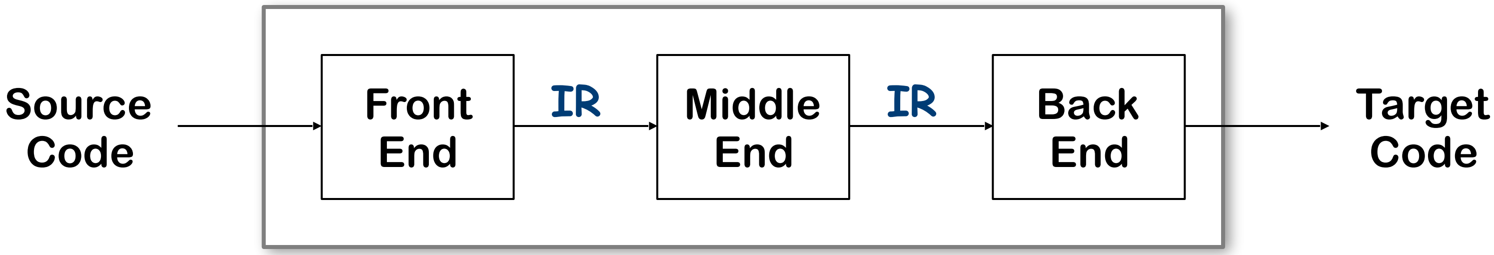
Intermediate Representations

Where are we?

- We are on the cusp of the art, science, & engineering of compilation
- Scanning & parsing are applications of automata theory
- Context-sensitive analysis, as covered in class, is mostly software engineering
- The mid-section of the course will focus on issues where the compiler writer needs to choose among alternatives
  – The choices matter; they affect the quality of compiled code
  – There may be no "best answer" or "best practice"

The fun begins at this point

Intermediate Representations



- Front end - produces an intermediate representation (IR)
- Middle end - transforms the IR into an equivalent IR that runs more efficiently
- Back end - transforms the IR into native code

- IR encodes the compiler's knowledge of the program
- Middle end usually consists of several passes

Intermediate Representations

- Decisions in IR design affect the speed and efficiency of the compiler

- Some important IR properties
  - Ease of generation
  - Ease of manipulation

- The importance of different properties varies between compilers
  - Selecting an appropriate IR for a compiler is critical

- Three axes of choice: structural organization, level of abstraction, naming discipline

Structural organization

Three major categories

- Structural
  - Graphically oriented
  - Heavily used in source-to-source translators
  - Tend to be large

Examples:
Trees, DAGs

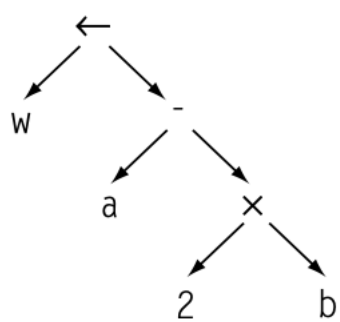
- Linear
  - Pseudo-code for an abstract machine
  - Level of abstraction varies
  - Simple, compact data structures
  - Easier to rearrange

Examples:
3 address code
Stack machine code

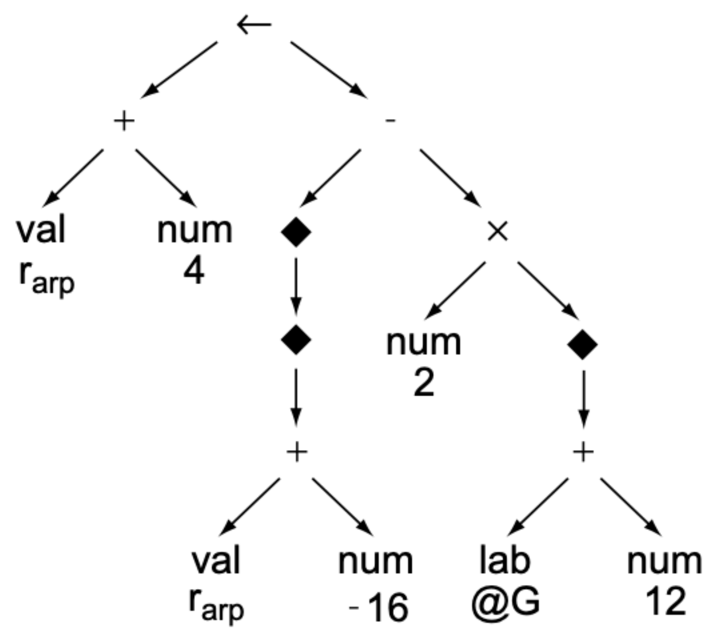
- Hybrid
  - Combination of graphs and linear code

Example:
Control-flow graph

AST with different level of abstraction



(a) Source-Level AST

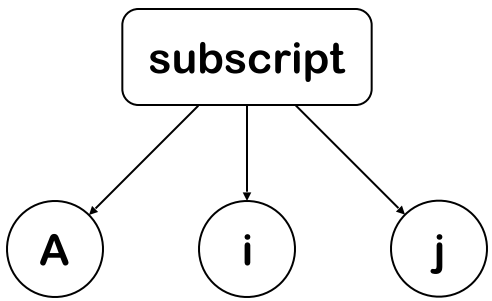


(b) Low-Level AST

Level of Abstraction

- The level of detail exposed in an IR influences the profitability and feasibility of different optimizations.
- Two different representations of an array reference:

array A[1…10,1…10] of 1 byte memorised in row-major order



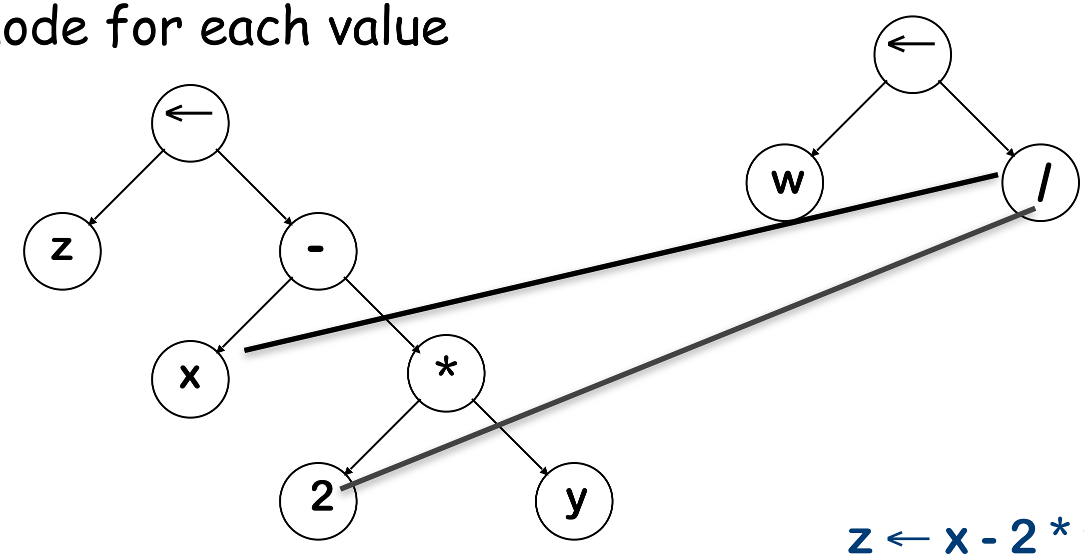
High level AST:
Good for memory disambiguation

```
loadI 1        => r1
sub   rj, r1   => r2
loadI 10       => r3
mult  r2, r3   => r4
sub   ri, r1   => r5
add   r4, r5   => r6
loadI @A       => r7
add   r7, r6   => r8
load  r8       => rAij
```

Low level linear code:
Good for further optimisation

Directed Acyclic Graph

A directed acyclic graph (DAG) is an AST with a unique node for each value



z ← x - 2 * y
w ← x / 2

- Makes sharing explicit
- Encodes redundancy

With two copies of the same expression, the compiler might be able to arrange the code to evaluate it only once.

The third axis : naming

- The compiler must choose names for a variety of distinct values.

- a-2*b   t1<-b
  t2<-2*t1
  t3<-a
  t4<-t3-t2   t1,t2,t3 and t4 are new names

If every subexpression has a name the compiler can reuse the value of the subexpression

Naming :Stack Machine Code

Originally used for stack-based computers, now Java

- Example:

x - 2 * y   becomes

```
push x
push 2
push y
multiply
subtract
```

Advantages

- Compact form
- Introduced names are implicit, not explicit
- Simple to generate and execute code

Useful where code is transmitted over slow communication links (the net)

Implicit names take up no space, where explicit ones do!

Three Address Code

Several different representations of three address code

- In general, three address code has statements of the form:

$$x \leftarrow y \ \underline{op} \ z$$

With 1 operator (op) and, at most, 3 names (x, y, & z)

Example:

$z \leftarrow x - 2 * y$ becomes

```
t ← 2 * y
z ← x - t
```

Advantages:

- Resembles many real machines
- Introduces a new set of names
- Compact form

Three Address Code: Quadruples

Naïve representation of three address code

- Table of k * 4 small integers
- Simple record structure
- Easy to reorder
- Explicit names

The original FORTRAN compiler used "quads"

```
load  r1, y
loadI r2, 2
mult  r3, r2, r1
load  r4, x
sub   r5, r4, r3
```

RISC assembly code

load	1	y	
loadi	2	2	
mult	3	2	1
load	4	x	
sub	5	4	3

Quadruples

Three Address Code: Triples

- Index used as implicit name
- 25% less space consumed than quads
- Much harder to reorder

(1)	load	y	
(2)	loadI	2	
(3)	mult	(1)	(2)
(4)	load	x	
(5)	sub	(4)	(3)

```
load  r1, y
loadI r2, 2
mult  r3, r2, r1
load  r4, x
sub   r5, r4, r3
```

Implicit names occupy no space

Remember, for a long time, 640Kb was a lot of RAM

- Major tradeoff between quads and triples is compactness versus ease of manipulation
  - In the past compile-time space was critical
  - Today, speed may be more important

Two Address Code

- Allows statements of the form

  $x \leftarrow x$ op y

  Has 1 operator (op) and, at most, 2 names (x and y)

Example:

$z \leftarrow x - 2 * y$ becomes

```
t1 ← 2
t2 ← load y
t2 ← t2 * t1
z  ← load x
z  ← z - t2
```

- Can be very compact

Problems

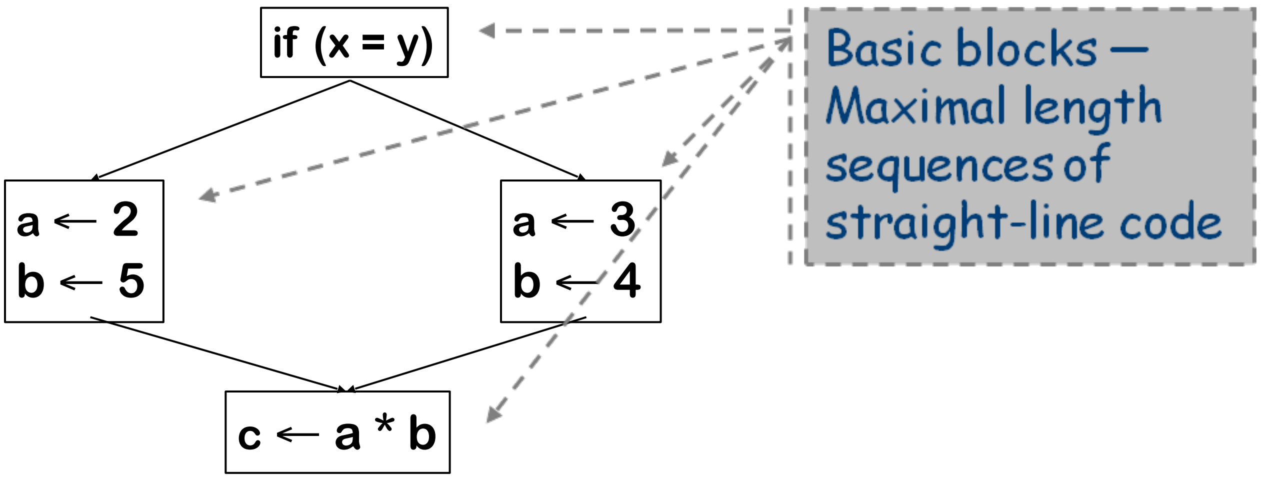
- Machines no longer rely on destructive operations
- Difficult name space
  - Destructive operations make reuse hard
  - Good model for machines with destructive ops (PDP-11)

Control-flow Graph

Models the transfer of control in the procedure

- Nodes in the graph are basic blocks
  - Can be represented with quads or any other linear representation
- Edges in the graph represent control flow

Example



Static Single Assignment Form

- The main idea: each name is defined exactly once
- The name refers to the variable in some program point
- Encodes both control and value flow
- Introduce ϕ-functions to make it work

Original

```
x ← …
y ← …
while (x < k)
    x ← x + 1
    y ← y + x
```

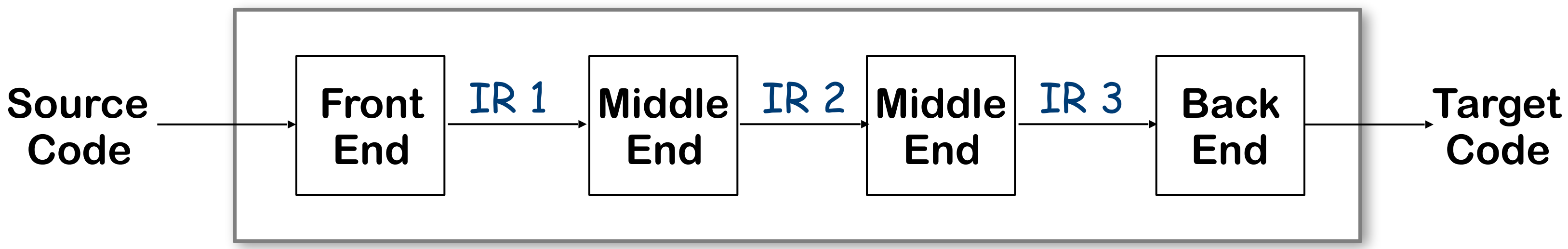
SSA-form

$$
\begin{array}{ll}
 & x_0 \leftarrow \ldots \\
 & y_0 \leftarrow \ldots \\
 & \texttt{if } (x_0 \texttt{ >= } k) \texttt{ goto next} \\
\texttt{loop:} & \quad x_1 \leftarrow \phi(x_0, x_2) \\
 & \quad y_1 \leftarrow \phi(y_0, y_2) \\
 & \quad x_2 \leftarrow x_1 + 1 \\
 & \quad y_2 \leftarrow y_1 + x_2 \\
 & \quad \texttt{if } (x_2 < k) \texttt{ goto loop} \\
\texttt{next:} & \quad \ldots
\end{array}
$$

Strengths of SSA-form

- each use refers to a single definition
- (sometimes) faster algorithms

Using Multiple Representations



- Repeatedly lower the level of the intermediate representation
  - Each intermediate representation is suited towards certain optimizations
- Example: the Open64 compiler
  - WHIRL intermediate format
    - → Consists of 5 different IRs that are progressively more detailed and less abstract

Memory Models

Two major models

- Register-to-register model
  - Keep all values that can legally be stored in a register in registers
  - Ignore machine limitations on number of registers
  - Compiler back-end must insert loads and stores

use virtual registers!

- Memory-to-memory model
  - Keep all values in memory
  - Only promote values to registers directly before they are used
  - Compiler back-end can remove loads and stores
- Compilers for RISC machines usually use register-to-register
  - Easier to determine when registers are used

The Rest of the Story…

Representing the code is only part of an IR

The compiler must discover and store many distinct kinds of information

For a variable it has to store its data type, storage class, the level of its declaring procedure, and a base and offset in memory

For an array also the number of dimension and the upper and lower bounds of each dimension

It often uses centralised information

- Symbol table
- Constant table

It needs an efficient and expandable way to realise them!

Symbol Tables

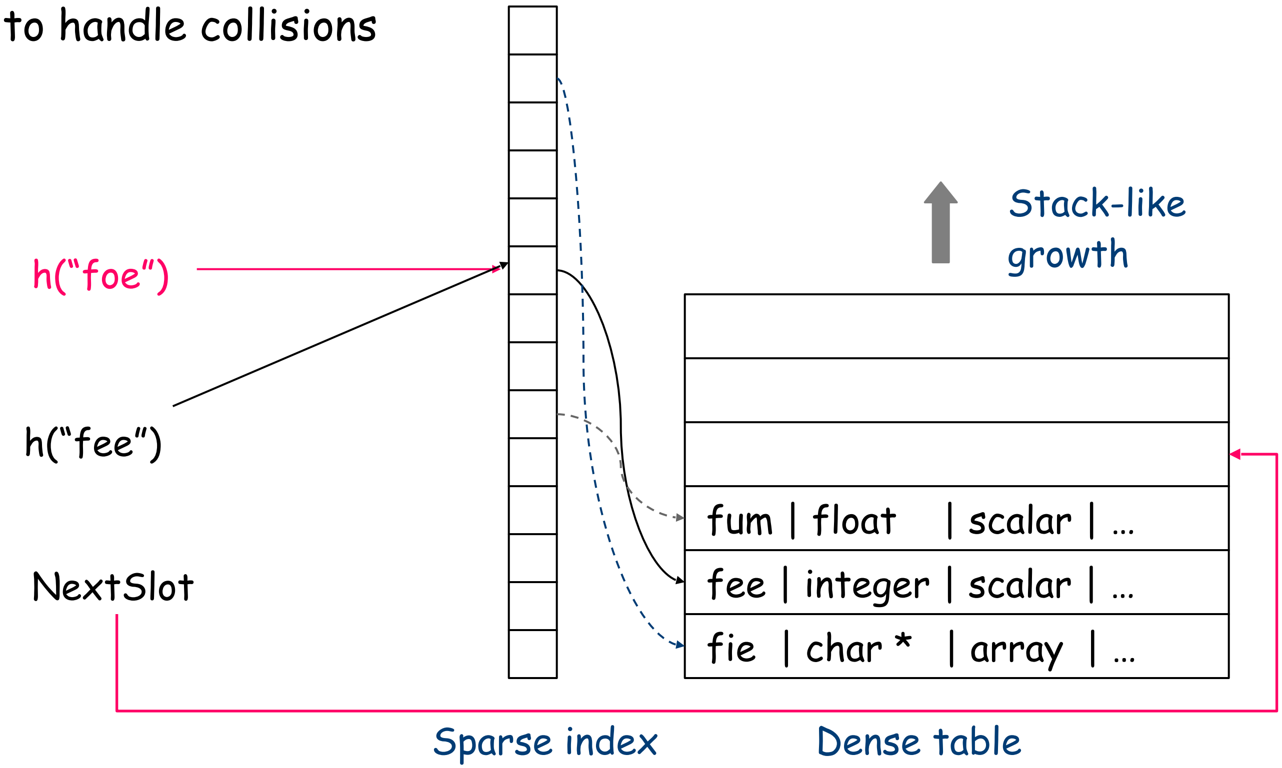
Classic approach uses hashing because we need a constant-time expected lookup

A two-table scheme

- Sparse index to reduce chance of collisions
- Dense table to hold actual data
  - → Easy to expand, to traverse, to read & write from/to files

• Use chains in index to handle collisions

Collision occurs when h() returns a slot in the sparse index that is already full.



Hash-less Symbol Tables

Classic approach to building a symbol table uses hashing

- Some concern about worst-case behavior
  - Collisions in the hash function can lead to linear search
  - Some authors advocate "perfect" hash for keyword lookup

Collision occurs when h() returns a slot in the sparse index that is already full.