# CS 314 Lecture 7

Recall the format

$\underbrace{(\text{Sign})1.\boxed{\text{mantissa}}}_{m} \times 2^{\overset{\text{Exponent}}{E}}$

$a = m \times 2^E$

$b = p \times 2^F$

If $E = F$

$a + b = (m+p) \times 2^E$

$a = 1.100_b \times 2^1 = 3$

$b = 1.000_b \times 2^1 = 2$

$(10.100_b \times 2^1$

Renorm

$1.0100_b \times 2^2 = 5$

If $E \neq F$

$a = 3 = 1.100_b \times 2^1$

$b = 3/4 = 1.100_b \times 2^{-1}$

$1.10000 \times 2^1$

$.01100 \times 2^1$

$1.11100 \times 2^1$

$= (\underbrace{1 + \tfrac{1}{2}}_{3} + \underbrace{\boxed{\tfrac{1}{4} + \tfrac{1}{8}}}_{3/4}) \times 2$

## What about Rounding?

$a = 1$

$b = (1.11)_b \times 2^{-23}$

(Assume IEEE Single)

add a+b

1.000 000 000 000 000 000 000 0000

0.000 000 000 000 000 000 000 0111

$1.\underbrace{\qquad\qquad}_{22\ 0\text{'s}}111$

How should we round?

- Floor
- Round to nearest.
- Truncate
- Ceil
- Round to an Even value.

## IEEE Rounding Modes

(§ 5.5 in book)

**Definition** a floating pt # is a value that can be represented exactly in a FP # System.

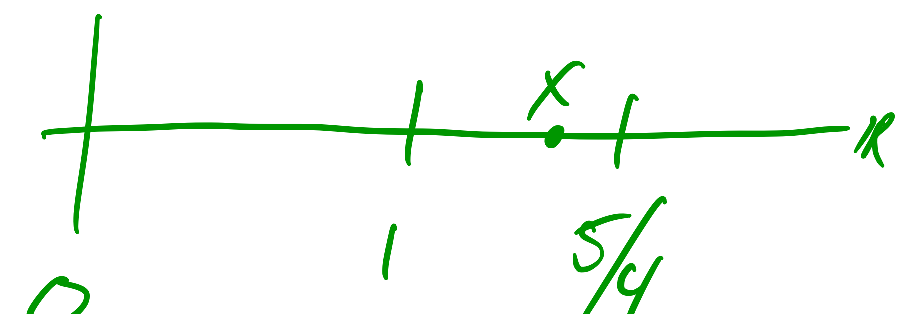
Rounding Modes

- Round to Nearest
- Round up
- Round down
- Round to zero

$\text{round}(x) = \text{FP}\#$

$x \in \mathbb{R}$

$\text{round}: \mathbb{R} \to \text{FP}\#$

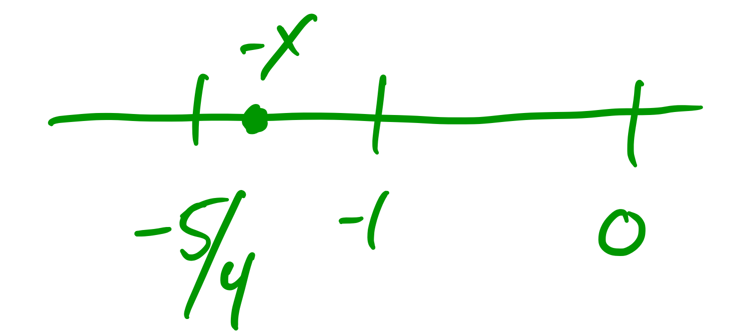


round(x) = 5/4 Nearest

5/4 up

1 down

1 zero



up = -1

zero = -1

down = -5/4

## Quiz

In our toy FP System

**Theorem** For all Rounding modes in IEEE

$\text{round}(\overset{\mathbb{R}}{x}) = \overset{\text{FP}\#}{\boxed{x(1+\delta)}}$

where $|\delta| \le \textcircled{\varepsilon}$

Machine $\varepsilon$.

Pf: Rounding modes only change the last bit.

More notation

$\overbrace{\underset{\uparrow\ \text{FP}\#}{a} \oplus \underset{\uparrow\ \text{FP}\#}{b}}^{\text{FP}\#} = \text{round}\underbrace{(a+b)}_{\text{May not be a FP}\#}$

$= (a+b)(1+\delta_1)$

$w/ |\delta_1| \le \varepsilon$

$a \oplus b$ is what you get on your Computer

IEEE guarantees

$a \oplus b = (a+b)(1+\delta_1)$

$a \ominus b = (a-b)(1+\delta_2)$

$a \otimes b = (ab)(1+\delta_3)$

$a \oslash b = (a/b)(1+\delta_4)$

Implementing this guarantee is $\boxed{\text{HARD}}$