Lecture Outline

1. Allocating temporaries in the activation record
   - Let’s optimize our code generator a bit
2. A deeper look into calling sequences
   - Caller/Callee responsibilities
3. Parameter passing mechanisms
   - Call-by-value
   - Call-by-reference
   - Call-by-value-result
   - Call-by-name
   - Call-by-need
4. Code generation for OO languages
   - Object memory layout
   - Dynamic dispatch

5. Code generation of data structure references
   - Address calculations
   - Array references

6. Code generation for logical expressions
   - Short-circuiting
An Optimization: Temporaries in the Activation Record

Topic 1
Review

- The stack machine has activation records and intermediate results interleaved on the stack
- The code generator must assign a location in the AR for each temporary

<table>
<thead>
<tr>
<th>AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporaries</td>
</tr>
<tr>
<td>AR</td>
</tr>
<tr>
<td>Temporaries</td>
</tr>
</tbody>
</table>

These get put here when we evaluate compound expressions like $e_1 + e_2$ (need to store $e_1$ while evaluating $e_2$)
Review (Cont.)

- Advantage: Simple code generation
- Disadvantage: Slow code
  - Storing/loading temporaries requires a store/load and \$sp adjustment

\[ cgen(e_1 + e_2) = cgen(e_1) \]
\[
\text{sw } \text{a0 0}($sp) \quad ; \text{save its value}
\]
\[
\text{addiu } \text{sp } \text{sp -4} \quad ; \text{adjust } \text{sp (!)}
\]
\[
cgen(e_2) \quad ; \text{eval } e_2
\]
\[
\text{lw } \text{t1 4}($sp) \quad ; \text{get } e_1
\]
\[
\text{add a0 t1 a0} \quad ; a0 = e_1 + e_2
\]
\[
\text{addiu } \text{sp } \text{sp 4} \quad ; \text{adjust } \text{sp (!)}
\]
An Optimization

• Idea: Predict how $sp$ will move at run time
  - Do this prediction at compile time
  - Move $sp$ to its limit, at the beginning

• The code generator must *statically* assign a location in the AR for each temporary
Improved Code

Old method
\[
c\text{gen}(e_1 + e_2) = \\
c\text{gen}(e_1) \\
\text{sw } a0 \ 0($sp) \\
\text{addiu } sp \ sp -4 \\
c\text{gen}(e_2) \\
\text{lw } t1 \ 4($sp) \\
\text{add } a0 \ t1 \ a0 \\
\text{addiu } sp \ sp \ 4
\]

New idea
\[
c\text{gen}(e_1 + e_2) = \\
c\text{gen}(e_1) \\
\text{sw } a0 \ ?($fp) \\
c\text{gen}(e_2) \\
\text{lw } t1 \ ?($fp) \\
\text{add } a0 \ t1 \ a0
\]

- statically allocate
Example

```plaintext
add(w,x,y,z)
begin
  x + (y + (z + (w + 42)))
end
```

- What intermediate values are placed on the stack?
- How many slots are needed in the AR to hold these values?
How Many Stack Slots?

- Let $NS(e) =$ # of slots needed to evaluate $e$
  - Includes slots for arguments to functions

- E.g: $NS(e_1 + e_2)$
  - Needs at least as many slots as $NS(e_1)$
  - Needs at least one slot to hold $e_1$, plus as many slots as $NS(e_2)$, i.e. $1 + NS(e_2)$

- Space used for temporaries in $e_1$ can be reused for temporaries in $e_2$
The Equations for the “Mini Bar” Language

\[
\begin{align*}
NS(e_1 + e_2) &= \max(NS(e_1), 1 + NS(e_2)) \\
NS(e_1 - e_2) &= \max(NS(e_1), 1 + NS(e_2)) \\
NS(\text{if } e_1 = e_2 \text{ then } e_3 \text{ else } e_4) &= \max(NS(e_1), 1 + NS(e_2), NS(e_3), NS(e_4)) \\
NS(f(e_1, \ldots, e_n)) &= \max(NS(e_1), 1 + NS(e_2), 2 + NS(e_3), \ldots, (n-1) + NS(e_n), n) \\
NS(\text{int}) &= 0 \\
NS(\text{id}) &= 0
\end{align*}
\]

Rule for \( f(e_1, \ldots, e_n) \): Each time we evaluate an argument, we put it on the stack
The Revised Activation Record

• For a function definition $f(x_1,\ldots,x_n)$ begin e end the AR has $2 + \text{NS}(e)$ elements
  - Return address
  - Frame pointer
  - $\text{NS}(e)$ locations for intermediate results

• Note that $f$’s arguments are now considered to be part of its caller’s AR
Picture: Activation Record

- **FP** increased by caller
- **FP−4** decreased by callee
- **x_1** pushed by caller
- **x_n** popped by callee
- **Old FP** pushed by callee
- **Return Addr.** saved by callee
- **Temp NS(e)** pushed by callee
- **Temp 1** saved by callee

*Note: This diagram disagrees slightly with previous lecture: here, the *callee* saves FP*
• Code generation must know how many slots are in use at each point

• Add a new argument to code generation: the position of the next available slot
Improved Code

Old method
\[
cgen(e_1 + e_2) = \\
cgen(e_1) \\
sw \text{ } 0($sp) \\
addiu \text{ } $sp \text{ } $sp -4 \\
cgen(e_2) \\
lw \text{ } 4($sp) \\
add \text{ } $a0 \text{ } $t1 \text{ } $a0 \\
addiu \text{ } $sp \text{ } $sp 4
\]

New method
\[
cgen(e_1 + e_2, \text{ } \text{ns}) = \\
cgen(e_1, \text{ } \text{ns}) \\
sw \text{ } ns($fp) \\
cgen(e_2, \text{ } \text{ns+4}) \\
lw \text{ } ns($fp) \\
add \text{ } $a0 \text{ } $t1 \text{ } $a0
\]

Compile-time prediction

Static allocation
Notes

• The slots for temporary values are still used like a stack, but we predict usage at compile time
  - This saves us from doing that work at run time
  - Allocate all needed slots at start of a function

Exerc. Write some code which runs slower after performing the optimization just presented
  - Hint: Think about memory usage (& caches, etc.)
A Deeper Look into Calling Sequences

Topic 2
Handling Procedure Calls and Returns

Calling sequence: a code sequence that sets up a procedure call
- allocates an activation record (model-dependent)
- loads actual parameters
- saves machine state (return address, etc.)
- transfers control to callee

Return sequence: a code sequence that handles the return from a procedure call
- deallocates the activation record
- sets up return value (if any)
- restores machine state (stack pointer, PC, etc.)
Calling Sequences: Division of Responsibilities

- The code in a calling sequence is often divided up between the caller and the callee

  Calling sequence code

  \[ \text{caller} \quad \text{callee} \]

- If there are \( m \) calls to a procedure, the instructions in the caller’s part of the calling sequence are repeated \( m \) times, while the callee’s part is repeated exactly once
  - This suggests that, for smaller code size, we should try to put as much of the calling sequence as possible in the callee
  - However, it may be possible to carry out more call-specific optimization by putting more of the code into the caller instead of the callee
Calling Sequences: Layout Issues

**General rule of thumb:**

Fields that are fixed early, are placed near the middle of the activation record

- The caller has to evaluate the actual parameters, and retrieve the return value
  - these fields should be located near the caller’s activation record

- The callee has to fill in machine status fields so that the callee can restore state on return
  - the caller should have easy access to this part of the callee’s activation record
Typical calling sequence:
1. caller evaluates actuals; pushes them on the stack
2. caller saves machine status on the stack (in the callee’s AR) and updates the stack pointer
3. caller transfers control to the callee
4. callee saves registers, initializes local data, and begins execution

Typical return sequence:
1. callee stores return value in the appropriate place
2. callee restores registers and old stack pointer
3. callee branches to the return address
Example Activation Record: The SPARC

Registers
- g0-g7  global registers
- o0-o7  outgoing args
- l0-l7  local registers
- i0-i7  incoming args

function return address
caller's o7/callee's i7

caller's frame
- locals and temporaries
- outgoing args not in o0-o5
- stack growth
- space to save o0-05 if necessary
- addr of return value
- space to save i0-i7 and l0-l7 if necessary

callee's frame
- current fp
- caller's sp
- current sp
- callee's fp
- high addresses
- low addresses

varies
- 6 words
- 1 word
- 16 words
Example Activation Record: Intel x86

- Caller's frame
  - Incoming arguments
  - Return address
  - Saved registers
  - Saved ebp
  - Locals and temporaries
- Callee's frame
  - Frame ptr
  - Ebp
  - Stack ptr
  - Esp

High addresses

Low addresses

Stack growth
Example Activation Record: MIPS R3000

- caller's frame
  - incoming arguments
  - locals and temporaries
  - callee-save registers
  - outgoing arguments
- callee's frame
- stack ptr $sp
- stack growth
- high addresses
- low addresses
Parameter Passing Mechanisms

Topic 3
Parameter Passing Mechanisms

• There are many semantic issues in programming languages centering on when values are computed and the scopes of names
  - Evaluation is the heart of computation
  - Names are most primitive abstraction mechanism

• We will focus on parameter passing
  - When are arguments of function calls evaluated?
  - What are formal parameters bound to?
First, an issue not discussed much...

Order of argument evaluation

- “Usually” not important for the execution of a program
- However, in languages that permit side-effects in call arguments, different evaluation orders may give different results

  e.g. a call $f(\text{++x}, x)$ in C

- A “standard” evaluation order is then specified
  
  C compilers typically evaluate their arguments right-to-left. Why?
Call-by-value

C uses call-by-value everywhere (except macros...) Default mechanism in Pascal and in Ada

callByValue(int y)
{
    y = y + 1;
    print(y);
}

main()
{
    int x = 42;
    print(x);
    callByValue(x);
    print(x);
}

output:
    x = 42
    y = 43
    x = 42

x’s value does not change when y’s value is changed
Call-by-reference

Available in C++ with the '&' type constructor (and in Pascal with the var keyword)

callByRef(int &y)
{
    y = y + 1;
    print(y);
}

main()
{
    int x = 42;
    print(x);
    callByRef(x);
    print(x);
}

output:
    x = 42
    y = 43
    x = 43

x’s value changes when y’s value is changed
Call-by-reference can be faked with pointers

C++:

```cpp
callByRef(int &y)
{
    y = y + 1;
    print(y);
}

main()
{
    int x = 42;
    print(x);
callByRef(x);
    print(x);
}
```

C:

```c
fakeCallByRef(int *y)
{
    *y = *y + 1;
    print(*y);
}

main()
{
    int x = 42;
    print(x);
fakeCallByRef(&x);
    print(x);
}
```

must explicitly pass the address of a local variable
Pointers to fake call-by-reference (cont.)

• It’s not *quite* the same
  - A pointer can be reassigned to point at something else; a C++ reference cannot

• The pointer itself was passed by value

• This is how arrays (they are implicitly pointers) and structures are passed in C
Call-by-value-result

Available in Ada for in out parameters
(code below in C syntax)

callByValueResult(int y, int z)
{
    y = y + 1;  z = z + 1;
    print(y);   print(z);
}

main()
{
    int x = 42;
    print(x);
callByValueResult(x, x);
    print(x);
}

output:
    x = 42
    y = 43
    z = 43
    x = 43

Note that x’s value is different from both using call-by-value and call-by-reference
What about Java?

• Primitive types (int, boolean, etc.) are always passed by value
• Objects are not quite -by-value nor -by-reference:
  - If you reassign an object reference, the caller’s argument does not get reassigned (like -by-value)
  - But if the object referred-to is modified, that modification is visible to the caller (like -by-reference)
• It’s really ordinary call-by-value with pointers, but the pointers are not syntactically obvious
Implementing Parameter Passing

**Call-by-value** (easy, no special compiler effort)
The arguments are evaluated at the time of the call and the value parameters are copied and either
- behave as *constant values* during the execution of the procedure (i.e., cannot be assigned to as in Ada), or
- are viewed as initialized *local* variables (in C or in Pascal)

**Call-by-reference**
The arguments must have allocated memory locations
The compiler passes the address of the variable, and the parameter becomes an *alias* for the argument
Local accesses to the parameter are turned into *indirect accesses*
Implementing Parameter Passing (Cont.)

Call-by-value-result

The arguments are evaluated at call time and the value parameters are copied (as in call-by-value) and used as a local variables.

The final values of these variables are copied back to the location of the arguments when the procedure exits (note that the activation record cannot be freed by the callee!)

Issues left unspecified:
- the order in which the results are copied back
- whether the locations of the arguments are calculated only on entry and stored, or whether they are recalculated on exit
Call-by-name

• Whole different ballgame: it’s like passing the text of the argument expression, unevaluated
  - The text of the argument is viewed as a function in its own right
  - Also passes the environment, so free variables are still bound according to rules of static scoping

• The argument is not evaluated until it is actually used, inside the callee
  - Might not get evaluated at all!

• An optimized version of call-by-name is used in some functional languages (e.g. Haskell, Miranda, Lazy-ML) under the names lazy evaluation (or call-by-need)
Call-by-name example (in “C++ Extra”)

callByName(int closure y)
{
    eval
    print(y);
    print(y);
    // => print(x = x+1)
}

main()
{
    int x = 42;
    print(x);
    callByName([[ x = x+1 ]]);
    print(x);
}
Code Generation for OO Languages

Topic 4
(probably not covered in lecture)
Object Layout

• Object-Oriented (OO) code generation and memory layout

• OO Slogan: If C (child) is a subclass of P (parent), then an instance of class C can be used wherever an instance of class P is expected

• This means that P’s methods should work with an instance of class C
Two Issues

• How are objects represented in memory?

• How is dynamic dispatch implemented?
Object Representation

class P {
    x : Int <- 17;
    y : String <- "Hi";
    f() : Int { x };
    z : Bool <- true;
    g() : String { y };
};

• Why method pointers?
• Why the tag?
class P { ..(same).. };

class C inherits P {
    w : Int <- 42;  // new
    f() : Int { w }; // override
    h() : Bool { z }; // new
};

Idea: Append new fields

To call f:

lw $t1 12($s0)
jalr $t1
Subclasses (Cont.)

• The offset for an attribute is the same in a class and all of its subclasses
  - Any method for an $A_1$ can be used on a subclass $A_2$
• Consider layout for $A_n < ... < A_3 < A_2 < A_1$

<table>
<thead>
<tr>
<th>Header</th>
<th>$A_1$ object</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$ attrs</td>
<td>$A_2$ object</td>
</tr>
<tr>
<td>$A_2$ attrs</td>
<td>$A_3$ object</td>
</tr>
<tr>
<td>$A_3$ attrs</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

What about multiple inheritance?
What’s the point?

• Simple
  - Just append subclass fields

• Efficient
  - Code can ignore dynamic type - just act as if it is the static type

• Supports overriding of methods
  - Just replace the appropriate dispatch pointers

• We implement type conformance (compile-time concept) with representation conformance (run-time concept)
An Optimization: Dispatch Tables

Consider 3 instances of class $\text{C}$:

<table>
<thead>
<tr>
<th>tag: C</th>
<th>x</th>
<th>y</th>
<th>f()</th>
<th>z</th>
<th>g()</th>
<th>w</th>
<th>h()</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

C.f: “return self [6]”

<table>
<thead>
<tr>
<th>tag: C</th>
<th>x</th>
<th>y</th>
<th>f()</th>
<th>z</th>
<th>g()</th>
<th>w</th>
<th>h()</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

P.g: “return self [2]”

<table>
<thead>
<tr>
<th>tag: C</th>
<th>x</th>
<th>y</th>
<th>f()</th>
<th>z</th>
<th>g()</th>
<th>w</th>
<th>h()</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Observation

- Every instance of a given class has the same values for all of its method pointers

- Space optimization: Put all method pointers for a given class into a common table, called the “dispatch table”
  - Each instance has a pointer to the dispatch table
Consider again 3 instances of C:

- Objects are smaller
- Dispatch is slower

Minor point: the offsets have changed since we removed the method pointers.
Subclassing, again

P.f: “return self [2]”

call f:
  lw $t1 4($s0)
lw $t1 0($t1)
jalr $t1

P.g: “return self [3]”

C.f: “return self [5]”

C.h: “return self [4]”
Real Object Layout

- Actually, the first 3 words of objects contain header information:

<table>
<thead>
<tr>
<th>Offset (in bytes)</th>
<th>Class Tag</th>
<th>Object Size</th>
<th>Dispatch Ptr</th>
<th>Attribute 1</th>
<th>Attribute 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
<td>8</td>
<td>12</td>
<td>16</td>
<td>...</td>
</tr>
</tbody>
</table>
Summary of Dispatch Tables

Pulled method pointers out, into separate table
- Makes objects smaller
- Makes (dynamic) dispatch slower

Q: Why don’t we do this for attributes?

Exerc. Write some code that is slower with dispatch tables (instead of embedded method pointers)
Exerc. Write some code that is faster with dispatch tables