Memory and Data Structures

Arrays, Stacks, Queues

(Ch 10 & 16)
Memory

- This is the "RAM" in a system
- We have seen labels and addresses point to pieces of memory storing:
  - words
  - bytes
  - strings
  - numbers
- Memory is just a collection of bits
  - We could use it to represent integers
  - Or as an arbitrary set of bits
Treat memory as a giant array

• Compiler or programmer decides what use to make of it.
• The element numbering starts at 0
• The element number is an address
• In “C” to allocate some memory:

```c
char m[size_of_array];
```
Storage of Data

- LC-3 architecture is “word addressable” meaning that all addresses are “word” addresses.
- This means the smallest unit of memory we can allocate is 16-bits, a word.
- Use ‘LD’ (load) and ‘ST’ (store) to access this unit (or LDR & STR).
Example

mychar
newline


BLKW 1
.FILL xA

LD R1, newline
GETC
ST R0, mychar
JSR Sub ; R2=R1-R0
BRz found_newline

found_newline

...
The data is placed in memory like this at start up (assuming data section starts at address 1). The “mychar” variable will change to the value of the character entered by the user once stored.
Pointers and Arrays

We've seen examples of both of these in our LC-3 programs, let's see how these work in "C"

Pointer
- Address of a variable in memory
- Allows us to indirectly access variables
  - in other words, we can talk about its address rather than its value

Array
- A list of values arranged sequentially in memory
- Example: a list of telephone numbers
- Expression a[4] refers to the 5th element of the array a
Arrays

Array implementation is very important

• Most assembly languages have only basic concept of arrays (BLKW)
• From an array, any other data structure we might want can be built
Properties of arrays:
- Each element is the same size
- Elements are stored contiguously
- First element at the smallest memory address

In assembly language we must
- Allocate correct amount of space for an array
- Map array addresses to memory addresses
LC-3 declarations of arrays within memory

To allocate a portion of memory (more than a single variable’s worth)

`variableName .BLKW numElements`

`numElements` is just that, numbering starts at 0 (as in C)
Array of Integers

Calculating the address of an array element

```c
int myarray[7] /* C */
```

• If base address of "myarray" is 25
• Address of myarray[4] = 25 + 4 = 29

```plaintext
0 1 2 3 4 5 6
myarray
25 26 27 28 29 2A 2B
```

• Which is base address + distance from the first element
How do you get the address of myarray?

• Use the “load effective address” instruction, “LEA”
• Keep clear the difference between an address and the contents of an address.
To get address of `myarray[4]` in LC-3, write the code...

```
LEA   R0, myarray
ADD   R1, R0, 4
```

Now, if we wanted to increment element number 5 by 1...

```
LDR   R4, R1, 0
ADD   R4, R4, 1
STR   R4, R1, 0
```
Address vs. Value

Sometimes we want to deal with the address of a memory location, rather than the value it contains.

Recall example from Chapter 6: adding a column of numbers.
- R2 contains address of first location.
- Read value, add to sum, and increment R2 until all numbers have been processed.

R2 is a pointer -- it contains the address of data we’re interested in.
2-Dimensional Arrays

2-Dimensional arrays are more complicated in assembly

- Memory is a 1-D array
- Must map 2-D array to 1-D array
- Arrays have rows and columns
  - $r \times c$ array
  - $r = \text{rows}$
  - $c = \text{columns}$
2-Dimensional Arrays

\[ N_{ij} = \text{start} + N \]

Two sensible ways to map 2-D to 1-D

Row major form:
(rows are all together)

\[
\begin{array}{l}
0,0 \\
0,1 \\
1,0 \\
1,1 \\
2,0 \\
2,1 \\
3,0 \\
3,1 \\
\end{array}
\]

Column major form:
(columns are all together)

\[
\begin{array}{l}
0,0 \\
1,0 \\
2,0 \\
3,0 \\
0,1 \\
1,1 \\
2,1 \\
3,1 \\
\end{array}
\]

4x2 example
$$\begin{bmatrix} 0 & 1 & 2 \\ 3 & 4 & 5 \\ 6 & 7 & 8 \end{bmatrix}$$
2-Dimensional Arrays

\[
\begin{bmatrix}
0 & 1 & 2 & 3 \\
1 & 4 & 5 & 6 \\
2 & 7 & 8 & 9
\end{bmatrix}
\]

How do you calculate addresses in a 2-D array?

• **Row Major:**

  Address \((r_i, c_i) = \) Base Address + \(((r_i \times \text{Number of Cols}) + c_i) \times \text{Element size})\)

• **Column Major:**

  Address \((r_i, c_i) = \) Base Address + \(((c_i \times \text{Number of Rows}) + r_i) \times \text{Element size})\)
Fill $x_5000$

Fill $x_{4000}$

Fill $x_{2000}$
Summary of 2D arrays

- Row/Column major (storage order)
- Base address
- Size of elements
- Dimensions of the array

How about 3-D arrays?
Bounds Checking

- Many HLL’s have bounds checking (not C!!!)
- Assembly languages have no implied bounds checking
- Your program is in total control of memory
- With a 5 x 3 array, what does the following address?

```
array .BLKW 15
LEA    R1, array
ADD    R1, R1, #15
LDR    R0, R1, #0

16th

```

- Bounds checking is often a good idea!!
- Most C development environments include optional bounds checking.
Heart bleed

Open SSL

\texttt{gimme(56)}

\texttt{gimme(1024)}
Stacks

A LIFO (last-in first-out) storage structure.
  - The first thing you put in is the last thing you take out.
  - The last thing you put in is the first thing you take out.

This means of access is what defines a stack, not the specific implementation.

Two main operations:

  **PUSH**: add an item to the stack
  **POP**: remove an item from the stack
A Physical Stack

Coin rest in the arm of an automobile

Initial State  After One Push  After Three More Pushes  After One Pop

First quarter out is the last quarter in.
A Hardware Implementation

Data items move between registers

Initial State

After One Push

After Three More Pushes

After Two Pops
A Software Implementation

Data items don't move in memory, just our idea about where the TOP of the stack is.

Initial State

After One Push

After Three More Pushes

After Two Pops

By convention, R6 holds the Top of Stack (TOS) pointer.
Delete Secret.txt
Basic Push and Pop Code

For our implementation, stack grows downward (when item added, TOS moves closer to 0)

Push
- ADD R6, R6, #~1 ; decrement stack ptr
- STR R0, R6, #0 ; store data (R0)

Pop
- LDR R0, R6, #0 ; load data from TOS
- ADD R6, R6, #1 ; increment stack ptr
Pop with Underflow Detection

If we try to pop too many items off the stack, an underflow condition occurs.

- Check for underflow by checking TOS before removing data.
- Return status code in R5 (0 for success, 1 for underflow)

```
POP
LD R1, EMPTY ; EMPTY = -x4000
ADD R2, R6, R1 ; Compare stack pointer
BRz FAIL ; to x4000 to see if empty
LDR R0, R6, #0
ADD R6, R6, #1
AND R5, R5, #0 ; SUCCESS: R5 = 0
RET
FAIL AND R5, R5, #0 ; FAIL: R5 = 1
ADD R5, R5, #1
RET
EMPTY .FILL xc000 ; 2SC rep of -x4000
```
Push with Overflow Detection

If we try to push too many items onto the stack, an **overflow** condition occurs. This example assumes stack has room for 5 items.

- Check for overflow by checking TOS before adding data.
- Return status code in R5 (0 for success, 1 for overflow)

```
PUSH    LD R1, MAX     ; MAX = -x3FFB
ADD R2, R6, R1 ; Compare stack pointer
BRz FAIL    ; top address to see if full
ADD R6, R6, #-1
STR R0, R6, #0
AND R5, R5, #0 ; SUCCESS: R5 = 0
RET
FAIL AND R5, R5, #0 ; FAIL: R5 = 1
ADD R5, R5, #1
RET
MAX .FILL xC005 ; 2SC of -x3FFB
```
Stack Example

- Printing out a positive integer, character by character
- Push LSB to MSB
- Pop MSB to LSB (LIFO)

```
integer = 1024
if integer == 0 then
  push '0'
else
  while integer != 0
    digit ← integer mod base
    char ← digit + 48
    push char onto stack
    integer ← integer div base
  while stack is not empty
    pop char
    print char
```

1024 mod 10

base 10: 1024 → 102 → 102

1024 / 10

Arithmetic Using a Stack

Instead of registers, some ISA’s use a stack for source and destination operations: a zero-address machine.

- Example:
  ADD instruction pops two numbers from the stack, adds them, and pushes the result to the stack.

Evaluating \((A+B) \cdot (C+D)\) using a stack:

1. push A
2. push B
3. ADD
4. push C
5. push D
6. ADD
7. MULTIPLY
8. pop result

Why use a stack?
- Limited registers.
- Convenient calling convention for subroutines.
- Algorithm naturally expressed using LIFO data structure.
Example: OpAdd

POP two values, ADD, then PUSH result.
Example: OpAdd

OpAdd

JSR POP  ; Get first operand.
ADD R5,R5,#0  ; Check for POP success.
BRp Exit  ; If error, bail.
ADD R1,R0,#0  ; Make room for second.
JSR POP  ; Get second operand.
ADD R5,R5,#0  ; Check for POP success.
BRp Restore1  ; If err, restore & bail.
ADD R0,R0,R1  ; Compute sum.
JSR RangeCheck  ; Check size.
BRp Restore2  ; If err, restore & bail.
JSR PUSH  ; Push sum onto stack.
RET

Restore2
ADD R6,R6,#-1  ; Decr stack ptr (undo POP)

Restore1
ADD R6,R6,#-1  ; Decr stack ptr

Exit
RET
Queues

A queue is a **FIFO** (First In, First Out).
- The classic analogy of a queue is a line.
  - Person gets on the end of the line (the **Tail**),
  - Waits,
  - Gets off at the front of the line (the **Head**).
- Getting into the queue is an operation called **enqueue**
- Taking something off the queue is an operation called **dequeue**.
- It takes 2 pointers to keep track of the data structure,
  - Head (let’s use R5)
  - Tail always points to empty element (R6)
Initial state:

After 1 enqueue operation:

After another enqueue operation:
After a dequeue operation:

Like stacks, when an item is removed from the data structure, it is physically still present, but correct use of the structure cannot access it.
Implementation of a queue

Storage:

queue .BLKW infinity ; assume infinite for now
LEA R5, queue ; head
LEA R6, queue ; tail

Enqueue (item):

STR R0, R6, #0 ; R0 has data to store
ADD R6, R6, #1

Dequeue (item):

JSR SUB
BRz queue_empty ; R0 = R5-R6
LDR R1, R5, #0 ; put data in R1
ADD R5, R5, #1
Circular Queues

- To avoid infinite array, wrap around from end to beginning.
- Head == Tail means empty
- Head points to first item (for next dequeue)
- Tail point to empty location (for next enqueue)

Example of an 8 element circular queue
Circular Queues

After “enqueue’ing” one element

After “enqueue’ing” another element
Circular Queues

After “dequeuing” an element

\[
\text{Modulo}
\]
Circular Queues

Storage and initialization:

```assembly
.queue                .BLKW queue_size
queue_end             .BLKW 1
LEA R5, queue         ; head
LEA R6, queue         ; tail
```

Enqueue (item)

```assembly
STR R0, R6, #0        ; data to enqueue is in R0
ADD R6, R6, #1
LEA R1, queue_end
JSR SUB               ; R1 = R1 - R6
BRp continue1
LEA R6, queue         ; wrap around
continue1
```
Circular Queues

Deque (item):

<table>
<thead>
<tr>
<th>Instruction</th>
<th>Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>JSR</td>
<td>SUB</td>
</tr>
<tr>
<td>BRz</td>
<td>queue_empty</td>
</tr>
<tr>
<td>LDR</td>
<td>R0, R5, #0</td>
</tr>
<tr>
<td>ADD</td>
<td>R5, R5, #1</td>
</tr>
<tr>
<td>LEA</td>
<td>R1, queue_end</td>
</tr>
<tr>
<td>JSR</td>
<td>SUB</td>
</tr>
<tr>
<td>BRn</td>
<td>continue2</td>
</tr>
<tr>
<td>LEA</td>
<td>R5, queue</td>
</tr>
<tr>
<td></td>
<td>wrap around</td>
</tr>
</tbody>
</table>

; R1 = R5 - R6

; R1 = R5 - R1
Summary of data structures

- All data structures are based on the simple array.
- 2D Arrays, Stacks, Queues.
- It is all about the implementation.
- Bounds checking is important.
- If not documented can become confusing.
Floating Point Numbers

\[ 2.3 \times 2 \]

\[ 3 \times 10^{24} \]

\[ 3.4 \times 5 \times 6 \times 7 \times 8 \times 9 \times 10^{-22} \]
Floating Point Numbers

\[ 2^{-3} \]

\[ 2.1 \ 2.2 \ 2.0004 \]

- Registers for real numbers usually contain 32 or 64 bits, allowing \(2^{32}\) or \(2^{64}\) numbers to be represented.
- Which reals to represent? There are an infinite number between 2 adjacent integers. (or two reals!!)
- Which bit patterns for reals selected?
- Answer: use scientific notation

\[ 1.01 \ 1.001 \]
Floating Point Numbers

Consider: $A \times 10^B$, where $A$ is one digit

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$A \times 10^B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>any</td>
<td>0</td>
</tr>
<tr>
<td>1..9</td>
<td>0</td>
<td>1..9</td>
</tr>
<tr>
<td>1..9</td>
<td>1</td>
<td>10..90</td>
</tr>
<tr>
<td>1..9</td>
<td>2</td>
<td>100..900</td>
</tr>
<tr>
<td>1..9</td>
<td>-1</td>
<td>0.1..0.9</td>
</tr>
<tr>
<td>1..9</td>
<td>-2</td>
<td>0.01..0.09</td>
</tr>
</tbody>
</table>

How to do scientific notation in binary?
Standard: **IEEE 754 Floating-Point**
IEEE 754 Single Precision Floating Point Format

Representation:

<table>
<thead>
<tr>
<th>S</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
</table>

- **S** is one bit representing the sign of the number
- **E** is an 8 bit biased integer representing the exponent
- **F** is an 23-bit unsigned integer

The true value represented is: \((-1)^s \times f \times 2^e\)

- **S** = sign bit
- **e** = **E** - bias
- **f** = \(F/2^n + 1\)
- for single precision numbers \(n=23\), bias=127
IEEE 754 Single Precision Floating Point Format

S, E, F are all fields within a representation. Each is just a bunch of bits.

**S** is the sign bit
• \((-1)^S \rightarrow (-1)^0 = +1\) and \((-1)^1 = -1\)
• Just a sign bit for signed magnitude

**E** is the exponent field
• The E field is a biased-127 representation.
• True exponent is \((E - \text{bias})\)
• The base (radix) is always 2 (implied).
• Some early machines used radix 4 or 16 (IBM)
**IEEE 754 Single Precision Floating Point Format**

![Binary representation of a floating point number: 101.1001]  

1.011001 . 2

**F** (or **M**) is the fractional or mantissa field.  
- It is in a strange form.  
- There are 23 bits for **F**.  
- A normalized FP number always has a leading 1.  
- No need to store the one, just assume it.  
- This MSB is called the HIDDEN BIT.
How to convert 64.2 into IEEE SP

1. Get a binary representation for 64.2
   - Binary of left of radix point is: \(10000000\)
   - Binary of right of radix:
     - \(.2 \times 2 = 0.4\) \(0\)
     - \(.4 \times 2 = 0.8\) \(0\)
     - \(.8 \times 2 = 1.6\) \(1\)
     - \(.6 \times 2 = 1.2\) \(1\)
   - Binary for \(.2\): \(1111111111111111\)
   - 64.2 is: \(1.0000000000000000\)

2. Normalize binary form
   - Produces: \(1.0000000000000000 \times 25\)
Floating Point

3. Turn true exponent into bias-127

4. Put it together:
   23-bit F is: 000000 0011 0011 0011 0011 0011
   S E F is: 01000010 000000 00110001
   In hex: 0x42806666

• Since floating point numbers are always stored in normal form, how do we represent 0?
  • 0x0000 0000 and 0x8000 0000 represent 0.
Other special values:

- $+ \frac{5}{0} = \infty$
- $+\infty = 0 \overline{11111111} 00000...$ (0x7f80 0000)
- $-\frac{7}{0} = -\infty$
- $-\infty = 1 \overline{11111111} 00000...$ (0xff80 0000)
- $0/0$ or $+\infty -\infty = \text{NaN}$ (Not a number)
- NaN ? $1\overline{11111111} \overline{??????}...$
  (S is either 0 or 1, E=0xff, and F is anything but all zeroes)
- Also de-normalized numbers (beyond scope)
IEEE Floating Point

$2^{23-16}$

What is the decimal value for this SP FP number $0x4228\ 0000$?

$0100\ 0010\ 0010\ 1000\ \text{Zero} \rightarrow 1$

$100\ 00100$

$128 + 4 = 132 - 727$

$\frac{3}{5}$

$1.010100000000$

$101010$

$42$
IEEE Floating Point

\[ \frac{1}{2} + \frac{1}{8} \]

What is \( 47.625_{10} \) in SP FP format?

1.011111101 \times 2^5

E: 5 + 127 = 132

10000100

010000100 011111101 00000000000000000000000

4 2 3 1 0

5 = 0
What do floating-point numbers represent?

- Rational numbers with non-repeating expansions in the given base within the specified exponent range.
- They do not represent repeating rational or irrational numbers, or any number too small or too large.

\[ 2^{-127} - 2^{126} \]
IEEE Double Precision FP

- IEEE Double Precision is similar to SP
  - 52-bit M
    - 53 bits of precision with hidden bit
  - 11-bit E, excess 1023, representing \(-1023 \rightarrow 2046\)
  - One sign bit
- Always use DP unless memory/file size is important unless on a microcontroller
  - SP \(\sim 10^{-38} \ldots 10^{38}\)
  - DP \(\sim 10^{-308} \ldots 10^{308}\)
- Be **very** careful of these ranges in numeric computation
Floating Point Arithmetic

Floating Point operations include
- Addition
- Subtraction
- Multiplication
- Division

They are complicated because...
Floating Point Addition

Decimal Review

\[ 9.997 \times 10^2 \]
\[ + 4.631 \times 10^{-1} \]

How do we do this?

1. Align decimal points
2. Add
\[ 9.997 \times 10^2 \]
\[ + 0.004631 \times 10^2 \]
\[ \underline{10.001631 \times 10^2} \]

3. Normalize the result
   - Often already normalized
   - Otherwise move one digit
\[ 1.0001631 \times 10^3 \]

4. Possibly round result
\[ 1.000 \times 10^3 \]
Floating Point Addition

Example: $0.25 + 100$ in SP FP

First step: get into SP FP if not already

$0.25 = 0 \ 01111101 \ 000000000000000000000000000000000$

$100 = 0 \ 10000101 \ 100100000000000000000000000000000$

Or with hidden bit

$0.25 = 0 \ 01111101 \ \boxed{1} \ 000000000000000000000000000000000$

$100 = 0 \ 10000101 \ \boxed{1} \ 100100000000000000000000000000000$

Hidden Bit
Floating Point Addition

Second step: Align radix points

- Shifting F left by 1 bit, \textbf{decreasing} \( e \) by 1
- Shifting F right by 1 bit, \textbf{increasing} \( e \) by 1
- Shift F right so least significant bits fall off
- Which of the two numbers should we shift?

\[
\begin{align*}
0.1000 & \quad \rightarrow \quad 0.1000 \\
1.0000 & \quad \rightarrow \quad 1.0010
\end{align*}
\]
Floating Point Addition

Second step: Align radix points cont.

Shift the .25 to increase its exponent so it matches that of 100.

\[\begin{align*}
0.25's \; e: & \quad 01111101 - 1111111 \; (127) = -2 \\
100's \; e: & \quad 10000101 - 1111111 \; (127) = 6 \\
\end{align*}\]

Shift .25 by 8 then.

Easier method: Bias cancels with subtraction, so

\[
\begin{align*}
(a + X) - (c + b) & \quad 10000101 - 01111101 \\
& \quad \overline{00001000}
\end{align*}
\]
Floating Point Addition

Carefully shifting the 0.25’s fraction

<table>
<thead>
<tr>
<th>S</th>
<th>E</th>
<th>HB</th>
<th>F</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>01111101</td>
<td>1</td>
<td>00000000000000000000000000000000</td>
<td>(original value)</td>
</tr>
<tr>
<td>0</td>
<td>01111110</td>
<td>0</td>
<td>10000000000000000000000000000000</td>
<td>(shifted by 1)</td>
</tr>
<tr>
<td>0</td>
<td>01111111</td>
<td>0</td>
<td>01000000000000000000000000000000</td>
<td>(shifted by 2)</td>
</tr>
<tr>
<td>0</td>
<td>10000000</td>
<td>0</td>
<td>00100000000000000000000000000000</td>
<td>(shifted by 3)</td>
</tr>
<tr>
<td>0</td>
<td>10000001</td>
<td>0</td>
<td>00010000000000000000000000000000</td>
<td>(shifted by 4)</td>
</tr>
<tr>
<td>0</td>
<td>10000010</td>
<td>0</td>
<td>00001000000000000000000000000000</td>
<td>(shifted by 5)</td>
</tr>
<tr>
<td>0</td>
<td>10000011</td>
<td>0</td>
<td>00000100000000000000000000000000</td>
<td>(shifted by 6)</td>
</tr>
<tr>
<td>0</td>
<td>10000100</td>
<td>0</td>
<td>00000010000000000000000000000000</td>
<td>(shifted by 7)</td>
</tr>
<tr>
<td>0</td>
<td>10000101</td>
<td>0</td>
<td>00000000100000000000000000000000</td>
<td>(shifted by 8)</td>
</tr>
</tbody>
</table>
Floating Point Addition

Third Step: Add fractions with hidden bit

\[
\begin{array}{ccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccc
Floating Point Addition

Normalization example

```
S  E  HB  F
0  011 1  1100
+ 0  011 1  1011
---
0  011 11  0111
```

Need to shift so that only a 1 in HB spot

```
0  100  1  1011  1 -> discarded
```
Floating Point Subtraction

- Mantissa's are sign-magnitude
- Watch out when the numbers are close

\[
\begin{align*}
1.23455 \times 10^2 \\
- 1.23456 \times 10^2 \\
\end{align*}
\]

- A many-digit normalization is possible
  This is why FP addition is in many ways more difficult than FP multiplication
Floating Point Subtraction

Steps to do subtraction

1. Align radix points
2. Perform sign-magnitude operand swap if needed
   - Compare magnitudes (with hidden bit)
   - Change sign bit if order of operands is changed.
3. Subtract
4. Normalize
5. Round
Floating Point Subtraction

Simple Example:

<table>
<thead>
<tr>
<th>S</th>
<th>E</th>
<th>HB</th>
<th>F</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>011</td>
<td>1</td>
<td>1011</td>
<td>smaller</td>
</tr>
<tr>
<td>-0</td>
<td>011</td>
<td>1</td>
<td>1101</td>
<td>bigger</td>
</tr>
</tbody>
</table>

switch order and make result negative

<table>
<thead>
<tr>
<th>S</th>
<th>E</th>
<th>HB</th>
<th>F</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>011</td>
<td>1</td>
<td>1101</td>
<td>bigger</td>
</tr>
<tr>
<td>-0</td>
<td>011</td>
<td>1</td>
<td>1011</td>
<td>smaller</td>
</tr>
</tbody>
</table>

|   | 011 | 0   | 0010 |

1  0000  1  00000 switched sign, renormalized
Floating Point Multiplication

Decimal example:

$$3.0 \times 10^1$$
$$\times 5.0 \times 10^2$$

How do we do this?

1. Multiply mantissas
   $$3.0$$
   $$\times 5.0$$
   $$15.00$$

2. Add exponents
   $$1 + 2 = 3$$

3. Combine
   $$15.00 \times 10^3$$

4. Normalize if needed
   $$1.50 \times 10^4$$
Floating Point Multiplication

Multiplication in binary (4-bit F)

\[
\begin{array}{c}
0 \ 10000100 \ 0100 \\
x \ 1 \ 00111100 \ 1100 \\
\hline
10.00110000
\end{array}
\]

Step 1: Multiply mantissas
(put hidden bit back first!!)

\[
\begin{array}{c}
1.0100 \\
x \ 1.1100 \\
\hline
00000 \\
00000 \\
10100 \\
10100 \\
+ \ 10100 \\
\hline
1000110000
\end{array}
\]
Floating Point Multiplication

Second step: Add exponents, subtract extra bias.

\[
\begin{align*}
10000100 & \quad 11000000 \\
+ 00111100 & \quad - 01111111 \\
\hline
11000000 & \quad (127) \\
01000001 & \\
\end{align*}
\]

Third step: Renormalize, correcting exponent

\[1 \ 01000001 \ 10.00110000\]

Becomes

\[1 \ 01000010 \ 1.000110000\]

Fourth step: Drop the hidden bit

\[1 \ 01000010 \ 000110000\]

\[= \ 0xA10C0000\]
Floating Point Multiplication

Multiply these SP FP numbers together

\[ 0x49FC0000 \times 0x4BE00000 \]
Floating Point Division

• True division
  • Unsigned, full-precision division on mantissas
    • This is much more costly (e.g. 4x) than mult.
  • Subtract exponents
• Faster division
  • Newton’s method to find reciprocal
  • Multiply dividend by reciprocal of divisor
  • May not yield exact result without some work
  • Similar speed as multiplication
• Not covered in this class!
Floating Point Summary

- Has 3 portions, S, E, F/M
- Do conversion in parts
- Arithmetic is signed magnitude
- Subtraction could require many shifts for renormalization
- Multiplication is easier since do not have to match exponents