Lesson 17: Floating-Point Numbers

Prof. John Wenskovitch

10/17/2014
Last Time

• The negative numbers lecture I should have given long ago.

Today

• How do we express floating-point numbers?
Floating Point Numbers

• We can add a “point” in binary notation:
  – 101.1010
  – Integer part = 101 = 5
  – Fractional part = $1 \times 2^{-1} + 1 \times 2^{-3} = 0.625$
  – Full number = 5.625

• Normal Form: Shift “point” so that there’s only a single leading 1 (Think of it like scientific notation)
  – 101.1010 = 1.011010 $\times 2^2$, shift left by 2 positions
  – 0.0001101 = 1.101 $\times 2^{-4}$, shift right 4 positions
Floating Point (con’t)

• Think of the decimal number: \(-1.4 \times 10^{-2}\)
• What components do we have?
  – Sign
  – Significand
  – Exponent
• More bits in significand = higher accuracy
• More bits in exponent = wider range
IEEE 754

• A standard for representing FP numbers in computers
  – Single precision (32-bit): 8-bit exponent, 23-bit significand
  – Double precision (64-bit): 11-bit exponent, 52-bit significand

• \((-1)^{\text{sign}} \times (1 \circ \cdots \circ \text{significand}) \times 2^{\text{exponent}}\)

• Leading 1 in significand is implicit
• Exponent is a biased number
“Biased Numbers”

• Yet another binary number representation
  – 000….000 is the smallest number
  – 111….111 is the largest number

• To get the “real” value, subtract a pre-determined bias from the unsigned evaluation of the bit pattern
  – In other words, \( \text{representation} = \text{value} + \text{bias} \)
  – Bias for the exponent field in IEEE 754:
    • Single-precision: 127
    • Double-precision: 1023
Example 1

\((-0.75)_{10} = (x)_2\)

- +0.75 = 0.11, shift it left 1 to get 1.1
- Sign bit is 1 – number is negative
- Significand is 1; the leading 1 is implicit
- Exponent is 126 (-1 + 127 = 126 in biased rep.)

\[1011111010000000000000000000000\]
Example 2

\[ (+15.625)_{10} = (x)_2 \]

- Integer part is 15 = 1111
- +0.625 = 0.101
- Now we have 1111.101, shift right 3 to get 1.111101
- Sign bit is 0 – number is positive
- Significand is 111101; the leading 1 is implicit
- Exponent is 130 (3 + 127 = 130 in biased rep.)

\[ 0100000101111010000000000000000000 \]
Adding/Subtracting

- Must equalize exponents first
  - Think of the decimal equivalent:
    - $3 \times 10^6 + 4 \times 10^6 = 7 \times 10^6$
    - $3 \times 10^6 + 4 \times 10^7 \neq 7 \times 10^6$ or $7 \times 10^7$

1. Align binary points, shift the smaller number right until the exponents match.
2. Add the significands
3. Shift sum back to restore only one leading 1
   - Remember to check for overflow/underflow!
Overflow & Underflow

• **Overflow** – the exponent is too large to fit in the exponent field

• **Underflow** – the exponent is too small to fit in the exponent field
  – Easy fix – rounding to 0
Multiplying/Dividing

• Multiply/divide significand, add/subtract exponents
  – Again, consider the decimal equivalent:
    • \(3 \times 10^6 \times 4 \times 10^6 = (3 \times 4) \times 10^{6+6} = 12 \times 10^{12} = 1.2 \times 10^{13}\)

1. Add the exponents, correcting for the bias
2. Multiply the significands
3. Shift product back to store only one leading 1
4. Set sign
Any Questions?