Floating point

Recap:

$$1 \leq \cdot \leq 2$$

$$217.3 = (\underbrace{1.\underline{1}\ \underline{0}\ \underline{1}\ \underline{1}}_{})_2 \cdot 2^{\underline{7}}$$

$$(1\ \_\ \_\ \_\ \_\ .)$$

Unrepresentable numbers?

- Can you think of a somewhat central number that we cannot represent as

$$x = (1.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_)_2 \cdot 2^{-p}?$$

0 → To store 0: significand: all zeros + special exp. −1024

Exp. range −1023 ... 1024

−1024

special meaning: "turn off 'the leading 1'"

Demo: Picking apart a floating point number

$$\underbrace{\text{Exponent}} = \underbrace{-1023}_{\substack{\text{implicit} \\ \text{(not stored}}} + \underbrace{\text{stored}}$$

$$(3)_{10} = (11)_2 = (1.1) \cdot 2^1 = (1.1) \cdot 2^{-1023+1024}$$

$$2^{-1022} = (1. \qquad\qquad) \cdot 2^{-1022}$$

normal

$$2^{-1023} = (\text{"}0.\boxed{1\,0\,0\,0\,0\,0\,0}\text{"}) \cdot 2^{-1023}$$

subnormal

$$0 = (0.0\,0\,0\,0\,0\,0\,0) \cdot 2^{-1023}$$

Subnormal Numbers

- What is the smallest normal representable number in an FP system with 4 stored bits in the significand and an exponent range of $[-7, 7]$?

$$(1.\underline{0}\,\underline{0}\,\underline{0}\,\underline{0})_2 \cdot 2^{\underbrace{-7}} = 2^{-7}$$

$$(0\,\underline{1}.\underline{0}\,\underline{0}\,\underline{0}) \cdot 2^{\underbrace{-8}} = 2^{-8}$$

not stored | stored

$$(\underline{0.001})_2 \cdot 2^{-8} = 2^{-11}$$

"FP assist" → working w/ subnormal numbers

→ super slow

Demo: Density of Floating Point Numbers
Demo: Floating Point vs. Program Logic

"Underflow"

"Subnormals" → "gradual underflow"

Floating Point and Rounding Error

What is the relative error produced by working with floating point numbers?

- What is smallest floating point number > 1? Assume 4 stored bits in the significand.

  $$(1.\overset{2^{-1}}{\_}\,\_\,\_\,\overset{1}{\_})_2 \approx 1 + \underbrace{2^{-4}}_{\text{machine eps}}$$

- What's the smallest FP number $>$ 1024 in that same system?

  ULP: unit in the last place

- Can we give that number a name?

- What does this say about the relative error incurred in floating point calculations?

  Rel. error introduced in every f.p. op. is $\sim$ machine eps

- What's that same number for double-precision floating point? (52 bits in the significand)

  2^{-52}

Implementing Arithmetic

- How is floating point addition implemented?
  Consider adding $a = (1.101)_2 \cdot 2^1$ and $b = (1.001)_2 \cdot 2^{-1}$ in a system with three bits in the significand.

shift onto same exponent

$$\begin{array}{l} (1.101)_2 \cdot 2^1 \\ + (1.001)_2 \cdot 2^{-1} \\ \hline \end{array}$$

$$\rightsquigarrow (1.101)_2 \cdot 2^1$$

$$\rightsquigarrow (0.010\cancel{01})_2 \cdot 2^1$$

$$(1.\underline{1}\,\underline{1}\,\underline{1}) \cdot 2^1$$

$$(1.\_\,\_\,\_)$$

Demo: Floating point and the harmonic series

Problems with FP Addition

- What happens if you subtract two numbers of very similar magnitude? As an example, consider $a = (1.1011)_2 \cdot 2^0$ and $b = (1.1010)_2 \cdot 2^0$.

Demo: Catastrophic Cancellation

In-class activity: Floating Point 2