CS 33

Data Representation (Part 4)
Floating Point

- **Single precision (float)**

<table>
<thead>
<tr>
<th>s</th>
<th>exp</th>
<th>frac</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8-bits</td>
<td>23-bits</td>
</tr>
</tbody>
</table>

- range: $\pm 1.8 \times 10^{-38} - \pm 3.4 \times 10^{38}$, ~7 decimal digits

- **Double Precision (double)**

<table>
<thead>
<tr>
<th>s</th>
<th>exp</th>
<th>frac</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11-bits</td>
<td>52-bits</td>
</tr>
</tbody>
</table>

- range: $\pm 2.23 \times 10^{-308} - \pm 1.8 \times 10^{308}$, ~16 decimal digits
Floating Point in C

- **Conversions/casting**
  - casting between `int`, `float`, and `double` changes bit representation
  - `double`/`float` → `int`
    » truncates fractional part
    » like rounding toward zero
    » not defined when out of range or NaN: generally sets to Tmin
  - `int` → `double`
    » exact conversion, as long as `int` has ≤ 53-bit word size
  - `int` → `float`
    » will round according to rounding mode
Quiz 1

Suppose \( f \), declared to be a float, is assigned the largest possible floating-point positive value (other than +\( \infty \)). What is the value of \( g = f + 1.0 \)?

a) \( f \)

b) +\( \infty \)

c) NaN

d) 0
Float is not Rational …

- Floating addition
  - commutative: \( a +^f b = b +^f a \)
    » yes!
  - associative: \( a +^f (b +^f c) = (a +^f b) +^f c \)
    » no!
    - \( 2 +^f (1e38 +^f -1e38) = 2 \)
    - \( (2 +^f 1e38) +^f -1e38 = 0 \)
Float is not Rational …

• Multiplication
  – commutative: \( a \times f b = b \times f a \)
    » yes!
  – associative: \( a \times f (b \times f c) = (a \times f b) \times f c \)
    » no!
    • \( 1e37 \times f (1e37 \times f 1e-37) = 1e38 \)
    • \( (1e37 \times f 1e37) \times f 1e-37 = +\infty \)
Float is not Rational …

• More …
  – multiplication distributes over addition:
    \[ a \times f (b + f c) = (a \times f b) + f (a \times f c) \]
    » no!
    » \[ 1e38 \times f (1e38 + f -1e38) = 0 \]
    » \[ (1e38 \times f 1e38) + f (1e38 \times f -1e38) = \text{NaN} \]
  – loss of significance:
    \[ x=y+1 \]
    \[ z=2/(x-y) \]
    \[ z==2? \]
    » not necessarily!
      • consider \( y = 1e38 \)
CS 33
Intro to Machine Programming
Machine Model

Processor (aka CPU) \(\xrightarrow{\text{instructions and data}}\) Memory (aka RAM) \(\xrightarrow{\text{data}}\)
Memory

Instructions

Data

or

Instructions are Data
Processor: Some Details

Execution engine

Instruction pointer

Condition codes
Processor: Basic Operation

while (forever) {
    fetch instruction IP points at
    decode instruction
    fetch operands
    execute
    store results
    update IP and condition code
}
Instructions ...

<table>
<thead>
<tr>
<th>Op code</th>
<th>Operand1</th>
<th>Operand2</th>
<th>...</th>
</tr>
</thead>
</table>

Operands

• Form
  – immediate vs. reference
    » value vs. address

• How many?
  – 3
    » add a, b, c
      • c = a + b
  – 2
    » add a, b
      • b += a
Operands (continued)

- **Accumulator**
  - special memory in the processor
    - known as a *register*
    - fast access
  - allows single-operand instructions
    - add a
      - acc += a
    - add b
      - acc += b
From C to Assembler ...

\[ a = (b + c) \times d; \]

\begin{verbatim}
mov b, %acc
add c, %acc
mul d, %acc
mov %acc, a
if (a<b)
c = 1;
else
d = 1;
cmp a, b
jge .L1
mov $1, c
jmp .L2
.L1
mov $1, d
.L2
\end{verbatim}
Condition Codes

- Set of flags giving status of most recent operation:
  - zero flag
    » result was zero
  - sign flag
    » for signed arithmetic interpretation: sign bit is set
  - overflow flag
    » for signed arithmetic interpretation
  - carry flag (generated by carry or borrow out of most-significant bit)
    » for unsigned arithmetic interpretation

- Set implicitly by arithmetic instructions

- Set explicitly by compare instruction
  - cmp a,b
    » sets flags based on result of b-a
Examples (1)

- Assume 32-bit arithmetic

- $x$ is 0x80000000
  - $TMIN$ if interpreted as two’s-complement
  - $2^{31}$ if interpreted as unsigned

- $x-1$ (0x7fffffff)
  - $TMAX$ if interpreted as two’s-complement
  - $2^{31}-1$ if interpreted as unsigned
  - zero flag is not set
  - sign flag is not set
  - overflow flag is set
  - carry flag is not set
Examples (2)

• **x is 0xffffffff**
  – -1 if interpreted as two’s-complement
  – UMAX \((2^{32}-1)\) if interpreted as unsigned

• **x+1 (0x00000000)**
  – zero under either interpretation
  – zero flag is set
  – sign flag is not set
  – overflow flag is not set
  – carry flag is set
Examples (3)

• x is 0xffffffff
  – -1 if interpreted as two’s-complement
  – UMAX (2^{32}-1) if interpreted as unsigned

• x+2 (0x00000001)
  – (+)1 under either interpretation
  – zero flag is not set
  – sign flag is not set
  – overflow flag is not set
  – carry flag is set
Quiz 2

• Set of flags giving status of most recent operation:
  – zero flag
    » result was zero
  – sign flag
    » for signed arithmetic interpretation: sign bit is set
  – overflow flag
    » for signed arithmetic interpretation
  – carry flag (generated by carry or borrow out of most-significant bit)
    » for unsigned arithmetic interpretation

• Set explicitly by compare instruction
  – cmp a,b
    » sets flags based on result of b-a

Which flags are set to one by “cmp 2,1”?

a) overflow flag only
b) carry flag only
c) sign and carry flags only
d) sign and overflow flags only
e) sign, overflow, and carry flags
Jump Instructions

• Unconditional jump
  – just do it

• Conditional jump
  – to jump or not to jump determined by condition-code flags
  – field in the op code indicates how this is computed
  – in assembler language, simply say
    » je
      • jump on equal
    » jne
      • jump on not equal
    » jgt
      • jump on greater than
    » etc.
Addresses

```c
int a, b, c, d;

int main() {
    a = (b + c) * d;
    ...
}
```

<table>
<thead>
<tr>
<th>Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>1012:</td>
</tr>
<tr>
<td>1008:</td>
</tr>
<tr>
<td>1004:</td>
</tr>
<tr>
<td>1000:</td>
</tr>
</tbody>
</table>

- d
- c
- b {global}
- a {variables}

- mov b, %acc
- add c, %acc
- mul d, %acc
- mov %acc, a
- mov 1004, %acc
- add 1008, %acc
- mul 1012, %acc
- mov %acc, 1000
### Addresses

```c
int b;

int func(int c, int d) {
    int a;
    a = (b + c) * d;
    ...
}
```

- One copy of $b$ for duration of program’s execution
- $b$‘s address is the same for each call to `func`
- Different copies of $a$, $c$, and $d$ for each call to `func`
- addresses are different in each call
Relative Addresses

- **Absolute address**
  - actual location in memory

- **Relative address**
  - offset from some other location

- **Blob's absolute address is 10000**
- **Datum's relative address (to Blob) is 100**
  - its absolute address is 10100
Base Registers

mov $10000, %base
mov $10, 100(%base)
### Addresses

```c
int b;

int func(int c, int d) {
    int a;
    a = (b + c) * d;
    ...
}
```

```asm
mov 1000,%acc
add -8(%base),%acc
mul -12(%base),%acc
mov %acc,-16(%base)
```
Quiz 3

Suppose the value in base is 10,000. What is the address of c?

a) 9992
b) 9996
c) 10,004
d) 10,008

mov 1000,%acc
add -8(%base),%acc
mul -12(%base),%acc
mov %acc,-16(%base)
Registers

- Instruction pointer
- Accumulator
- Base register
- More
- Condition codes

Execution engine

Interchangeable
Registers vs. Memory

- Execution engine
  - Instruction pointer
  - Accumulator
  - Base register
  - More
  - Condition codes

- Memory (aka RAM)
  - Instructions and data
  - Data
  - A relatively long distance