So Far

- Can do logical, add, subtract, multiply, divide, ...
- But........
  - what about fractions?
  - what about really large numbers?

Binary Fractions

1011₂ = 1x2³ + 0x2² + 1x2¹ + 1x2⁰
so...
101.011₂ = 1x2² + 0x2¹ + 1x2⁰ + 0x2⁻¹ + 1x2⁻² + 1x2⁻³
e.g.,
.75 = 3/4 = 3/2² = 1/2 + 1/4 = .11

Recall Scientific Notation

<table>
<thead>
<tr>
<th>sign</th>
<th>exponent</th>
<th>mantissa</th>
</tr>
</thead>
<tbody>
<tr>
<td>+6.02</td>
<td>23</td>
<td>1.673 x 10²⁴</td>
</tr>
</tbody>
</table>

Issues:
- Arithmetic (+, -, *, /)
- Representation, Normal form
- Range and Precision
- Rounding
- Exceptions (e.g., divide by zero, overflow, underflow)
- Errors
- Properties (negation, inversion, if A = B then A - B = 0)

Floating-Point Numbers

Representation of floating point numbers in IEEE 754 standard:

<table>
<thead>
<tr>
<th>sign</th>
<th>exponent</th>
<th>mantissa</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>E</td>
<td>M</td>
</tr>
</tbody>
</table>

- single precision

N = (-1)ˢ 2^(E-127) (1.M)

0 = 00000000 0 . . . 0
-1.5 = 10111111 10 . . . 0
0.325 = 101000101 X 2⁻² = 1.01000101 X 2⁻²
=.02 = .001100110100... X 2⁻³ = 1.1001101100... X 2⁻³

- range of about 2 X 10⁻³⁸ to 2 X 10³⁸
- always normalized (so always leading 1, thus never shown)
- special representation of 0 (E = 00000000) (why?)
- can do integer compare for greater-than, sign
What do you notice?

- 0
- 1.5 * 2^{-100}
- 1.75 * 2^{-100}
- 1.5 * 2^{100}
- 1.75 * 2^{100}

- Does this work with negative numbers, as well?

Double Precision Floating Point

Representation of floating point numbers in IEEE 754 standard:

\[
N = (-1)^S \times 2^{E-1023} \times (1.M)_{10}
\]

- range of about \(2 \times 10^{-308}\) to \(2 \times 10^{308}\)

Floating Point Addition

- How do you add in scientific notation?
  \(9.962 \times 10^4 + 5.231 \times 10^2\)

- Basic Algorithm
  1. 
  2. 
  3. 
  4. 

FP Addition Hardware
Floating Point Multiplication

- How do you multiply in scientific notation?
  \((9.9 \times 10^4)(5.2 \times 10^2) = 5.148 \times 10^7\)

- Basic Algorithm
  1. Add exponents
  2.
  3.
  4.
  5. Set Sign

FP Accuracy

- Extremely important in scientific calculations
- Very tiny errors can accumulate over time
- IEEE 754 FP standard has four rounding modes
  - always round up (toward \(+\infty\))
  - always round down (toward \(-\infty\))
  - truncate
  - round to nearest
    => in case of tie, round to nearest even
- Requires extra bits in intermediate representations

Extra Bits for FP Accuracy

- *Guard bits* -- bits to the right of the least significant bit of
  the significand computed for use in normalization (could become significant at that point) and rounding.
- IEEE 754 has two extra bits and calls them _______ and _______.

Key Points

- Floating Point extends the range of numbers that can be represented, at the expense of precision (accuracy).
- FP operations are very similar to integer, but with pre- and post-processing.
- Rounding implementation is critical to accuracy over time.