Floating Point

Math 426

University of Alaska Fairbanks

September 25, 2020

Rounding modes

9-bit format
mantissa: 4 bits

Compute $(1.0000)_2 \times 2^0 + (1.11)_2 \times 2^{-3}$

$$\begin{array}{r} {}^{1} \\ 1.0011 \\ 0.0001 \\ \hline 1.0100 \end{array}$$

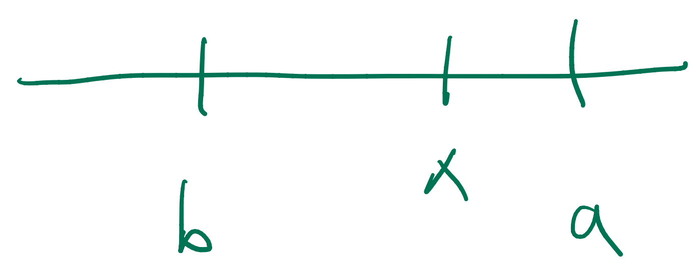
1.00111

Modes:

1. up ($\to \infty$):
2. down ($\to -\infty$):
3. zero ($\to 0$):
4. nearest (default)

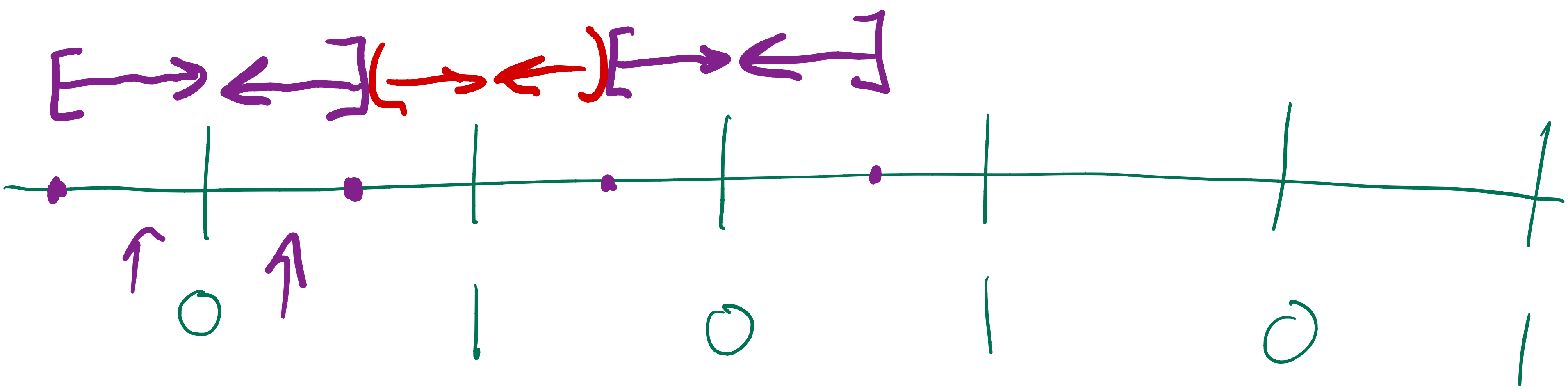
1.0100 a
1.0011 b

$$\begin{array}{r} {}^{1\,1} \\ 1.0011 \\ 0.0001 \\ \hline 1.0100 \end{array}$$



Round to nearest

Special rule if the number is exactly halfway between the two nearest representable numbers: result is the unique nearby representable number with a 0 in its least significant digit.



Rounding Error

$\varepsilon \to$ machine ε

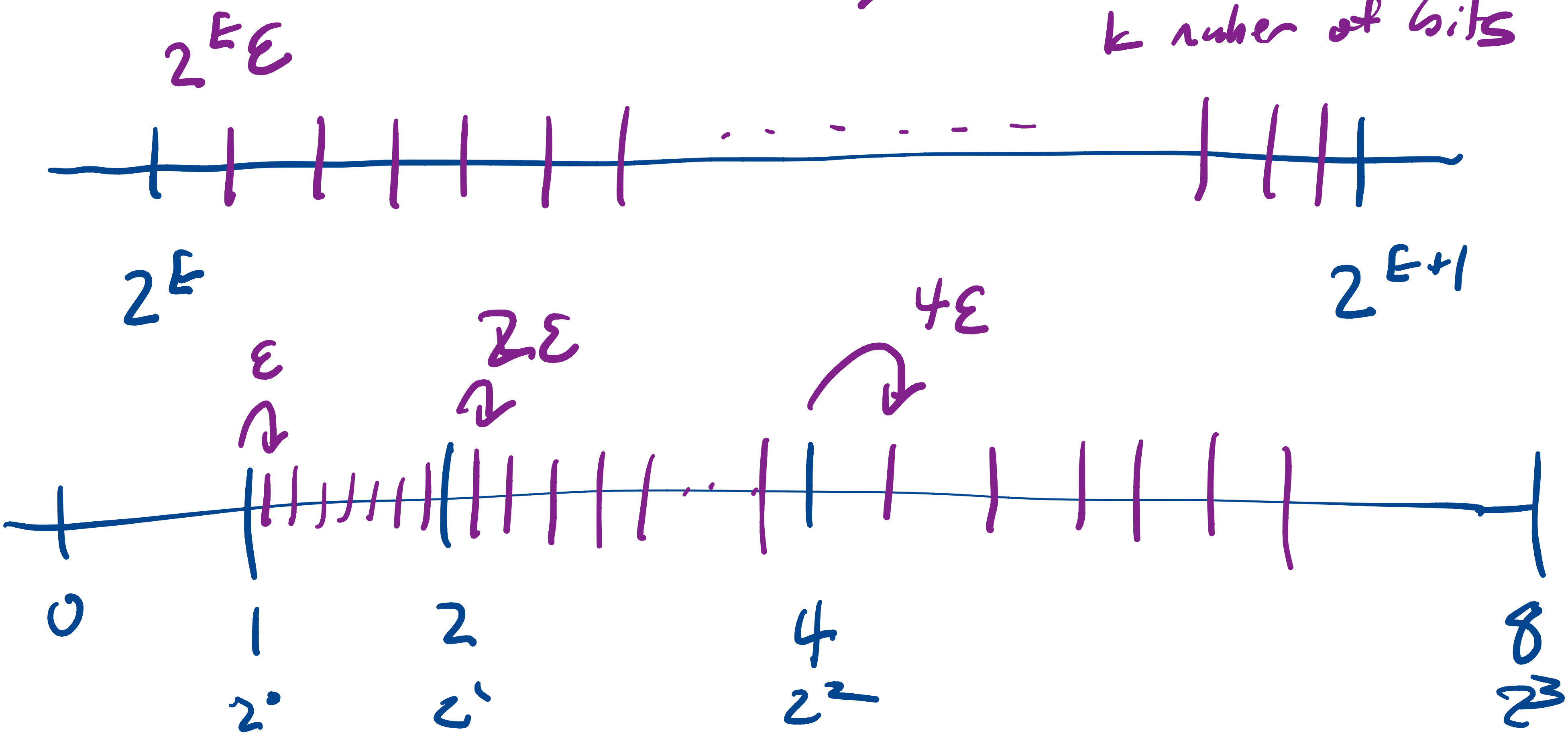
Suppose $2^E \le x < 2^{E+1}$.

$(-1)^s\, m \cdot 2^E$

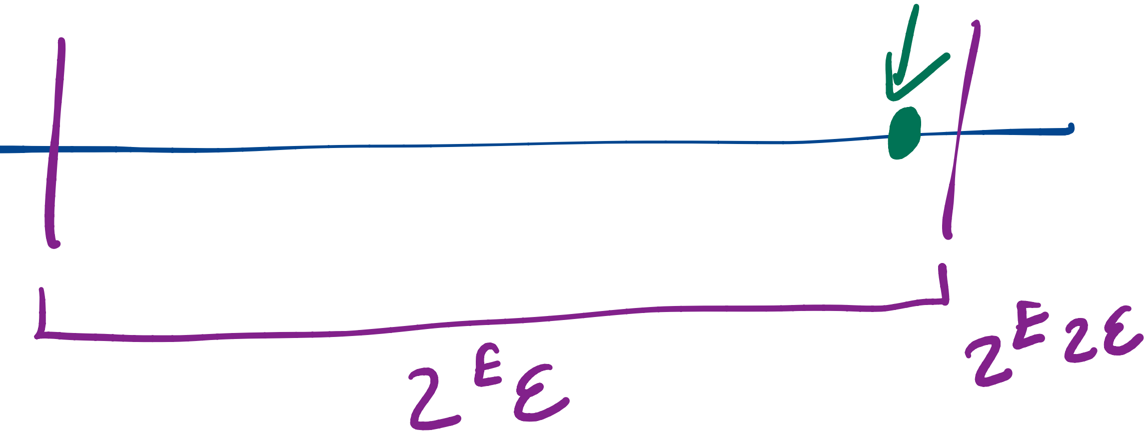
↳ # of choices 2^k

k number of bits

Number line:

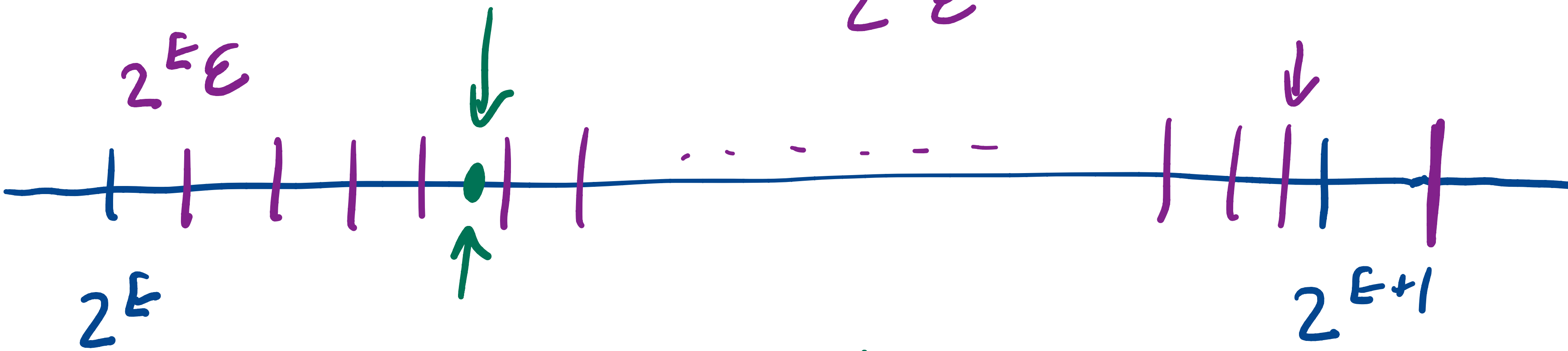


Rounding Error



Suppose $2^E \le x < 2^{E+1}$.

Number line:



all other modes → $|x - \mathrm{round}(x)| \le 2^E \varepsilon$

round to nearest → $|x - \mathrm{round}(x)| \le 2^E \varepsilon / 2$

Rounding Error

Suppose $2^E \le x < 2^{E+1}$.

Number line:

So: $|x - \text{round}(x)| \le \epsilon 2^E$.

(or $\epsilon/2 2^E$ for round to nearest)

Rounding Error

$$\frac{|x - \mathrm{round}(x)|}{|x|}$$

Relative error

Suppose $2^E \le x < 2^{E+1}$.

$$|x - \mathrm{round}(x)| \le 2^E \varepsilon$$

$$2^E \le |x|$$

$$\frac{1}{|x|} \le \frac{1}{2^E}$$

$$\frac{|x - \mathrm{round}(x)|}{|x|} \le \frac{2^E \varepsilon}{|x|} \le \frac{2^E \varepsilon}{2^E} = \varepsilon$$

Rounding Error

$$\frac{|x - \mathrm{round}(x)|}{|x|}$$

Suppose $2^E \leq x < 2^{E+1}$.

$$\frac{|x - \mathrm{round}(x)|}{|x|} \leq \varepsilon$$

$\left(\leq \frac{\varepsilon}{2} \right)$ for round to nearest

Rounding Error

$$\frac{|x - \mathrm{round}(x)|}{|x|}$$

Suppose $2^E \leq x < 2^{E+1}$.

$1/|x| \leq 2^{-E}$

Rounding Error

$$\frac{|x - \mathrm{round}(x)|}{|x|}$$

Suppose $2^E \le x < 2^{E+1}$.

$1/|x| \le 2^{-E}$

$$\frac{|x - \mathrm{round}(x)|}{|x|} \le \epsilon 2^E 2^{-E} = \epsilon$$

(or $\epsilon/2$ for round to nearest)

IEEE 754 arithmetic

The result of a floating point operation $(+, -, \cdot, /)$ is the correctly rounded value of the exact result.

$$x \oplus y := \mathrm{round}(x + y) = x + y + \mathrm{error}$$

$$\varepsilon \geq \frac{|\mathrm{error}|}{|x+y|} = \frac{|\mathrm{round}(x+y) - x+y|}{|x+y|}$$

$$|\mathrm{round}(x+y) - (x+y)| \leq \varepsilon\,|x+y|$$

IEEE 754 arithmetic

$\ominus \quad \otimes \quad \oslash \qquad x \otimes y = (x \cdot y)(1+\delta)$

The result of a floating point operation $(+, -, \cdot, /)$ is the correctly rounded value of the exact result.

$$x \oplus y := \underbrace{\text{round}(x+y)}_{} = \overbrace{x + y + \text{error}}^{(x+y)(1+\delta)}$$

$$x \oplus y = (x+y)(1+\delta) \quad \text{for some number } \delta \text{ with } |\delta| < \varepsilon/2$$

$$= (x+y) + \underbrace{(x+y)\delta}_{\text{error}}$$

$$\frac{\text{error}}{x+y} = \delta$$

IEEE 754 arithmetic

The result of a floating point operation $(+, -, \cdot, /)$ is the correctly rounded value of the exact result.

$$x \oplus y := \mathrm{round}(x + y) = x + y + \mathrm{error}$$

Similar operations: $\ominus, \otimes, \oslash$.

Rounding Isn't Easy

Compute $1.0000_2 - 0.00001010_2$ versus $1.0000_2 - 0.00001000_2$ under round to nearest.

$$\begin{array}{r} 1.0000 \\ -\ 0.0000 \\ \hline 0.1111111 \\ \hline 0.0000001 \\ \hline 1.0000000 \end{array}$$

$$\begin{array}{r} 0.1111 \\ 0.0001 \\ \hline 1.0000 \end{array}$$

$$\begin{array}{r} 0.11111100 \\ -\ 0.00000010 \\ \hline 0.11111010 \end{array}$$

0.11111

1.1111×2^{-1}

vs. 1.0000×2^{0}

Rounding Isn't Easy

Compute $1.0000_2 - 0.00001010_2$ versus $1.0000_2 - 0.00001000_2$ under round to nearest.

Requires extra bits (2 suffice for most: guard bits). Special cases requrie more (sticky bit for flag).