Floating Point Arithmetic

Most scientific and engineering computations require "decimal" arithmetic, i.e. numbers containing a decimal point. Floating point is the computer implementation of "real" arithmetic with limited precision. Until recently, only the largest computers had floating point hardware as standard equipment.

Terminology

Scientific Notation: \(3.1557 \times 10^9\), \(3.1416 \times 10^6\), \(3.1557 \times 10^{-9}\)

Normalized Number: \(31.557 \times 10^8\) ---\> \(0.31557 \times 10^{10}\)

- General Form of Floating Point: \(1.fffffffffff \times 2^{eee}\)
- Constituents are: sign, significand or mantissa and exponent

Biased Representation

- Signed exponents would complicate comparisons
- In biased notation the most negative number is \(000...0\) and the most positive is \(111...1\)
- Since the single precision exponent field is 8 bits, allowing 256 different configurations, the bias for sp fp is 127
  - \(+2\) is presented as \(2+127 = 129 = 1000\,0001\)
  - \(-2\) is represented as \(-2+127 = 125 = 0111\,1101\)
- The bias for double precision is 1023
- The formula: \((-1)^{\text{sign}} \times (1+\text{mantissa}) \times 2^{(\text{exponent}-\text{bias})}\)

Example Representations

- Find floating point for 5.125
  - \(5.125 = 5 + 0.125 = 5+1/8 = 5+1\times2^{-3}\)
  - \(= 101_2 + .001_2 = 1.001_2\)
  - Normalize: \(1.001_2 \times 2^{0} \rightarrow 1.001 \times 2^{3}\)
  - Thus 5.125 = \((-1)^0 \times 1 + 0.0111 \times 2^{0}\)
- In reverse, what floating point number is
  - \((-1)^0 \times (1.0111 000 0...)_2 \times 2^{128}\)
  - In binary scientific notation it is \(-1.0111 \times 2^{128}\)
- Recall that fp is scientific notation, so arithmetic is logarithmic
  - Add the exponents (reduce by the bias), multiply the mantissas and renormalize if needed

Multiplying Floating Point Numbers

- \(0.75 \times 2^4 = 0.11 \times 11000\)
- \(1.1x2^7\) times \(1.1x2^4\)
  - Add exponents: \(-1 + 4 = 3\)
  - Multiply fractions:
    - \(1.1\)
    - \(1.1\)
    - Result: \(10.010 \times 2^{10}\) or \(1.0010 \times 2^{13}\)
- \(0.75 \times 2^4 = 0.11 \times 11000\)
  - \(1.1x2^{12}\) times \(1.1x2^{12}\)
  - Add exponents:
    - \(12 + 12 = 257 - 127 = 130\)
  - Multiply fractions:
    - \(1.1\)
    - \(1.1\)
    - Result: \(10.01x2^5\) as normalized
Adding Floating Point Numbers

- Requires that the binary points be aligned
- Equivalent to having the same exponent
- Shift the mantissa of the smaller right, raising its exponent

$$1.000 \times 2^{-1} + 1.011 \times 2^{2} = (0.5 + 5.5)$$

Shift smaller right: $$1.000x2^{-1} = 0.100x2^0 = 0.0100x2^1 = 0.001x2^2$$
Renormalize: $$1.100 \times 2^2$$ ... it's OK
Result: $$1.100 \times 2^2$$

Floating Point Instructions

- There are 32 fp registers: $$f0, f1, \ldots f31$$
  - The even numbered registers are used for sp
  - An even/odd pair is used for dp. With the odd numbered register holding the lsb mantissa bits
- Special load/store instructions move fp data to/from mem
  - l.s, s.s, l.d, s.d
- Arithmetic operations (R-type) come in sp/dp forms
  - add.s, add.d, sub.s, sub.d, mul.s, mul.d
- Comparisons make direct tests and set a condition bit
  - c.le.s, c.lt.s, c.eq.s, c.ne.s, c.gt.s, c.ge.s
  - c.le.d, c.lt.d, c.eq.d, c.ne.d, c.gt.d, c.ge.d
- Branch if true, bc1t, and branch if false, bc1f

To Infinity and Beyond

- IEEE 754 reserves certain representations for extreme conditions:

<table>
<thead>
<tr>
<th>Single</th>
<th>Double</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp</td>
<td>Signif</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>nonzero</td>
</tr>
<tr>
<td>1-254</td>
<td>anything</td>
</tr>
<tr>
<td>255</td>
<td>0</td>
</tr>
<tr>
<td>255</td>
<td>nonzero</td>
</tr>
</tbody>
</table>