
  - Many transformations are driven by heuristics
- Many types of optimizations are well-known[1]

5.2. Dead Code Elimination

[Slide 139] Dead Block Elimination

- CFG not necessarily connected
- E.g., consequence of optimization
  - Conditional branch → unconditional branch
- Removing dead blocks is trivial
  1. DFS traversal of CFG from entry, mark visited blocks
  2. Remove unmarked blocks

[Slide 140] Optimization Example 1

```
define i32 @fac(i32 %0) {
  br label %for.header
for.header: ; preds = %for.body, %1
  %a = phi i32 [ 1, %1 ], [ %a.new, %for.body ]
  %b = phi i32 [ 0, %1 ], [ %b.new, %for.body ]
  %i = phi i32 [ 0, %1 ], [ %i.new, %for.body ]
  %cond = icmp sle i32 %i, %0
```

[1] FE Allen and J Cocke. *A catalogue of optimizing transformations*. 1971. URL: https://www.clear.rice.edu/comp512/Lectures/Papers/1971-allen-catalog.pdf.

```
  br i1 %cond, label %for.body, label %exit
for.body: ; preds = %for.header
  %a.new = mul i32 %a, %i
  %b.new = add i32 %b, %i
  %i.new = add i32 %i, 1
  br label %for.header
exit: ; preds = %for.header
  %absum = add i32 %a, %b
  ret i32 %a
}
```

[Slide 141] Simple Dead Code Elimination (DCE)

- Look for trivially dead instructions
  - No users or side-effects
  - Calls *might* be removed

1. Add all instructions to work queue
2. While work queue not empty:
   a) Check for deadness (zero users, no side-effects)
   b) If dead, remove and add all operands to work queue

Warning: Don't implement it this naively, this is inefficient

[Slide 142] Applying Simple DCE

```
          define i32 @fac(i32 %0) {
eff.: cf   br label %for.header
          for.header: ; preds = %for.body, %1
users: 3   %a = phi i32 [ 1, %1 ], [ %a.new, %for.body ]
users: 2   %b = phi i32 [ 0, %1 ], [ %b.new, %for.body ]
users: 4   %i = phi i32 [ 0, %1 ], [ %i.new, %for.body ]
users: 1   %cond = icmp sle i32 %i, %0
eff.: cf   br i1 %cond, label %for.body, label %exit
          for.body: ; preds = %for.header
users: 1   %a.new = mul i32 %a, %i
users: 1   %b.new = add i32 %b, %i
users: 1   %i.new = add i32 %i, 1
eff.: cf   br label %for.header
          exit: ; preds = %for.header
users: 0   %absum = add i32 %a, %b
eff.: cf   ret i32 %a
          }
```

> In this example, the instruction `%abssum` can be removed. This reduces the number of users of `%a` and `%b` by 1. As no other instructions have a user count of 0 after this change, the algorithm terminates.

[Slide 143] Dead Code Elimination

- Problem: unused value cycles

- Idea: find "value sinks" and mark all needed values as live unmarked values can be removed
  - Sink: instruction with side effects (e.g., store, control flow)
  - Alternative formulation: assume instruction is dead unless proven otherwise

1. Only mark instrs. with side effects as live
2. Populate work list with newly added live instrs.
3. While work list not empty:
   a) Mark dead operand instructions as live and add to work list
4. Remove instructions not marked as live

[Slide 144] Applying Liveness-based DCE

```
define i32 @fac(i32 %0) {
live   br1 label %for.header
for.header: ; preds = %for.body, %1
live   %a = phi i32 [ 1, %1 ], [ %a.new, %for.body ]

live   %i = phi i32 [ 0, %1 ], [ %i.new, %for.body ]
live   %cond = icmp sle i32 %i, %0
live   br2 i1 %cond, label %for.body, label %exit
for.body: ; preds = %for.header
live   %a.new = mul i32 %a, %i

live   %i.new = add i32 %i, 1
live   br3 label %for.header
exit: ; preds = %for.header

live   ret i32 %a
}
```

Work list (stack)

> This algorithm finds the dead value cycle pf `%b` from the previous example. (Refer to the slide deck for the animated version.)

5.3. Sparse Conditional Constant Propagation

[Slide 145] Optimization Example 2

```
define i32 @trivial() {
entry:
  br label %for.cond
for.cond:
  %x = phi i32 [ 0, %entry ], [ %inc, %for.inc ]
  %cmp = icmp ugt i32 %x, 0
  br i1 %cmp, label %for.inc, label %ret
for.inc:
  %inc = add i32 %x, 1
  br label %for.cond
ret:
```

```
  ret i32 %x
}
```

[Slide 146] Sparse Conditional Constant Propagation[2]

- Constant folding and dead code elimination separately insufficient
- Idea: be optimistic about reachability and constantness
  - Assume block is dead/value is constant unless proven othewise
  - Ignore edges that are not proven as executable
- For every value: store lattice element
  - Lattice: $\top$ (any value) $>$ *constant* (single constant value) $>$ $\bot$ (dynamic)
  - $\top \wedge x = x, \bot \wedge x = \bot, x \wedge y = (x \text{ if } x = y \text{ else } \bot)$
- For every edge: store executability
  - ϕ-nodes: edges that are not (yet) marked executable get $\top$

[Slide 147] SCCP Algorithm

State:

- EdgeWorklist
- InstrWorklist
- EdgeExecutable
  - Initialize `false`
- InstrLattice
  - Initialize $\top$
- EdgeWorklist $\leftarrow$ edge to entry block
- Get item from any work list; stop when empty
- Edge – if not yet visited, mark visited and:
  - Visited block first time: *visit* all instrs (incl. ϕ)
  - Otherwise: *visit* just ϕ-nodes
- Instr – if block is reachable: *visit*
- At end: replace instrs with const lattice values

[Slide 148] SCCP Algorithm II

visit

- ϕ-node: update lattice, meet of visited edge values
- Other instructions: try constant fold
- If lattice value changed:
  - Add all users to instruction worklist
  - Branch with constant/no condition: add selected edge to worklist
  - Branch with $\bot$: add all edges to worklist

[2] MN Wegman and FK Zadeck. "Constant propagation with conditional branches". In: *TOPLAS* 13.2 (1991), pp. 181–210. URL: `https://dl.acm.org/doi/pdf/10.1145/103135.103136`.

[Slide 149]

In-Class Exercise:

```
define i32 @fn() {
entry:
  br label %loop
loop:
  %j2 = phi i32 [ 1, %entry ], [ %j4, %ifmerge ]
  %k2 = phi i32 [ 0, %entry ], [ %k4, %ifmerge ]
  %kcond = icmp slt i32 %k2, 100
  br i1 %kcond, label %loopbody, label %ret
loopbody:
  %jcond = icmp slt i32 %j2, 20
  br i1 %jcond, label %then, label %else
then:
  %k3 = add i32 %k2, 1
  br label %ifmerge
else:
  %k5 = add i32 %k2, 1
  br label %ifmerge
ifmerge:
  %j4 = phi i32 [ 1, %then ], [ %k2, %else ]
  %k4 = phi i32 [ %k3, %then ], [ %k5, %else ]
  br label %loop
ret:
  ret i32 %j2
}
```

Apply SCCP algorithm and fold constants. Then, remove dead instructions and blocks.

In-Class Exercise:

Code attribution: Adapted from `llvm/test/Transforms/SCCP/sccptest.ll` from the LLVM Project, licensed under Apache-2.0 WITH LLVM-Exception.

Solution on page 86.

[Slide 150] SCCP: Misc

- Strictly more powerful than seperate constant folding + DCE
- Runtime: $\mathcal{O}(\#\text{CFG-edges} + \#\text{instr.-operands})$
  - Every CFG edge is processed at most once
  - Every def–use relation is processed at most twice lattice can lower at most twice ($\top \to const, const \to \bot$)
- Can also be used as inter-procedural optimization reaching into module-internal functions

5.4. Simple Common Subexpression Elimination

[Slide 151] Optimization Example 3

```
define i32 @foo(i32 %0, ptr %1, ptr %2) {
  %4 = zext i32 %0 to i64
```

```
  %5 = getelementptr i32, ptr %1, i64 %4
  %6 = load i32, ptr %5, align 4
  %7 = zext i32 %0 to i64
  %8 = getelementptr i32, ptr %2, i64 %7
  %9 = load i32, ptr %8, align 4
  %10 = add nsw i32 %6, %9
  ret i32 %10
}
```

[Slide 152] Common Subexpression Elimination (CSE) – Attempt 1

- Idea: find/eliminate redundant computation of same value
- Keep track of previously seen values in hash map
- Iterate over all instructions
  - If found in map, remove and replace references
  - Otherwise add to map
- Easy, right?

> This algorithm only works for chains of blocks, but not for merging control flow. It is therefore rarely useful, see later.

[Slide 153] CSE Attempt 1 – Example 1

```
          define i32 @foo(i32 %0, ptr %1, ptr %2) {
→ ht       %4 = zext i32 %0 to i64
→ ht       %5 = getelementptr i32, ptr %1, i64 %4
→ ht       %6 = load i32, ptr %5, align 4
dup %4     %7 = zext i32 %0 to i64
→ ht       %8 = getelementptr i32, ptr %2, i64 %7%4
→ ht       %9 = load i32, ptr %8, align 4
→ ht       %10 = add nsw i32 %6, %9
→ ht       ret i32 %10
          }
```

- Obsolete instr. can be killed immediately, or in a later DCE

[Slide 154] CSE Attempt 1 – Example 2

```
              define i32 @square(i32 %a, i32 %b) {
              entry:
→ ht           %cmp = icmp slt i32 %a, %b
→ ht           br i1 %cmp, label %if.then, label %if.end
              if.then: ; preds = %entry
→ ht           %add1 = add i32 %a, %b
→ ht           br label %if.end
              if.end: ; preds = %if.then, %entry
→ ht           %condvar = phi i32 [ %add1, %if.then ], [ %a, %entry ]
dup %add1      %add2 = add i32 %a, %b
→ ht           %res = add i32 %condvar, %add2%add1
→ ht           ret i32 %res
              }
```

```
Instruction does not dominate all uses!    error: input module is broken!
```

5.5. Dominator Tree

[Slide 155] Domination

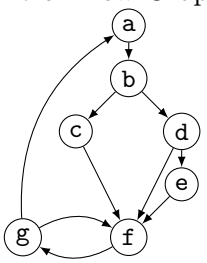
- Remember: CFG $G = (N, E, s)$ with digraph (N, E) and entry $s \in N$
- Dominate: d dom n iff every path from s to n contains d
  - Dominators of n: $DOM(n) = \{d | d \text{ dom } n\}$
- Strictly dominate: d sdom $n \Leftrightarrow d$ dom $n \wedge d \neq n$
- Immediate dominator: idom $(n) = d : d$ sdom $n \wedge \nexists d'.d$ sdom $d' \wedge d'$ sdom n

$\Rightarrow$ All strict dominators are always executed before the block
$\Rightarrow$ All values from dominators available/usable
$\Rightarrow$ All values not from dominators **not** usable

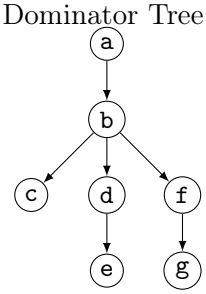
[Slide 156] Dominator Tree

- Tree of immediate dominators
- Allows to iterate over blocks in pre-order/post-order
- Answer a sdom b quickly

Control Flow Graph



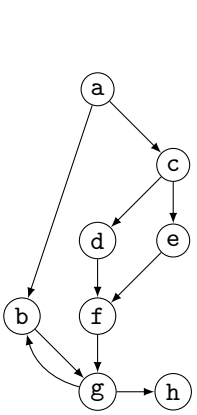
Dominator Tree

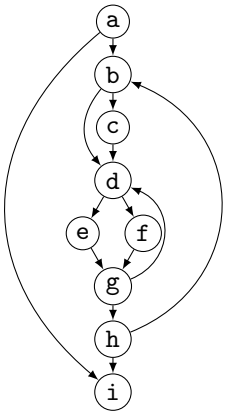


[Slide 157] Dominator Tree – Example

In-Class Exercise:

Construct the dominator tree for the following CFGs (entry at `a`):





Solution on page 87.

[Slide 158] Dominator Tree: Construction

- Naive: inefficient (but reasonably simple)[3]
  - For each block: find a path from the root – superset of dominators
  - Remove last block on path and check for alternative path
  - If no alternative path exists, last block is idom
- Lengauer–Tarjan: more efficient methods[4]
  - Simple method in $\mathcal{O}(m \log n)$; sophisticated method in $\mathcal{O}(m \cdot \alpha(m, n))$ ($\alpha(m, n)$ is the inverse Ackermann function, grows *extremely* slowly)
  - Used in some compilers[5]
- Semi-NCA: $\mathcal{O}(n^2)$, but lower constant factors[6]

> Most notable, LLVM doesn't use the Lengauer–Tarjan algorithm. Instead, they use the Semi-NCA algorithm, which has $\mathcal{O}(n^2)$ runtime, but lower constant factors and is therefore substantially faster for certain (typical) inputs[a].
>
> [a] J Kuderski. "Dominator Trees and incremental updates that transcend times". In: *LLVM Dev Meeting*. Oct. 2017. URL: `https://llvm.org/devmtg/2017-10/slides/Kuderski-Dominator_Trees.pdf`.

[Slide 159] Dominator Tree: Implementation

- Per node store: *idom*, idom-children, DFS pre-order/post-order number
- Get immediate dominator: ...lookup *idom*
- Iterate over all dominators/dominated by: ...trivial

[3] ES Lowry and CW Medlock. "Object code optimization". In: *CACM* 12.1 (1969), pp. 13–22. URL: `https://dl.acm.org/doi/pdf/10.1145/362835.362838`.

[4] T Lengauer and RE Tarjan. "A fast algorithm for finding dominators in a flowgraph". In: *TOPLAS* 1.1 (1979), pp. 121–141. URL: `https://dl.acm.org/doi/pdf/10.1145/357062.357071`

[5] Example: `https://github.com/WebKit/WebKit/blob/aabfacb/Source/WTF/wtf/Dominators.h`

[6] L Georgiadis. "Linear-Time Algorithms for Dominators and Related Problems". PhD thesis. Princeton University, Nov. 2005. URL: `https://www.cs.princeton.edu/techreports/2005/737.pdf`

- Check whether a sdom b[7]
  - $a.preNum < b.preNum \ \wedge \ a.postNum > b.postNum$
  - After updates, numbers might be invalid: recompute or walk tree
- Problem: dominance of unreachable blocks ill-defined $\rightsquigarrow$ special handling

5.6. Common Subexpression Elimination

[Slide 160] CSE Attempt 2

- Option 1:
  - For identical instructions, store all
  - Add dominance check before replacing
  - Visit nodes in reverse post-order (i.e., topological order)
- Option 2 (Global Value Numbering):[8]
  - Do a DFS over dominator tree
  - Use scoped hashmap to track available values

Does this work? Yes.

[Slide 161] CSE: Hashing an Instruction (and Beyond)

- Needs hash function *and* "relaxed" equality
- Idea: combine opcode and operands/constants into hash value
  - E.g. assign number to values (hence "value numbering")
  - Alternatively: use pointer or index for instruction result operands
- Canonicalize commutative operations
  - Order operands deterministically, e.g., by number/address
- Identities: `a+(b+c)` vs. `(a+b)+c`

[Slide 162] More Complex Global Value Numbering

- Hash-based approach only catches trivially removable duplicates
- Alternative: partition values into *congruence classes*
  - Congruent values are guaranteed to always have the same value
- Optimistic approach: values are congruent unless proven otherwise
- Pessimistic approach: values are not congruent unless proven
- Combinable with: reassociation, DCE, constant folding
- Rather complex, but can be highly beneficial[9]

[7] PF Dietz. "Maintaining order in a linked list". In: *STOC*. 1982, pp. 122–127. URL: `https://dl.acm.org/doi/pdf/10.1145/800070.802184`.

[8] P Briggs, KD Cooper, and LT Simpson. *Value numbering*. Tech. rep. CRPC-TR94517-S. Rice University, 1997. URL: `https://www.cs.rice.edu/~keith/Promo/CRPC-TR94517.pdf.gz`.

[9] K Gargi. "A sparse algorithm for predicated global value numbering". In: *PLDI*. 2002, pp. 45–56.

5.7. Inlining

[Slide 163] Inlining

- Inlining: copy function body in place of the call
  - Historically also referred to as "procedure integration"
- Benefits:
  - More optimization opportunities: constant propagation, dead code elimination, better register allocation, etc.
  - Avoids call overhead: stack frame setup, register saving, etc.
- Inlining is crucial for performance in many programming languages
  - "Zero-cost" abstractions typically rely on small functions
  - Critical in e.g. C++/Rust; not too important for C

[Slide 164] Inlining: Decision

- First: determine whether function is inlinable

> Not all functions are inlinable. For example, in LLVM, functions with computed `goto` cannot be inlined, because basic blocks whose address was taken cannot be duplicated. Another example functions compiled from different languages with different exception handling mechanisms (the exception handling routine needs language-specific metadata, which in general cannot be merged; see later).

- Easy way: delegate to programmer (`__attribute__((always_inline))`)

> This attribute is, in fact, considered as somewhat problematic, because it is not just a very strong inlining hint, but guarantees inlining even with optimizations disabled. Therefore, even at `-O0`, compilers must execute an optimization pass to specifically handle functions annotated with this attribute.

- Otherwise: very difficult problem
- Problem 1: ordering
  - Inlining big `a()` into `b()` might make `b()` too big to inline but inlining `b()` into `c()` might provide more opportunities
- Problem 2: cost model and thresholds
  - Estimate cost of inlining, e.g. additional code size, more branches
  - Estimate optimization gains and hotness
  - Some attempts to use machine learning here

> Typical heuristics are based on somewhat arbitrary thresholds, which are typically higher for higher optimization levels. There're also approaches to use machine learning

to guide inlining decisions[a].

[a] M Trofin et al. *MLGO: a Machine Learning Guided Compiler Optimizations Framework*. 2021. arXiv: 2101.04808 [cs.PL]. URL: https://arxiv.org/abs/2101.04808.

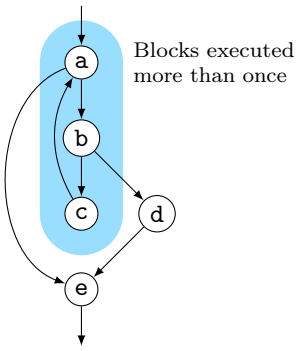
[Slide 165] Inlining Transformation

- Copy original function in place of the call
  - Split basic block containing function call
  - Needs map of old to new value
  - Constant fold/simplify instructions along the way
  - Replace constant conditional branches ⇝ fewer copying
- Replace returns with branches and ϕ-node to/at continuation point
- Move static `alloca` to beginning
- Dynamic `alloca`: save/restore stack pointer
  - Prevent unbounded stack growth in loops
  - LLVM provides `stacksave`/`stackrestore` intrinsics
- Exceptions may need special treatment

5.8. Loop Analysis

[Slide 166] What is a Loop?

```
void func() {
  while (a()) {
    if (b()) {
      d();
      break;
    }
    c();
  }
  e();
}
```



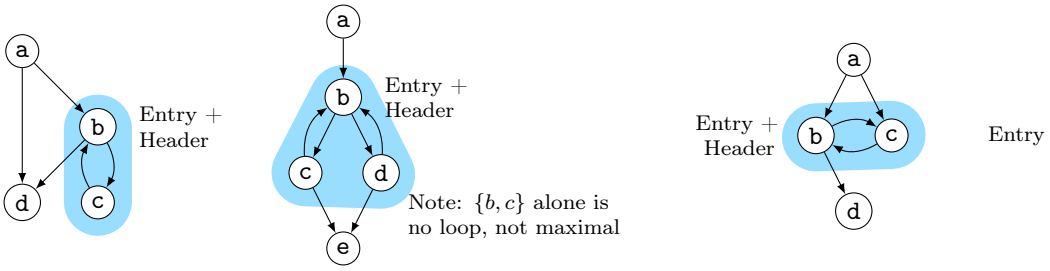
- Loops in source code ≠ loops in CFG
- `d` is *not* part of loop: executed at most once

⇝ Need algorithm to find loops in CFG

[Slide 167] Loops

- Loop: maximal SCC L with at least one internal edge[10] (strongly connected component (SCC): all blocks reachable from each other)
  - Entry: block with an edge from outside of L
  - Header h: first entry found (might be ambiguous)
- Loop nested in L: loop in subgraph $L \setminus \{h\}$

[10] P Havlak. "Nesting of reducible and irreducible loops". In: *TOPLAS* 19.4 (1997), pp. 557–567. URL: https://dl.acm.org/doi/pdf/10.1145/262004.262005.



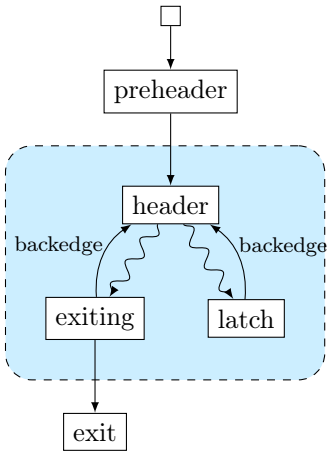
[Slide 168] Natural Loops

- Natural Loop: loop with single entry
  - $\Rightarrow$ Header is unique
  - $\Rightarrow$ Header dominates all block
  - $\Rightarrow$ Loop is reducible
- Backedge: edge from block to header
- Predecessor: block with edge into loop
- Preheader: unique predecessor

Formal Definition

Loop L is reducible iff $\exists h \in L \,.\, \forall n \in L \,.\, h$ dom n

CFG is reducible iff all loops are reducible



[Slide 169] Finding Natural Loops

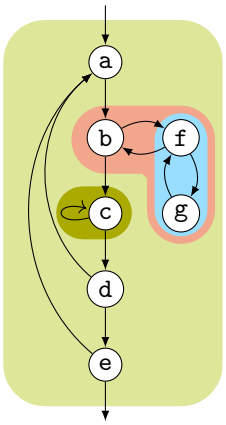
- Modified version[11] of Tarjan's algorithm[12]
- Iterate over dominator tree in post order
- Each block: find predecessors dominated by the block
  - None $\rightsquigarrow$ no loop header, continue
  - Any $\rightsquigarrow$ loop header, these edges *must* be backedges
- Walk through predecessors until reaching header again
  - All blocks on the way must be part of the loop body
  - Might encounter nested loops, update loop parent

[11] G Ramalingam. "Identifying loops in almost linear time". In: *TOPLAS* 21.2 (1999), pp. 175–188. URL: `https://dl.acm.org/doi/pdf/10.1145/316686.316687`.

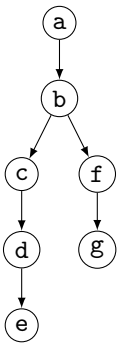
[12] R Tarjan. "Testing flow graph reducibility". In: *STOC*. 1973, pp. 96–107. URL: `https://dl.acm.org/doi/pdf/10.1145/800125.804040`.

[Slide 170] Finding Natural Loops: Example

Control Flow Graph



Dominator Tree



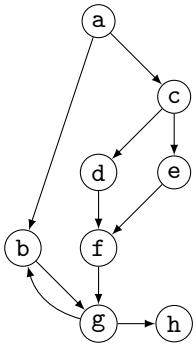
Loop Info

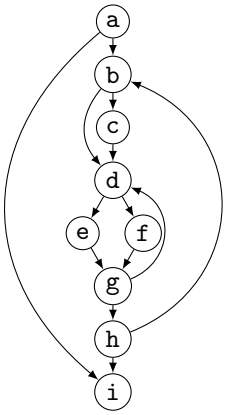
Loop A: {c}
header: c; parent: D
Loop B: {f,g}
header: f; parent: C
Loop C: {b,f,g}
header: b; parent: D
Loop D: {a,b,c,d,e,f,g}
header: a; parent: NULL

[Slide 171] Loop Analysis – Example

In-Class Exercise:

Apply the previous algorithm to find loops in the following CFGs (entry at a):





Solution on page 87.

5.9. Loop Optimizations

[Slide 172] Loop Invariant Code Motion (LICM)

- Analyze loops, iterate over loop tree in post-order
  - I.e., visit inner loops first

↑ Hoist:[13] iterate over blocks of loop in reverse post-order

[13] https://github.com/bytecodealliance/wasmtime/blob/bd6fe11/cranelift/codegen/src/licm.rs

  - For each movable inst., check for loop-defined operands
  - If not, move to preheader (create one, if not existent)
  - Otherwise, add inst. to set of values defined inside loop

↓ Sink: Iterate over blocks of loop in post-order

  - For each movable inst., check for users inside loop
  - If none, move to unique exit (if existent)

[Slide 173] Loop Unswitch

- Hoist loop-invariant conditionals outside loop

> After loop-invariant code is moved outside of the loop, it is trivial to determine whether a condition is loop-invariant.

```
while (...)
  if (loopInvariantCond)
    A
  else
    B
  C
```

→

```
if (loopInvariantCond)
  while (...)
    A
    C
else
  while (...)
    B
    C
```

- Might duplicate parts of the loop ⇝ needs heuristics
- Might lead to exponential code size increase

> For n independently unswitchable conditions, fully unswitching the loop leads to 2^n loops.

[Slide 174] Loop Rotation

- Rotate loop condition to bottom
  - Essentially, convert `while{}` to `if{do{}while}`

> This reduces the number of branches per loop trip from 2 to 1. Additionally, if the loop is known to have a trip count of at least 1, the initial condition can be eliminated.

```
define void @src() {
  ; ...
loopheader:
  %iv = phi ...
  %cond = ...
  br i1 %cond, %body, %exit
body:
  ; ...
  br %loopheader
exit: ; ...
}
```

→

```
define void @tgt() {
  ; ...
  %cond1 = ...
  br i1 %cond1, %body, %exit
body:
  %iv = phi ...
  ; ...
  %cond2 = ...
  br i1 %cond2, %body, %exit
exit: ; ...
}
```

[Slide 175] Loop Strength Reduction

- Loop induction variables (e.g., loop counters) are often multiplied
  - E.g., for array indexing
- Strength reduction: replace "strong" multiply with "weak" addition
- Needs: analysis of loop induction variables and their recurrence
- Rewrite through replaced/extra loop induction variable
  - E.g., replace scaled index operation with pointer addition
- Information about ISA addressing modes beneficial

[Slide 176] Analyses and Transformations – Summary

- Program Transformation critical for performance improvement
- Code not necessarily better
- Analyses are important to drive transformations
  - Dominator tree, loop detection, value liveness
- Important optimizations
  - Dead code elimination, sparse conditional constant propagation common sub-expression elimination, loop-invariant code motion, loop unswitching, loop rotation, loop strength reduction

[Slide 177] Analyses and Transformations – Questions

- Why is "optimization" a misleading name for a transformation?
- How to find unused code sections in a function's CFG?
- Why is a liveness-based DCE better than a simple, user-based DCE?
- Why is SCCP more powerful than separate DCE/constant prop.?
- What is a dominator tree useful for?
- What is the difference between an irreducible and a natural loop?
- How to find natural loops in a CFG?
- How does the algorithm handle irreducible loops?
- Why is sinking a loop-invariant inst. harder than hoisting?

6. Optimizations in LLVM

6.1. Overview

[Slide 179] Optimizations in LLVM

- Analysis: collect information about IR
  - Examples: loop info, alias analysis, branch probabilities
  - Results are cached, re-run when outdated
- Pass: perform transformation on module/CGSCC/function/loop
  - Examples: instruction combine, simplify CFG, global value numbering, scalar replacement of aggregates (SROA; promotes `alloca` to SSA)
  - Input and output IR must be valid
  - Can use results of analyses; invalidates (some) analyses on changes
- Pass Manager: execute series of passes of same granularity
  - Otherwise, use adaptor: `createFunctionToLoopPassAdaptor`

[Slide 180] LLVM Optimization Pipeline (Simplified)[1]

Module Pipeline (`buildPerModuleDefaultPipeline`): two parts
Module Simplification Pipeline (`buildModuleSimplificationPipeline`)

- Early function simplification (SimplifyCFG, SROA, EarlyCSE)
- CGSCC (post-order) (repeated after devirtualization):
  - Inliner
  - Function Simplification Pipeline (`buildFunctionSimplificationPipeline`)
    - SROA, SimplifyCFG, InstCombine, Reassociate, GVN, DCE, LICM
    - Most passes are here

> Inside a call graph strongly-connected component (CGSCC), functions are simplified in post-order (i.e., leaf functions first). The inliner therefore makes decisions based on the simplified callee.
> Other acronyms: SROA is Scalable Replacement Of Aggregates; GVN is Global Value Numbering (elimination of common expressions); DCE is Dead Code Elimination; LICM is Loop Invariant Code Motion (hoisting instructions that compute the same value inside the loop out of the loop).

Module Optimization Pipeline (`buildModuleOptimizationPipeline`)

[1] See `llvm/lib/Passes/PassBuilderPipelines.cpp` for full details.

- Primarily hard-to-reverse loop optimizations that are not simplifications
- E.g. vectorization, loop unrolling, loop distribution

6.2. Canonicalization

[Slide 181] SSA Construction: Mem2Reg and SROA

- Front-ends typically emit `alloca`s for variables
  - Easier; also simplifies debugging and debug information

    > Variables stored in `alloca`s have a single location, which greatly simplifies debug information, as the variable does not need to be tracked through registers. Additionally, as the variable is loaded on every use, modifications of the variable in the debugger are possible and in most cases immediately propagated.

- Mem2reg: promote `alloca` to SSA values/phis
  - Condition: only `load`/`store`, no address taken
  - Essentially just SSA construction

    > In code, this pass is called `PromotePass`.

- SROA: scalar replacement of aggregate
  - Separate structure fields into separate variables
  - Also promote them to SSA, subsumes mem2reg

> SROA is run multiple times in the pipeline, as further simplifications might enable the promotion of more variables (e.g., when escapes of the variable are eliminated).

[Slide 182] Simplification

- Many operations can be expressed in various ways
  - E.g.: `sub i32 %x, 1`, `add i32 %x, -1`, `add i32 -1, %x`
- Problem: always need to consider all equivalent cases

⇒ Canonicalize IR: one single preferred form
  - Single instructions: e.g. operand order
  - Instruction sequences: e.g. series of additions, useless PHIs
  - CFG: e.g. branches, loop layout
- E.g.: constants go to the right, constant `sub` is `add`, loops with fixed trip count turned into `for (i = 0; i != 25; ++i)`, ...

> Canonicalization is one of LLVM's core approaches to program optimizations. In addition to simplifying the implementation of pattern matches (passes can assume the IR to be canonical form, although they must work correctly even on non-canonical IR), the canonicalizations also simplify and thereby optimize the program.

> In contrast to other optimizations, canonicalizations are not based on heuristics and should always work in the same direction (i.e., one pass should not undo the transformation of a previous pass). Inlining is a notable exception of a heuristics-based pass run during the simplification pipeline.
>
> Note that some canonicalization done on the IR are undone in the back-end. Examples include tail duplication (duplicating the function returns) and moving certain instructions back into loops (e.g., to reduce register pressure).
>
> Unfortunately, there is no documentation on the canonical form of IR; it is defined implicitly by what the passes (currently) produce.

[Slide 183] InstCombine

- Main instruction canonicalization and simplification pass
- Includes many arithmetic-logical patterns, folding of PHIs, ...
- Implementation:
  - Add instructions to work list
    - Trivially dead/constant instructions eliminated immediately
    - Blocks traversed in reverse post-order
  - Patterns applied to instructions; on match, replace all uses
  - Changed instruction re-enqueues all users to work list

  > The block ordering in and adding instructions in the order as they appear inside the basic block causes definitions to be visited before their uses (apart from ϕ-nodes). Therefore, adding users to the work list frequently has no effect, as it already contains all users.

- Historically: fixed-point iteration, now just one iter. per run

  > It turned out that most combines are performed in the first run. Very often, the second iteration would not change the IR but would still need to inspect it, causing easily avoidable compile time costs.
  > As the pass is executed multiple times throughout the default pass pipeline, combines missed in the first iteration are then simply performed in a later run of the pass.

- Covers many patterns, many patterns missing, >50kLOC C++

[Slide 184] IR Pattern Matching

- IR patterns are typically DAG matches on def–use graph
- Typically implemented via helper functions[2]

```
// Excerpt from InstCombinerImpl::visitAdd

// (add (xor A, B) (and A, B)) --> (or A, B)
```

[2] llvm/IR/PatternMatch.h

```
// (add (and A, B) (xor A, B)) --> (or A, B)
if (match(&I, m_c_BinOp(m_Xor(m_Value(A), m_Value(B)),
                        m_c_And(m_Deferred(A), m_Deferred(B)))))
  return BinaryOperator::CreateOr(A, B);
```

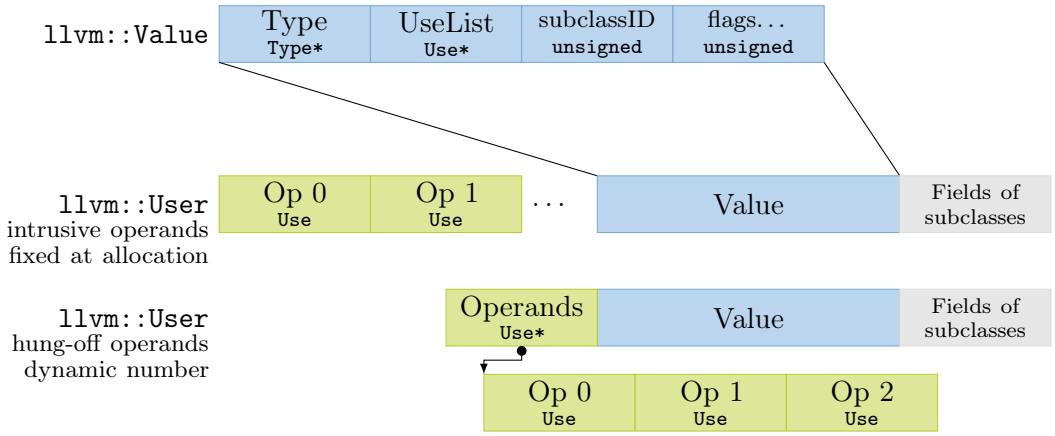
[Slide 185] LLVM: Use Tracking

- Values track their users

```
llvm::Value* v = /* ... */;
for (llvm::User* u : v->users())
  if (auto i = llvm::dyn_cast<llvm::Instruction>(u))
    // ...
```

- Allows for easy replacement:
  - `inst->replaceAllUsesWith(replVal);`

[Slide 186] LLVM IR Implementation: Value/User



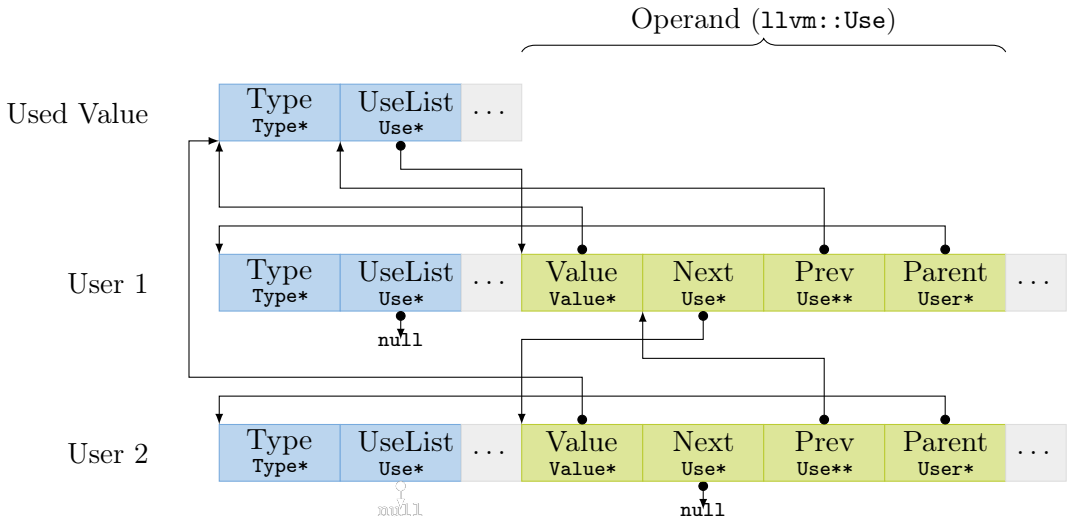
`PHINode` additionally stores n `BasicBlock*` after the operands, but aren't users of blocks.

Every LLVM `Instruction` is a separate heap allocation. As the number of operands is typically known when constructing the instruction, they are allocated *before* the instruction data structure (this is implemented by `User`).

It can happen that the number of operands increases beyond the allocated storage, for example, when a `PHINode` gets more operands than initially expected. In such cases, the operand list gets hung off into a separate allocation.

As a special case, `PHINode` needs to store the associated `BasicBlocks` in addition to the merged values. The blocks are stored after the operands, but are not operands themselves. Therefore, `PHINode` operands are always hung off.

[Slide 187] LLVM IR Implementation: Use



The use list is a doubly-linked list. Starting from `Value::UseList`, one can find all used by following the `Use::Next` pointer. The `Use::Prev` pointer does not point to the previous `Use`, but the previous `Use::Next` pointer or the `Value::UseList` — this way, unlinking does not need to distinguish the special case of the beginning of the use list.

A `Use` also has a pointer to the actual `Value`, so that when inspecting an operand one can actually find the operand itself.

There is also a `Parent` pointer, which points to the `User` which owns the operand: when iterating over the use list, this is the only way to find out which instruction (`User`) uses the value.

In sum, an LLVM-IR operand is quite large, using 32 bytes on a 64-bit system. In addition to every instruction being a separate heap allocation and every operand update requires updating the use list (less data locality), the IR data structures are (in absolute terms) not very efficient — despite being fairly optimized for the use cases they serve.

[Slide 188] LLVM IR Implementation: Instructions/Blocks

- `Instruction` and `BasicBlock` have pointers to parent and next/prev
  - Linked list updated on changes and used for iteration
  - Instructions have cached *order* (integer) for fast "comes before"
- `BasicBlock` successors: blocks used by terminator
- `BasicBlock` predecessors:
  - Iterate over users of block – these are terminators (and `blockaddress`)
  - Ignore non-terminators, parent of using terminator is predecessor
  - Same predecessor might be duplicated ($\rightsquigarrow$ `getUniquePredecessor()`)
- Finding first non-ϕ requires iterating over ϕ-nodes

[Slide 189] Valid Transforms?

In-Class Exercise:

Which of these transformations `src`→`tgt` are valid?

```
define float @src1(float %x) {
  %a = fadd float %x, 0.0
  ret float %a
}
define i1 @src2(i8 %x) {
  %xm1 = add i8 %x, -1
  %y = xor i8 %x, %xm1
  %r = icmp ule i8 %x, %y
  ret i1 %r
}
define i8 @src3(i8 %x) {
  %div = sdiv i8 %x, 4
  ret i8 %div
}
```

```
define float @tgt1(float %x) {
  ret float %x
}

define i1 @tgt2(i8 %x) {
  %ctpop = call i8 @llvm.ctpop(i8 %x)
  %r = icmp ult i8 %ctpop, 2
  ret i1 %r
}

define i8 @tgt3(i8 %x) {
  %div = ashr i8 %x, 2
  ret i8 %div
}
```

Solution on page 87.

[Slide 190] Alive2

- Seemingly correct transformations might be wrong in edge cases
- Unexpected transformations can be correct

⇒ Prove correctness of transformations

- Alive2[3]: translation validation tool for LLVM-IR
- Convert LLVM-IR into SMT formulas
- Proves that target is a refinement of source function

> Alive2 transforms LLVM-IR functions, including memory accesses and control flow, into SMT formulas and uses the Z3 SMT solver to prove that the target function is indeed a refinement of the source function, i.e., that the target is more defined than the source. There is limited support for `undef` values (which are not tracked per-bit but just per-value) and loops (which are completely unrolled for a fixed number of iterations).

- InstCombine transformations require Alive2 proofs nowadays[4]

[Slide 191] Value Tracking

- Many transformations benefit from value ranges and known bits
  - E.g., fold `trunc`+shift into smaller shift if shift amount is known small
- `computeKnownBits`: recursively collect information from instrs
  - Arithmetic-logical instrs, casts, several intrinsics, etc.

[3] NP Lopes et al. "Alive2: bounded translation validation for LLVM". in: *PLDI*. 2021, pp. 65–79. URL: `users.cs.utah.edu/~regehr/alive2-pldi21.pdf`.

[4] Random Example: `https://alive2.llvm.org/ce/z/xe_vb2`

- Bitwise instructions easier to reason about ⇝ preferred canonicalization
  - E.g. `(1 << X) - 1` → `(-1 << X) ^ -1`
  - E.g. `sext(X)` → `zext(X)` if `X` is known non-negative
- Also use knowledge from front-end (but how encoded?)

6.3. Flags

[Slide 192] Attribute, Flags, Assume

- Paramter/return attributes: constrain value ranges
  - `range(from, to)`, `nonnull`, `noundef`, ...
  - Values outside of the range are `poison`
- Instruction flags: encode additional knowledge about operation
  - `add`/`sub`/`mul`/`trunc` `nsw`/`nuw`: no signed/unsigned wrap

    > This can be used, e.g., to restrict the value range or get more precise information about loop trip counts, memory access offsets.

  - `or` `disjoint`: combination of disjoint bits

    > This permits back-ends to replace the `or` with an `add`, which can be folded into other operations (e.g., `lea` on x86).

  - `udiv`/`sdiv`/`lshr`/`ashr` `exact`: divisor multiple of dividend

    > Permit replacing division with shift and implies that the bits that are shifted out are zero.

  - `getelementptr` `nuw`/`nusw`/`inbounds`: no wrap/stays inside allocation

    > Bounds information primarily improves alias analysis: if two base pointers are known to refer to different allocations, they will never be the same after two `inbounds` computations.

  - `zext`/`uitofp` `nneg`: operand is non-negative

    > When a value is known to be non-negative, the instruction is converted into the canonical unsigned form. However, some targets might want to undo this transformation in the back-end, e.g., because sign-extension is cheaper (RISC-V) or the floating-point conversion is cheaper for signed integers (x86), but it might be difficult to re-prove the non-negativeness after other transformations. This flag retains this information and permits relaxing the operations later on.

  - `icmp` `samesign`: operands have the same sign

> When both operands are known to have the same sign, `icmp` is canonicalized into the unsigned comparison predicate. However, when there are multiple comparisons of the same value with some being signed while others are unsigned, it is more difficult to the aggregate information. This flag retains this information and indicates that the signed and unsigned variants of the comparison predicate can be used interchangeably.

- Violation is `poison`

[Slide 193] Assume

- Intrinsic `llvm.assume(i1)`
- Conditional immediate undefined behavior if argument is `false`
- Requires extra instruction for condition ⇝ might be problematic
  - Instructions use other values, etc. ⇝ prevent optimizations
- Also supports operand bundles for `nonnull`, `align`, `separate_storage`

Here's an example which demonstrate the use of `separate_storage` assumptions:

```
struct Buffer {
  char *data;
  void write(char c) {
    *data++ = c;
  }
};
void writeLE(unsigned x, Buffer &buf) {
  buf.write(x);
  buf.write(x >> 8);
  buf.write(x >> 16);
  buf.write(x >> 24);
}
```

The resulting machine code will repeatedly load the `data` pointer from `buf`, because `data` might alias with `buf`, because in C/C++ `char*` can alias with everything. When exposing the information that inside `write()` the pointers `this` and `this->data` do not alias, the four stores can get combined.

```
// Inside write:
__builtin_assume_separate_storage(this, data);
// LLVM IR: call void @llvm.assume(i1 true) [ "separate_storage"(ptr %this, ptr %data) ]
```

Full example: `https://godbolt.org/z/TMGrTco9T`

[Slide 194] Valid Transforms?

In-Class Exercise:

Which of these transformations `src`→`tgt` are valid?

```
define i8 @src1(i8 %x) {
  %div = udiv exact i8 1, %x
  ret i8 %div
}
define i1 @src2(i8 %x, i8 %y) {
  %cmp = icmp sge i8 %x, %y
  %cmpeq = icmp samesign eq i8 %x, 127
  %r = select i1 %cmp, i1 %cmpeq, i1 false
  ret i1 %r
}
define i8 @src3(i8 %x, i8 %y, i8 %z) {
  %xy = add nsw i8 %x, %y
  %xyz = add nsw i8 %xy, %z
  ret i8 %xyz
}
```

```
define i8 @tgt1(i8 %x) {
  ret i8 1
}

define i1 @tgt2(i8 %x, i8 %y) {
  %cmpeq = icmp samesign eq i8 %x, 127
  ret i1 %cmpeq
}

define i8 @tgt3(i8 %x, i8 %y, i8 %z) {
  %xz = add nsw i8 %x, %z
  %xyz = add nsw i8 %xz, %y
  ret i8 %xyz
}
```

Solution on page 87.

[Slide 195] Fast-Math Flags

- Floating-point arithmetic typically follows IEEE 754
  - Except for NaN, which is more relaxed

    > The behavior of floating-point instructions w.r.t. NaN varies strongly between different targets (e.g., taking one of the operands, returning a canonical NaN, converting signalling NaN to quiet NaN). Therefore, the semantics are specified rather generic for LLVM-IR.
- Implication: almost impossible to optimize
  - non-associative, NaNs, infinity, negative zero, no reciprocals, no contraction
  - Transformations can result in vastly different results!
- Additionally: floating-point exceptions
  - Typically ignored, but some users might be interested
- Value/operation flags to permit transformations locally
- `strictfp` function attribute allows more control through constrained floating-point intrinsics

[Slide 196] Branch Weights

- Annotation of branch with expected probabilities
  - Sources: programmer annotations, execution profiles
- Added as metadata to branch instructions
- Can compute block execution frequencies

```
define i32 @f(i32 %x) {
  ; ...
  %cmp = icmp eq i32 %x, 0
  br i1 %cmp, label %then, label %else, !prof !5
  ; ...
}

!5 = !{!"branch_weights", !"expected", i32 1, i32 2000}
```

6.4. Inter-Procedural Optimizations

[Slide 197] Function Argument Optimizations

- Inferring argument attributes
- Eliminate dead arguments
  - Optimistic liveness analysis of arguments (assume dead)
  - Create new function with fewer arguments
- Promote pointer arguments to value arguments
  - Esp. relevant for C++ pass-by-reference
  - Only possible when both caller/callee CPU features match

[Slide 198] Other Inter-Procedural Optimizations

- Inlining / Outlining
- Function Specialization
- Optimization of global variables
  - Marking as constant, delete dead variables, attributes, etc.
- Inter-procedural SCCP
- Devirtualization / call-target speculation
- Hot-Cold Splitting
- ...

[Slide 199] Optimization Remarks

- Understanding optimization decisions can be difficult
  - Complex code base, abstraction, heuristics, ...
- Optimization remarks: passes can report information
  - Inlining decisions: provide cost values/thresholds for decisions
  - Others: GVN, LICM, FastISel
- Output also provided as YAML or a binary format for analysis
- Lots of verbose output, hard to identify important points
- Clang: `-Rpass=<pass> -Rpass-analysis=<pass> -Rpass-missed=<pass>`

[Slide 200] Other Aspects not Covered Here

- Analyses: alias analysis, MemorySSA, Scalar Evolution (SCEV), target-specific analyses/heuristics, ...
- Optimizations: various loop transformations (distribute, fuse, flatten, polyhedral optimizations), value propagation, vectorization, reassociation, ...
- Loop-Closed SSA canonical form
  - Values defined inside loop only used in loop or in ϕ-node in exit
  - Simplifies tracking uses of values defined in loops

6.5. Running and Writing LLVM Passes

[Slide 201] Using LLVM (New) Pass Manager

```
void optimize(llvm::Function* fn) {
  llvm::PassBuilder pb;
  llvm::LoopAnalysisManager lam{};
  llvm::FunctionAnalysisManager fam{};
  llvm::CGSCCAnalysisManager cgam{};
  llvm::ModuleAnalysisManager mam{};
  pb.registerModuleAnalyses(mam);
  pb.registerCGSCCAnalyses(cgam);
  pb.registerFunctionAnalyses(fam);
  pb.registerLoopAnalyses(lam);
  pb.crossRegisterProxies(lam, fam, cgam, mam);

  llvm::FunctionPassManager fpm{};
  fpm.addPass(llvm::DCEPass());
  fpm.addPass(llvm::createFunctionToLoopPassAdaptor(llvm::LoopRotatePass()));
  fpm.run(*fn, fam);
}
```

[Slide 202] Writing a Pass for LLVM's New PM – Part 1

```
#include "llvm/IR/PassManager.h"
#include "llvm/Passes/PassBuilder.h"
#include "llvm/Passes/PassPlugin.h"

class TestPass : public llvm::PassInfoMixin<TestPass> {
public:
  llvm::PreservedAnalyses run(llvm::Function &F,
                              llvm::FunctionAnalysisManager &AM) {
    // Do some magic
    llvm::DominatorTree *DT = &AM.getResult<llvm::DominatorTreeAnalysis>(F);
    // ...
    llvm::errs() << F.getName() << "\n";
    return llvm::PreservedAnalyses::all();
  }
};
// ...
```

[Slide 203] Writing a Pass for LLVM's New PM – Part 2

```
extern "C" ::llvm::PassPluginLibraryInfo LLVM_ATTRIBUTE_WEAK
llvmGetPassPluginInfo() {
  return { LLVM_PLUGIN_API_VERSION, "TestPass", "v1",
    [] (llvm::PassBuilder &PB) {
      PB.registerPipelineParsingCallback(
        [] (llvm::StringRef Name, llvm::FunctionPassManager &FPM,
            llvm::ArrayRef<llvm::PassBuilder::PipelineElement>) {
          if (Name == "testpass") {
            FPM.addPass(TestPass());
            return true;
          }
          return false;
```

```
      });
   } };
}
  c++ -shared -o testpass.so testpass.cc -lLLVM -fPIC
  opt -S -load-pass-plugin=$PWD/testpass.so -passes=testpass input.ll
```

[Slide 204] Optimizations in LLVM – Summary

- LLVM provides a wide range of analyses and optimizations
- Optimizations organized in flexibly composeable passes
- Simplification centered around canonicalization
- Much knowledge inferred from assumptions (and `poison`/UB)
- SSA simplifies many transforms through explicit def–use chains
- Only few non-simplifying transformation passes

[Slide 205] Optimizations in LLVM – Questions

- Why are some passes executed multiple times during optimization?
- How does the simplification pipeline relate to inlining?
- What is the key benefit of canonicalizing the IR?
- How does LLVM's `replaceAllUsesWith` work?
- What is the benefit of instr. flags like `nsw`?
- Why are floating-point optimizations problematic?
- How is C++ `[[likely]]` represented in LLVM-IR?

A. Exercise Solutions

[Slide 22]

Some suggestions:

- Sequences of `+`/`-`/`<`/`>`
- `[-]` → set zero
- `[>]` → find next zero (`memchr`)
- `[->+>+<<]` → add to next two siblings, set zero
- `[->+++<]` → add 3 times to next sibling, set zero
- ...

[Slide 28]

Indirect/direct threading and computed `goto`/tail calls are orthogonal.

- Computed `goto` is a non-standard feature. A lot of state can be kept in registers across different bytecode operations, this is done automatically by the compiler. The possibility of indirect jumps inside the function may prevent some compiler optimizations.
- Tail calls (largely) rely on standard features and the code is typically more maintainable (e.g., operations can be implemented in separate functions). Some state can be kept in registers across operations, but this has to be implemented manually; the usable number of preservable registers is defined by the calling convention and is architecture-dependent. Non-standard calling convention (e.g., ghccc) can be used to increase this limit. Compilers cannot perform optimization across the different functions.
- Indirect threading permits a small opcode (e.g., one byte as in this example), which leads to higher code density and therefore increased data locality.
- Direct threading avoids the indirection to the opcode handler at the cost of larger bytecode (in the example, an operation grows from 2 to 16 bytes). On modern out-of-order CPUs, avoiding the indirection might not lead to significant performance differences.

[Slide 33]

As a simple approach, check the array boundaries at every pointer move, when reaching bounds, re-allocate to larger buffer and copy the elements. As pointer moves occur often, the bounds checks are rather expensive.

To improve efficiency, it is possible to use virtual memory by padding the buffer with known-unmapped pages (guard pages) and intercepting accesses to these pages (e.g., through a signal handler). On an access, the buffer can grow and/or be moved (e.g.,

through `mremap`). Care must be taken to ensure that the guard pages are large enough (the length of the program should be sufficient).

[Slide 52]

Example for input: `a = 3 * 2 + 1;`

	Rec. Depth 1	Rec. Depth 2	Rec. Depth 3

`minPrec`
`lhs`
`op` (prec/assoc)
	1
	`a`
	= (2/r)		

`minPrec`
`lhs`
`op` (prec/assoc)
	1
	`a`
	= (2/r)
		2
		`3`
		* (12/l)	

`minPrec`
`lhs`
`op` (prec/assoc)
	1
	`a`
	= (2/r)
		2
		`3`
		* (12/l)
			13
			`2`
			+ (11/l)

`minPrec`
`lhs`
`op` (prec/assoc)
	1
	`a`
	= (2/r)
		2
		`3*2`
		+ (11/l)	

`minPrec`
`lhs`
`op` (prec/assoc)
	1
	`a`
	= (2/r)
		2
		`3*2`
		+ (11/l)
			12
			1
			; (0/–)

`minPrec`
`lhs`
`op` (prec/assoc)
	1
	`a`
	= (2/r)
		2
		`(3*2)+1`
		; (0/–)	

`minPrec`
`lhs`
`op` (prec/assoc)
	1
	`a=((3*2)+1)`
	; (0/–)		

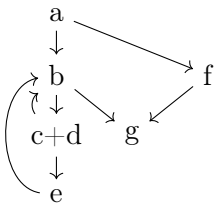
[Slide 60]

There are two ways of implementing a scoped hash table:

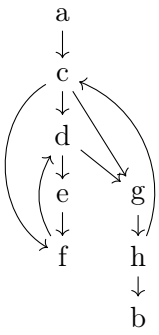
- Chain of hash maps: $Scope = (Map[Name \to Type]\ names, Scope\ parent)$. This is, however, very slow for deeply nested scopes, as all hash maps of the parent scopes must be queried. Hash map lookups are fairly expensive.
- Hash map of lists: $Map[Name \to List[Tuple[Depth, Type]]]$. For every identifier, the type at a given scope nesting depth is stored. Invalidation can be implemented with an epoch counter for every depth. The downside is that this hash map can grow very large, as entries are never removed.

[Slide 83]

Control flow graph for `fn1`:



Control flow graph for `fn2`:



[Slide 99]

```
phis(v1, v2) {
entry:
    v3 = mul v1, v2 // a * b
    v4 = mul v2, v2 // b * b
    v5 = cmpgt v1, v4 // a > b * b
    br v5, ifthen, ifelse
ifthen:
    br header
header:
    phi1 = phi(ifthen: v1, body: v7) // a
    phi2 = phi(ifthen: 1, body: phi2) // c
    v6 = cmpgt phi1, 0 // a > 0
    br v6, body, cont
body:
    v7 = sub phi1, phi2 // a - c
    br header
cont:
    br ret
ifelse:
    v8 = mul v2, v2 // b * b
    br ret
ret:
    phi3 = phi(cont: phi1, ifelse: v8) // a
    ret phi3
}

swap(v1, v2, v3) {
entry:
    br header
header:
    phi1 = phi(entry: v3, body: v5) // c
    // Note: all phi nodes execute concurrently. Therefore, these two phi nodes swap a and b.
    phi2 = phi(entry: v1, body: phi3) // a
    phi3 = phi(entry: v2, body: phi2) // b
    v4 = cmpgt phi1, 0 // a > 0
    br v4, body, cont
body:
```

```
    v5 = sub phi1, 1 // c - 1
    br header
cont:
    ret phi2
}
```

[Slide 128]

```
int sw(int x) {
  switch (x) {
  case 4:
  case 100: return 12;
  case 5: return 32;
  case 8: return 9;
  default: return x;
  }
}
```

[Slide 129]

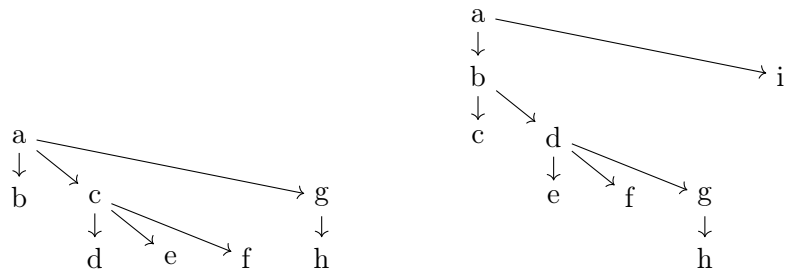
```
int sw(int x) {
  switch (x) {
  case 4:
  case 10: return 12;
  case 5: return 32;
  case 8: return 9;
  default: return x;
  }
}
```

[Slide 149]

The algorithm never visits the `%else` block and therefore can prove that `%j4` and thus `%j2` are always the constant 1.

```
define i32 @fn() {
entry:
  br label %loop
loop:
  %k2 = phi i32 [ 0, %entry ], [ %k3, %ifmerge ]
  %kcond = icmp slt i32 %k2, 100
  br i1 %kcond, label %loopbody, label %ret
loopbody:
  br label %then
then:
  %k3 = add i32 %k2, 1
  br label %ifmerge
ifmerge:
  br label %loop
ret:
  ret i32 1
}
```

[Slide 157]



[Slide 171]

The first CFG has no reducible loop, so the algorithm as written will not find any loops.
The second CFG contains two reducible loops:

- Loop 1: `d` (header), `e`, `f`, `g`
- Loop 2: `b` (header), `c`, `d`, `e`, `f`, `g`, `h`

[Slide 189]

You can use Alive2 as validation tool: `https://alive2.llvm.org/ce/z/_PuqaK`

1. The neutral element of floating-point addition is -0.0; adding $+0.0$ to -0.0 results in $+0.0$, which is not a generally correct transformation.
2. This transformation, although non-obvious, is correct.
3. `sdiv` rounds towards zero, but `ashr`, so for e.g. -3, `src3` returns 0 while `tgt3` return -1.

[Slide 194]

You can use Alive2 as validation tool: `https://alive2.llvm.org/ce/z/HgpRmK`

1. Correct: `udiv exact` results in `poison` when 1 is not a multiple of the parameter, i.e., when the parameter is not 1. Thus, always returning 1 makes the target function more defined than the source function.
2. Not correct: `%cmpeq` is `poison` if `%x` is negative. The transformation would be correct if the `samesign` flag was dropped.
3. Not correct: integer addition is associative, but the `nsw` flag needs to be dropped.

Bibliography

[AC71] FE Allen and J Cocke. *A catalogue of optimizing transformations*. 1971. URL: https://www.clear.rice.edu/comp512/Lectures/Papers/1971-allen-catalog.pdf.

[BCS97] P Briggs, KD Cooper, and LT Simpson. *Value numbering*. Tech. rep. CRPC-TR94517-S. Rice University, 1997. URL: https://www.cs.rice.edu/~keith/Promo/CRPC-TR94517.pdf.gz.

[Bel73] JR Bell. "Threaded Code". In: *CACM* 16.6 (1973), pp. 370–372. URL: https://dl.acm.org/doi/pdf/10.1145/362248.362270.

[Bra+13] M Braun et al. "Simple and efficient construction of static single assignment form". In: *CC*. 2013, pp. 102–122. URL: https://link.springer.com/content/pdf/10.1007/978-3-642-37051-9_6.pdf.

[Cyt+91] R Cytron et al. "Efficiently computing static single assignment form and the control dependence graph". In: *TOPLAS* 13.4 (1991), pp. 451–490. URL: https://dl.acm.org/doi/pdf/10.1145/115372.115320.

[Die82] PF Dietz. "Maintaining order in a linked list". In: *STOC*. 1982, pp. 122–127. URL: https://dl.acm.org/doi/pdf/10.1145/800070.802184.

[EG03] MA Ertl and D Gregg. "The structure and performance of efficient interpreters". In: *JILP* 5 (2003), pp. 1–25. URL: http://www.jilp.org/vol5/v5paper12.pdf.

[Gar02] K Gargi. "A sparse algorithm for predicated global value numbering". In: *PLDI*. 2002, pp. 45–56.

[Geo05] L Georgiadis. "Linear-Time Algorithms for Dominators and Related Problems". PhD thesis. Princeton University, Nov. 2005. URL: https://www.cs.princeton.edu/techreports/2005/737.pdf.

[Hab13] J Haberman. *Parsing C++ is literally undecidable*. 2013. URL: https://blog.reverberate.org/2013/08/parsing-c-is-literally-undecidable.html.

[Hav97] P Havlak. "Nesting of reducible and irreducible loops". In: *TOPLAS* 19.4 (1997), pp. 557–567. URL: https://dl.acm.org/doi/pdf/10.1145/262004.262005.

[Kud17] J Kuderski. "Dominator Trees and incremental updates that transcend times". In: *LLVM Dev Meeting*. Oct. 2017. URL: https://llvm.org/devmtg/2017-10/slides/Kuderski-Dominator_Trees.pdf.

[LA04] C Lattner and V Adve. "LLVM: A compilation framework for lifelong program analysis & transformation". In: *CGO*. 2004, pp. 75–86. URL: http://www.llvm.org/pubs/2004-01-30-CGO-LLVM.pdf.

[LM69] ES Lowry and CW Medlock. "Object code optimization". In: *CACM* 12.1 (1969), pp. 13–22. URL: https://dl.acm.org/doi/pdf/10.1145/362835.362838.

[Lop+21] NP Lopes et al. "Alive2: bounded translation validation for LLVM". In: *PLDI*. 2021, pp. 65–79. URL: users.cs.utah.edu/~regehr/alive2-pldi21.pdf.

[LT79] T Lengauer and RE Tarjan. "A fast algorithm for finding dominators in a flowgraph". In: *TOPLAS* 1.1 (1979), pp. 121–141. URL: https://dl.acm.org/doi/pdf/10.1145/357062.357071.

[Ram99] G Ramalingam. "Identifying loops in almost linear time". In: *TOPLAS* 21.2 (1999), pp. 175–188. URL: https://dl.acm.org/doi/pdf/10.1145/316686.316687.

[Tar73] R Tarjan. "Testing flow graph reducibility". In: *STOC*. 1973, pp. 96–107. URL: https://dl.acm.org/doi/pdf/10.1145/800125.804040.

[Tro+21] M Trofin et al. *MLGO: a Machine Learning Guided Compiler Optimizations Framework*. 2021. arXiv: 2101.04808 [cs.PL]. URL: https://arxiv.org/abs/2101.04808.

[Vel03] TL Veldhuizen. *C++ templates are Turing complete*. 2003. URL: http://port70.net/~nsz/c/c%2B%2B/turing.pdf.

[WZ91] MN Wegman and FK Zadeck. "Constant propagation with conditional branches". In: *TOPLAS* 13.2 (1991), pp. 181–210. URL: https://dl.acm.org/doi/pdf/10.1145/103135.103136.