Overview

- FP
- Inclass
-

- IEF debrief (next Tue)
- Examlet 2
  (extended!)
- HW4 (out soon)

$4.2 = 1.000\,0110? \cdot 2^2$

$- 4.1 = 1.000\,0001? \cdot 2^2$

$0.000\,0101$

$1.??\cdot 2^{-3}$

2^{-52}

2^0

$1.\underbrace{\qquad\qquad}_{52} \cdot 2^?$

$1.23 \cdot 10^2$

$1.23\underbrace{45}$

Unum

↑

Universal number

$$(1.1011)_2 \cdot 2^5$$

$$\left[\cancel{+\ 1.0010 \cdot 2^3}\right]$$

$$\cancel{+\ 0.0100\,|\,10 \quad 2^5}$$

$$
\begin{array}{r}
+\ 0.0101 \cdot 2^5 \\
\hline
10.0000
\end{array}
$$

$$1.\underbrace{\qquad\qquad}_{\text{fraction}} \cdot 2^{e_{meaning}}$$

$-1023 \to -1022$ turn off ld. 1

$$e_{meaning} = \begin{cases} -1023 + e_{stored} & e_{meaning} > -1023 \\ \underline{-1022} & e_{meaning} \le -1023 \end{cases}$$

turn off ld. 1

store: -1023

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

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~~ 3 stored bits in the significand.

$$x = 1.\underbrace{0}\underbrace{0}\underbrace{1} \cdot 2^0$$

$$\uparrow 2^{-3}$$

$$\frac{x-1}{1} = 2^{-3} = 2^{-3}$$

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

$$y = 1.\underbrace{0}\underbrace{0}\underbrace{1} \cdot 2^{\underbrace{10}}$$

$$1024 = 2^{10}$$

$$\frac{y - 1024}{1024} = \frac{2^{10} \cdot 2^{-3}}{1024} = 2^{-3}$$

> Can we give that number a name?

value of last digit of significand:

machine epsilon

What does this say about relative error?

Relative error in rounding to FP: machine epsilon

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.

$\checkmark$

Demo: Floating point and the harmonic series

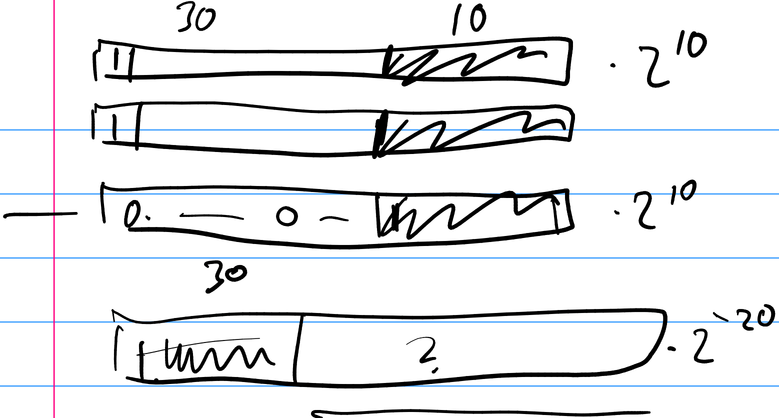
$$\sum_{i=1}^{\infty} \frac{1}{i}$$

$$1.\underline{\qquad\qquad\qquad\qquad}0 \cdot 2^{\underline{exp}}$$

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 $b = (1.1010)_2 \cdot 2^0$.

$\checkmark$



$(1. \quad)_2$ wrong

1000.1

1000.9

4

10^{-4}

Demo: Catastrophic Cancellation
In-class activity: Floating Point 2

Outline