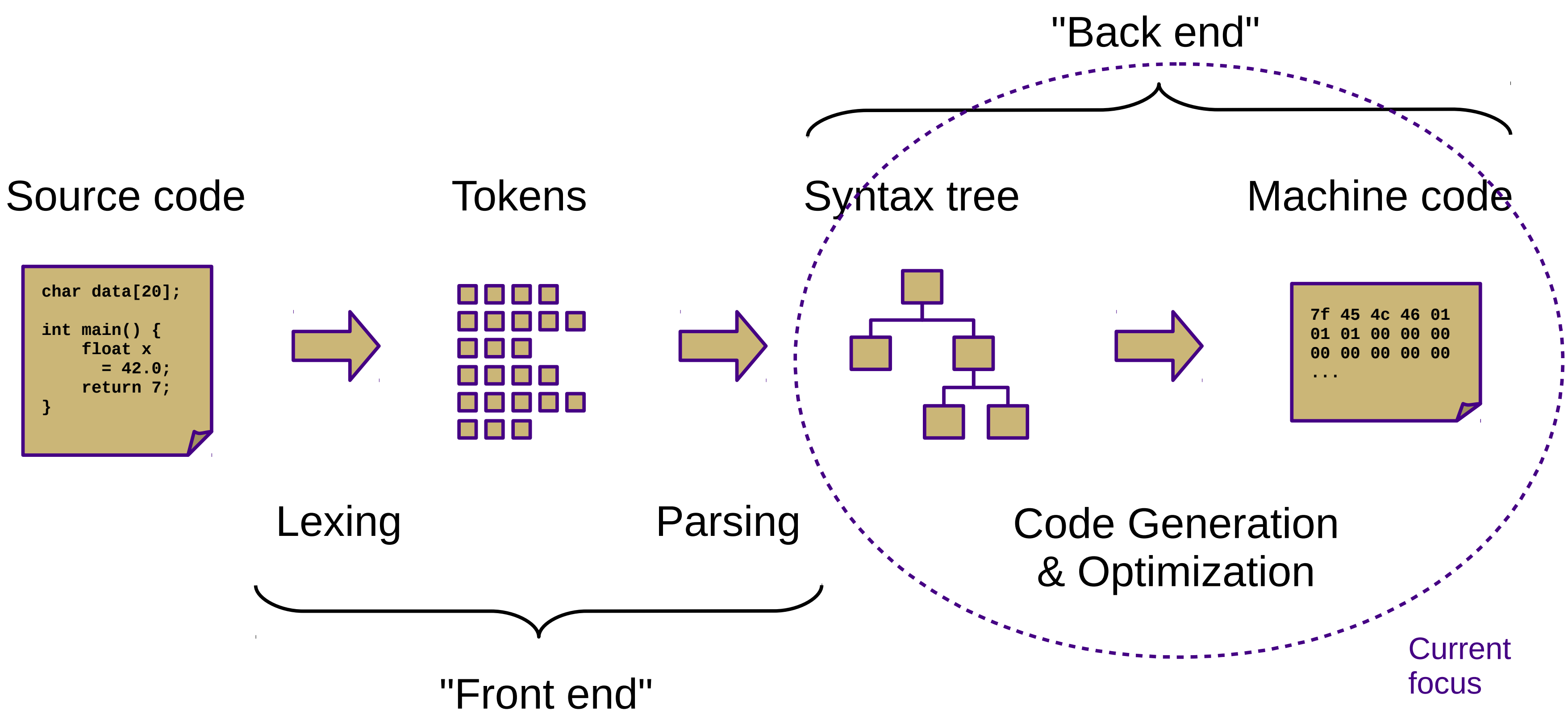
CS 432
Fall 2015

Mike Lam, Professor

Code Generation

Compilers


"Back end"
Source code
Tokens
Syntax tree
Machine code
char data[20];
int main() {
float x
= 42.0;
return 7;
}
7f 45 4c 46 01
01 01 00 00 00
00 00 00 00 00
...
Lexing
Parsing
Code Generation
& Optimization
Current
focus
"Front end"

Our Project

- Current status: type-checked AST
- Next step: convert to ILOC
  - This step is called *code generation*
  - Convert from a tree-based IR to a linear IR
    - (or directly to machine code)
    - Use a tree traversal to “linearize” the program
- But first, more general code gen topics

Goals

- Code generator outputs
  - Stack code (`push a, push b, multiply, pop c`)
  - Three-address code (`c = a + b`)
  - Machine code (`movq a, %eax; addq b, %eax; movq %eax, c`)
- Code generator requirements
  - Must preserve semantics
  - Should produce efficient code
  - Should run efficiently

Obstacles

- Generating the most optimal code is undecidable
  - Unlike front-end transformations
    - (e.g., lexing & parsing)
  - Must use heuristics and approximation algorithms
  - This is why most compilers research since 1960s has been on the back end

Phases

- Instruction selection
  - Map IR to target instructions
  - Difficulty is directly related to uniformity and completeness of target instruction set
- Register allocation/assignment
  - Allocation: selecting which variables to store in registers
  - Assignment: selecting which register to use for each variable
  - General problem is NP-complete
- Instruction scheduling
  - Optimize for pipelined architectures w/ caching
  - Take advantage of speculative execution

Syntax-Directed Translation

- Similar to attribute grammars (Figure 4.15)
- Associate bits of code with each production
  - This code performs the translation or code gen
  - Save intermediate results in temporary registers for now
- In our project, we will use a visitor
  - Still syntax-based (actually AST-based)
  - Not dependent on original grammar

ILOC

- Linear IR based on research compiler from Rice
- See Appendix A (and ILOCInstruction.java)
- I have made some modifications
  - Removed most immediate instructions (i.e., `subI`)
  - Removed binary shift instructions
  - Removed character-based instructions
  - Removed jump tables
  - Removed comparison-based conditional jumps
  - Added labels and function call mechanisms (`call`, `param`, `return`)
  - Added symbol address referencing (`loadS`)
  - Added binary `not` and arithmetic `neg`
  - Added `print` and `nop` instructions

SSA Form

- Static single-assignment
  - Naming convention that uses a unique name for each newly-calculated value
  - Values are collapsed at control flow points using Φ-functions
    - (not actual executed!)
  - Useful for various types of analysis

```
cmp_LT r1, r2 → r3
cbr r3 → l1, l2
```

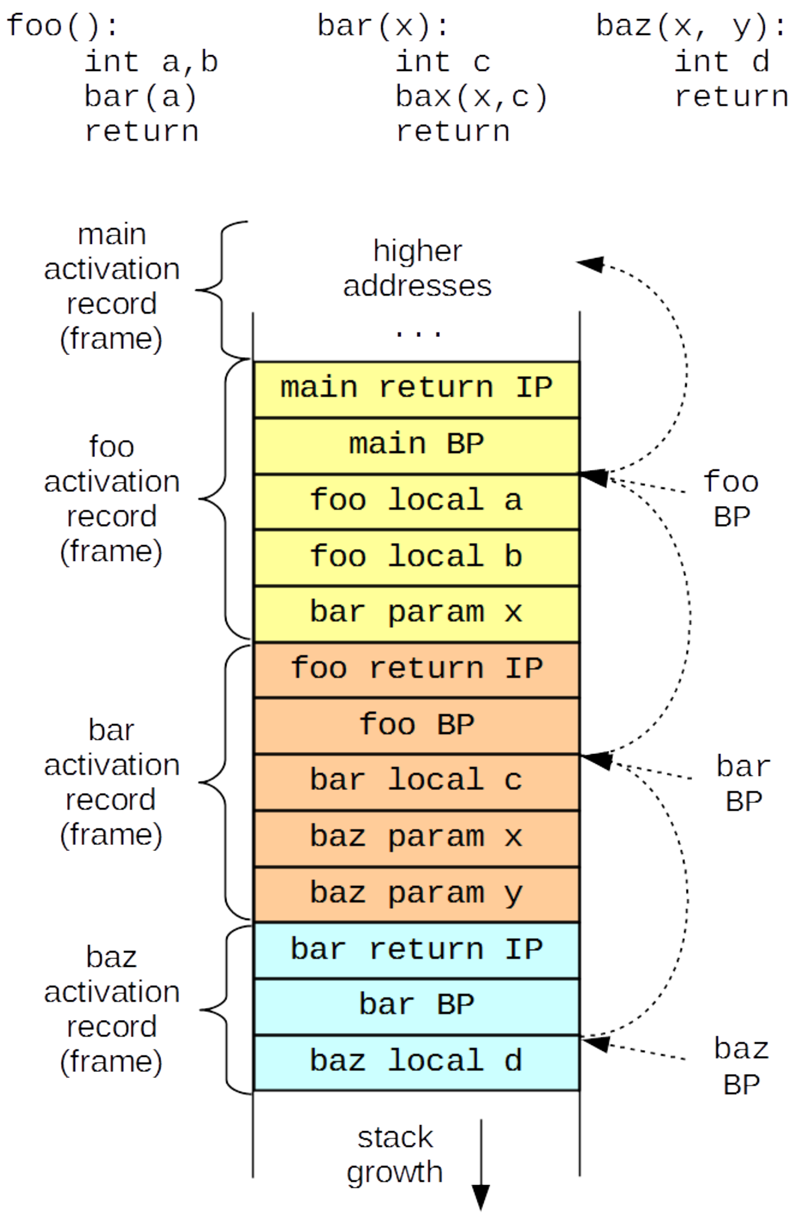
```
l1:
  loadI 4 → r4
  jmp l3
```

```
l2:
  loadI 8 → r5
  jmp l3
```

```
l3:
  r6 = Φ(r4, r5)
```

Assigning Storage Locations

- Memory regions
  - Code ("text")
  - Static ("data")
  - Heap
  - Stack
- Registers
  - General
  - Special



Boolean Encoding

- Integers: 0 for false, 1 for true
- Difference from book
  - No comparison-based conditional branches
  - Conditional branching uses boolean values instead
- Short-circuiting
  - Not in Decaf!

String Handling

- Arrays of chars vs. encapsulated type
  - Former is faster, latter is easier/safer
  - C uses the former, Java uses the latter
- Mutable vs. immutable
  - Former is more intuitive, latter is (sometimes) faster
  - C uses the former, Java uses the latter
- Decaf: immutable string constants only
  - No string variables

Array Accesses

- Generalization to multidimensions:
  - `base + (i_1 * w_1) + (i_2 * w_2) + ... + (i_k * w_k)`
- Alternate definition:
  - 1d: `base + width * (i_1)`
  - 2d: `base + width * (i_1 * n_2 + i_2)`
  - nd: `base + width * (( ... ((i_1 * n_2 + i_2) * n_3 + i_3) ... ) * n_k + i_k) * width`
- Row-major vs. column-major
- In Decaf: row-major one-dimensional global arrays

Struct and Record Types

- How to access member values?
  - Static offsets from base of struct/record
- OO adds another level of complexity
  - Now classes have methods
  - Class instance records and virtual method tables
- In Decaf: no structs or classes

Control Flow

- Introduce program labels
  - Named location in the program
  - Generated sequentially using static `newlabel`() call
- Generate goto instructions using templates
  - Also called "jumps" or "branches"
  - In ILOC: “`cbr`” instruction
  - Templates are composable

Control Flow

if statement: **if (E) B1**

```
        rE = << E code >>
        cbr rE → b1, skip
    b1:
        << B1 code >>
    skip:
```

Control Flow

if statement: **if (E) B1 else B2**

```
        rE = << E code >>
        cbr rE → b1, b2
b1:
        << B1 code >>
        jmp done
b2:
        << B2 code >>
done:
```

Control Flow

while loop: **while (E) B**

```
cond:                                   ; CONTINUE target
    rE = << E code >>
    cbr rE → body, done
body:
    << B code >>
    jmp cond
done:                                   ; BREAK target
```

Control Flow

for loop: **for V in E1, E2 B**

NOT CURRENTLY IN DECAF

```
    rX = << E1 code >>
    rY = << E2 code >>
    rV = rX
cond:
    cmp_GE rV, rY → rC
    cbr rC → done, body
body:
    << B code >>
    rV = rV + 1                    ; CONTINUE target
    jmp cond
done:                              ; BREAK target
```

Control Flow

switch statement:

switch (E) {

case V1: B1

case V2: B2

default: BD

}

NOT CURRENTLY IN DECAF

```
    rE = << E code >>
    if rE == V1 goto b1
    if rE == V2 goto b2
    << BD code >>
    jmp end
b1:
    << B1 code >>
    jmp end
b2:
    << B2 code >>
    jmp end
l3:
```

Control Flow

For sequential values starting with constant (C):

("jump table")

```
        rE = << E code >>
        jmp jt(rE)
    jt: jmp l1
        jmp l2
    (...)
```

Procedure Calls

- These are harder
  - We'll talk about them next week