- Form invites - one question / concor per post - numeriand linedy is norms/erves/ contr. C7 50 VIN 800 "1 error de Aigil" -> 0.1 2-52 .001 "Laanrok digids" > 0.01 0.000314 7777 000 FEE 115 000,

Rounding Modes

How is rounding performed? (Imagine trying to represent π .)



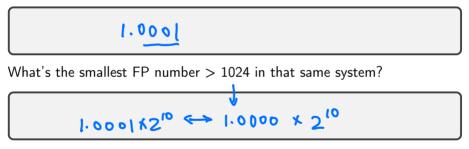
representable

What is done in case of a tie? $0.5 = (0.1)_2$ ("Nearest"?)

Demo: Density of Floating Point Numbers [cleared] **Demo:** Floating Point vs Program Logic [cleared]

Smallest Numbers Above...

▶ What is smallest FP number > 1? Assume 4 bits in the significand.



Can we give that number a name?

Unit Roundoff

[.100[1.1010

Unit roundoff or machine precision or machine epsilon or $\varepsilon_{\rm mach}$ is the smallest number such that

 $\mathsf{float}(1+\varepsilon) > 1.$

- Assuming round-to-nearest, in the above system, $\varepsilon_{mach} = (0.0001)_2$.
- Note the extra zero.

Another, related, quantity is ULP, or unit in the last place. ($\varepsilon_{mach} = 0.5 \text{ ULP}$)

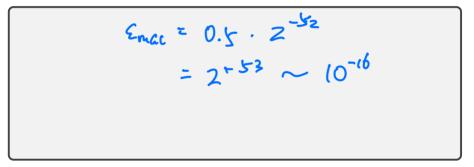
FP: Relative Rounding Error

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

 $\frac{\pi}{2} - \frac{\pi}{2} (1+\xi)$ = 121 = Emoch $\left|\widehat{\chi}\right| = \left|\chi(1+\varepsilon)\right|$

FP: Machine Epsilon

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



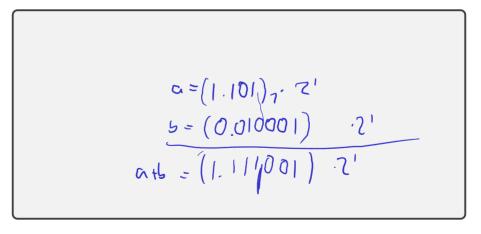
Demo: Floating Point and the Harmonic Series [cleared]

In-Class Activity: Floating Point

In-class activity: Floating Point

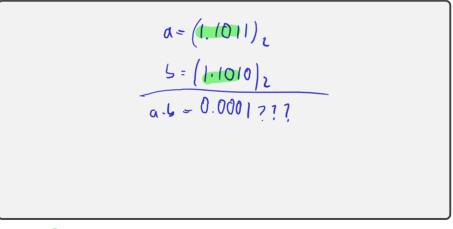
Implementing Arithmetic

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.



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 [cleared]

Supplementary Material

- ► Josh Haberman, Floating Point Demystified, Part 1
- David Goldberg, What every computer programmer should know about floating point