Floating point & Datalab

15-213 (18-213): Introduction to Computer Systems
Recitation 3, Jan. 28 2013

Jimmy Guo, Section H
FAQ page

- [http://www.cs.cmu.edu/~213/faq.html](http://www.cs.cmu.edu/~213/faq.html)
- **General/assignment issues**
  - E.g. Why do I get “permission denied” error?
  - Read through before you ask questions
  - Check back for any future update
Outline

- Integer
- IEEE floating point
  - Overview
  - Examples
- Datalab
- Questions
Integers

- **Power-of-2 Multiply/Divide with shift**
  - Multiply
    - Left shift by k
  - Divide
    - Negative number needs bias
    - \((x + (1 \ll k) - 1) \gg k\)

- **Representation in memory**
  - Big Endian vs. Little Endian
  - Keep in mind when doing Bomblab
Floating point

**Encoding**
- **Sign**
  - Symmetric on axis (thus there are both +0 and -0)
- **Exponent**
  - 3(normalized, denormalized, special) cases based on exp
  - More bits --> wider range
- **Fraction (Mantissa)**
  - Implied leading 1 (for normalized number)
  - More bits --> higher precision
Floating point

- Bias
  - Bias = $2^{k-1} - 1$, where $k$ is number of exponent bits

- Normalized vs. Denormalized
  - exp?
  - Implied leading 1 vs. Implied leading 0
  - $E = \text{exp} - \text{Bias}$ vs. $E = 1 - \text{Bias}$
  - Denser nearer the origin vs. evenly spaced

- Special case (exp = 111...1)
  - Infinity
  - NaN
Floating point

- Rounding
  - Round to even
  - Why?
    - Avoid statistical bias
  - How?
    - 1.1011 (All round to nearest 1/4)
    - 1.1010
    - 1.0101
    - 1.1110

- Addition & Multiplication
  - Associativity/distributivity may not hold
    - 3.14 + (1e20 - 1e20) vs. (3.14 + 1e20) - 1e20
Floating point (Examples)

1.a Consider the following 5-bit floating point representation based on the IEEE floating point format. This format does not have a sign bit — it can only represent nonnegative numbers.

- There are $k = 3$ exponent bits.
- There are $n = 2$ fraction bits.

What is the...

- Bias?
- The largest denormalized number?
- The smallest normalized number?
- Largest finite number it can represent?
- Smallest non-zero value it can represent?
Floating point (Examples)

1.b For the same problem, you are given some decimal values below, and your task it to encode them in floating point format. In addition, you should give the rounded value of the encoded floating point number.

<table>
<thead>
<tr>
<th>Value</th>
<th>Floating Point Bits</th>
<th>Rounded value</th>
</tr>
</thead>
<tbody>
<tr>
<td>9/32</td>
<td>001 00</td>
<td>1/4</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3/16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15/2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Floating point (Examples)

1.b For the same problem, you are given some decimal values below, and your task it to encode them in floating point format. In addition, you should give the rounded value of the encoded floating point number.

<table>
<thead>
<tr>
<th>Value</th>
<th>Floating Point Bits</th>
<th>Rounded value</th>
</tr>
</thead>
<tbody>
<tr>
<td>9/32</td>
<td>001 00</td>
<td>1/4</td>
</tr>
<tr>
<td>3</td>
<td>100 10</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>110 00</td>
<td>8</td>
</tr>
<tr>
<td>3/16</td>
<td>000 11</td>
<td>3/16</td>
</tr>
<tr>
<td>15/2</td>
<td>110 00</td>
<td>8</td>
</tr>
</tbody>
</table>
Floating point (Examples)

2. Consider the following two 7-bit floating point representation based on the IEEE floating point format. Neither has a sign bit – they can only represent nonnegative numbers.

Format A

- There are $k = 3$ exponent bits. The exponent bias is 3.
- There are $n = 4$ fraction bits.

Format B

- There are $k = 4$ exponent bits. The exponent bias is 7.
- There are $n = 3$ fraction bits.

Convert these bit patterns to the closest value in Format B.

<table>
<thead>
<tr>
<th>Format A</th>
<th>Format B</th>
</tr>
</thead>
<tbody>
<tr>
<td>011 0000</td>
<td>0111 000</td>
</tr>
<tr>
<td>101 1110</td>
<td></td>
</tr>
<tr>
<td>010 1001</td>
<td></td>
</tr>
<tr>
<td>110 1111</td>
<td></td>
</tr>
<tr>
<td>000 0001</td>
<td></td>
</tr>
</tbody>
</table>
Floating point (Examples)

2. Consider the following two 7-bit floating point representation based on the IEEE floating point format. Neither has a sign bit – they can only represent nonnegative numbers.

Format A
• There are $k = 3$ exponent bits. The exponent bias is 3.
• There are $n = 4$ fraction bits.

Format B
• There are $k = 4$ exponent bits. The exponent bias is 7.
• There are $n = 3$ fraction bits.

Convert these bit patterns to the closest value in Format B.

<table>
<thead>
<tr>
<th>Format A</th>
<th>Format B</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>011 0000</td>
<td>0111 000</td>
<td></td>
</tr>
<tr>
<td>101 1110</td>
<td>1001 111</td>
<td></td>
</tr>
<tr>
<td>010 1001</td>
<td>0110 100</td>
<td>Round down</td>
</tr>
<tr>
<td>110 1111</td>
<td>1011 000</td>
<td>Round up</td>
</tr>
<tr>
<td>000 0001</td>
<td>0001 000</td>
<td>Denorm -&gt; norm</td>
</tr>
</tbody>
</table>
Datalab Tips

- **Operator precedence**
  - $z = x << 2 + y$  

- **Edge cases**
  - $0? T_{\text{min}}?$
  - Shift by 32? (Undefined behavior!)

- **SubOK()?**

- **Use bddcheck & driver.pl**
  - Test thoroughly and provide more details on failure
  - Please do that early to avoid any grading surprise
  - Declare variables at very beginning of each function
Questions?