Floating Point Representation

$$\left(0.\underbrace{d_1 d_2 d_3 \cdots\cdots d_m}_{\text{mantissa bit / fraction}}\right)_{\beta} \times \beta^{e}$$

exponent ← e

β → Base

(1) Convention 1 (Standard Form)

$$\pm\left(0.\underbrace{d_1}_{d_1=1} d_2 d_3 \cdots\cdots d_m\right)_{\beta} \times \beta^{e}$$

(2) Convention 2 (Normalized Form)

$$\pm\left(1.\underbrace{d_1}_{d_1=[1,\,0]} d_2 d_3 \cdots\cdots d_m\right)_{\beta} \times \beta^{e}$$

(3) Convention 3 (Denormalized Form)

$$\pm\left(0.1\,\underbrace{d_1}_{d_1=[1,\,0]} d_2 d_3 \cdots\cdots d_m\right)_{\beta} \times \beta^{e}$$

Ex 1. Find the Largest Positive Numer (Highest Number) using m = 2

$\beta = 2.\quad e_{min} = -1 \quad \& \quad e_{max} = 2$

(1) Convention 1

$$(0.11)_2 \times 2^{2}$$

(2) Normalized

$$(1.11)_2 \times 2^{2}$$

(3) Denormalized

$$(0.111)_2 \times 2^{2}$$

Ex 2. Find the Lowest Positive Numer using m = 2

$\beta = 2.\quad e_{min} = -1 \quad \& \quad e_{max} = 2$

(1) Convention 1

$$(0.10)_2 \times 2^{-1}$$

(2) Normalized

$$(1.00)_2 \times 2^{-1}$$

(3) Denormalized

$$(0.100)_2 \times 2^{-1}$$

Ex 3. Find the Lowest Negative Numer using m = 2

$\beta = 2.\quad e_{min} = -1 \quad \& \quad e_{max} = 2$

(1) Convention 1

$$-(0.10)_2 \times 2^{-1}$$

(2) Normalized

$$-(1.00)_2 \times 2^{-1}$$

(3) Denormalized

$$-(0.100)_2 \times 2^{-1}$$

Ex 4. Find the Largest Negative Numer (Lowest Number) using m = 2

$\beta = 2.\quad e_{min} = -1 \quad \& \quad e_{max} = 2$

(1) Convention 1

$$-(0.11)_2 \times 2^{2}$$

(2) Normalized

$$-(1.11)_2 \times 2^{2}$$

(3) Denormalized

$$-(0.111)_2 \times 2^{2}$$

Ex 5. Find the Number of Non-Negative Combination using m = 2

$\beta = 2.\quad e_{min} = -1 \quad \& \quad e_{max} = 2$

$$e \rightarrow \begin{bmatrix} -1 \\ 0 \\ 1 \\ 2 \end{bmatrix} \rightarrow e^{num} = 4$$

(1) Convention 1

$0.1\,\_ \rightarrow 2^{1} \times 4 \qquad 2^{m-1} \times e^{num}$

(2) Normalized

$1.\,\_\,\_ \rightarrow 2^{2} \times 4 \qquad 2^{m} \times e^{num}$

(3) Denormalized

$0.1\,\_\,\_ \rightarrow 2^{2} \times 4 \qquad 2^{m} \times e^{num}$

For negative and positive combination, we have to muliply 2 with each value.

Decimal Shift

Shifting Decimal Left by m-> power/exponent INCREASE by m

Shifting Decimal Right by m-> power/exponent DECREASE by m

IEEE Format (64 bit)

1	11	52
sign	exponent ↓ 2^{11}	fraction/mantissa

Only positive : $0 \rightarrow 2047$

positive + negative: $-1022 \rightarrow 0 \rightarrow 1025$

Highest exponent to represent infinite -> 2^{1025}

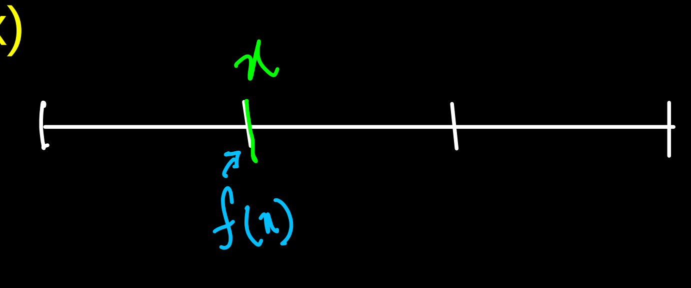
Lowest exponent to represent zero -> 2^{-1022}

Rounding & Error

Actual Value --> x

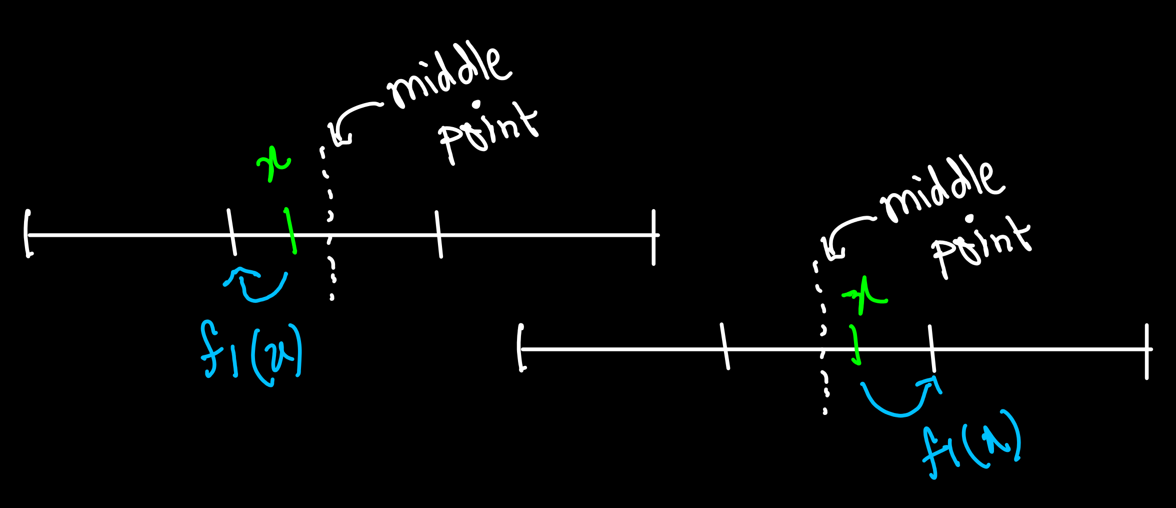
Floating Point Representation (F.P.R.) --> $fl(x)$

(1) x = fl(x)



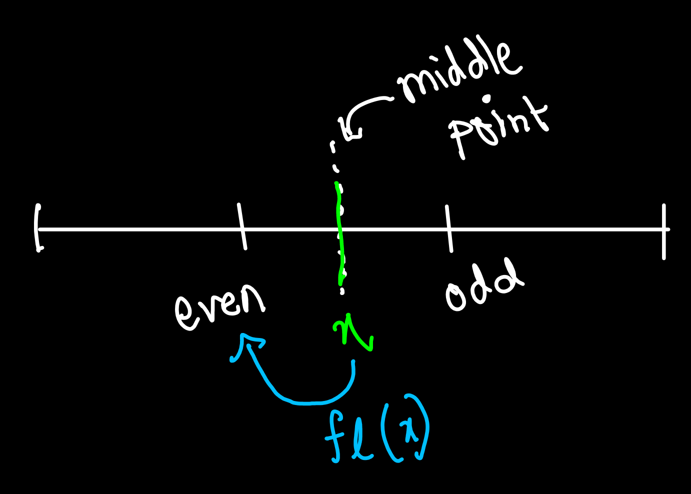
Since x = fl(x), so x is mapped fl(x). No Error.

(2) x not equals to fl(x)



x is mapped to the closest fl(x). Some Error.

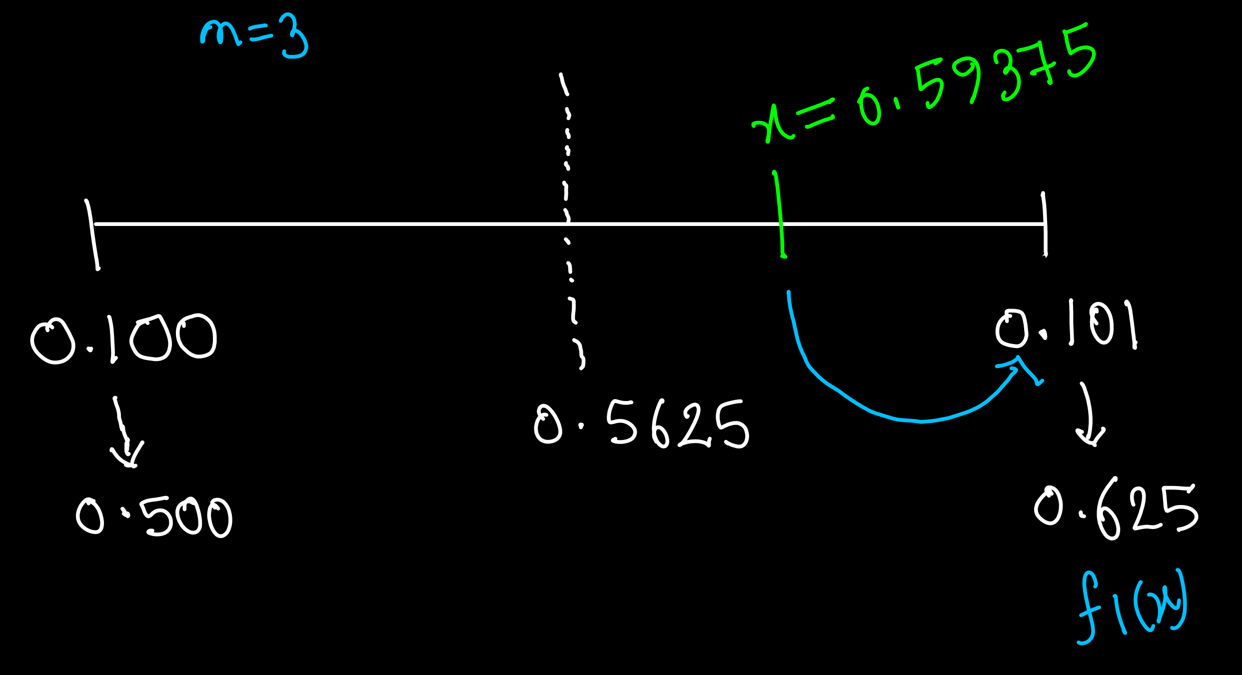
(3) x is on the middle point of two adjacent floating point



if fl(x) is exactly on the middle
of two floating point numbers,
then x is mapped to the EVEN number.
NB: Even number in binary have LSB (Rightmost bit)

Ex. 1 Using Convention 1, $m=3$ & $\beta=2$. Round $(0.10011)_2$ to the appropiate f.p.r.

$\underbrace{0.100}_{m=3}11 \rightarrow 0.100$



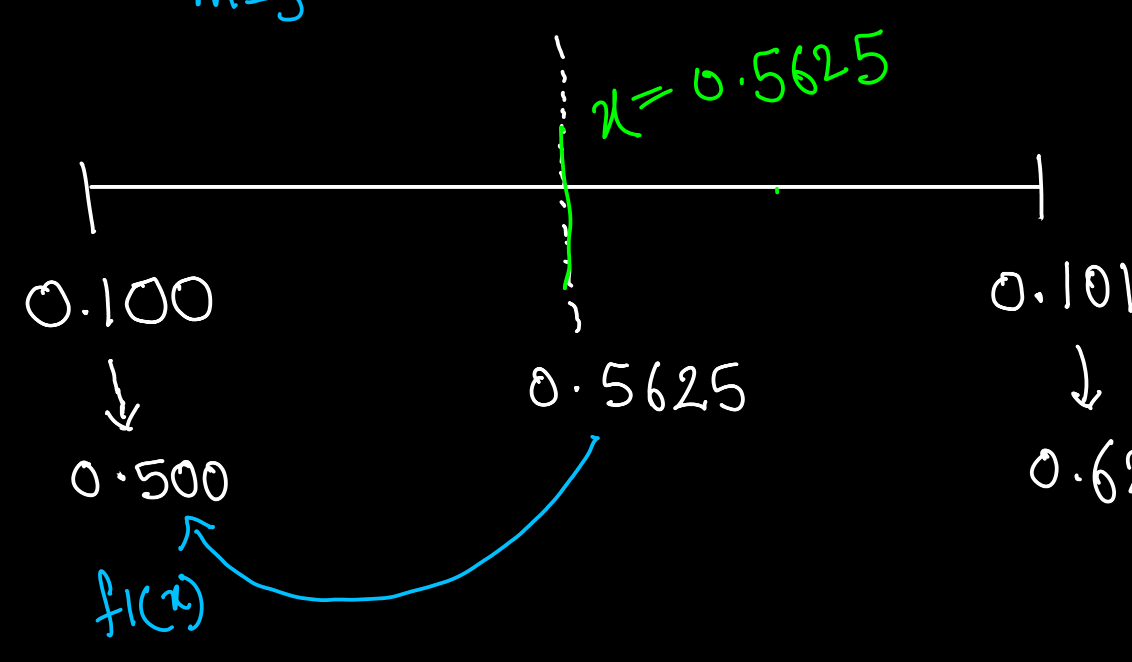
$$\text{Avg} = \frac{0.5 + 0.625}{2} = 0.5625$$

$$(0.10011)_2 \rightarrow (0.59375)_{10}$$

Floating point representation $= (0.101)_2$

Ex 2. Using Convention 1, m = 3 and $\beta=2$.
Round $(0.1001)_2$ to the appropiate f.p.r.

$\underbrace{0.100}_{m=3}1 \rightarrow 0.100$



$$\text{Avg} = \frac{0.5 + 0.625}{2} = 0.5625$$

$$(0.1001)_2 \rightarrow (0.5625)_{10}$$

Floating point representation $= (0.100)_2$

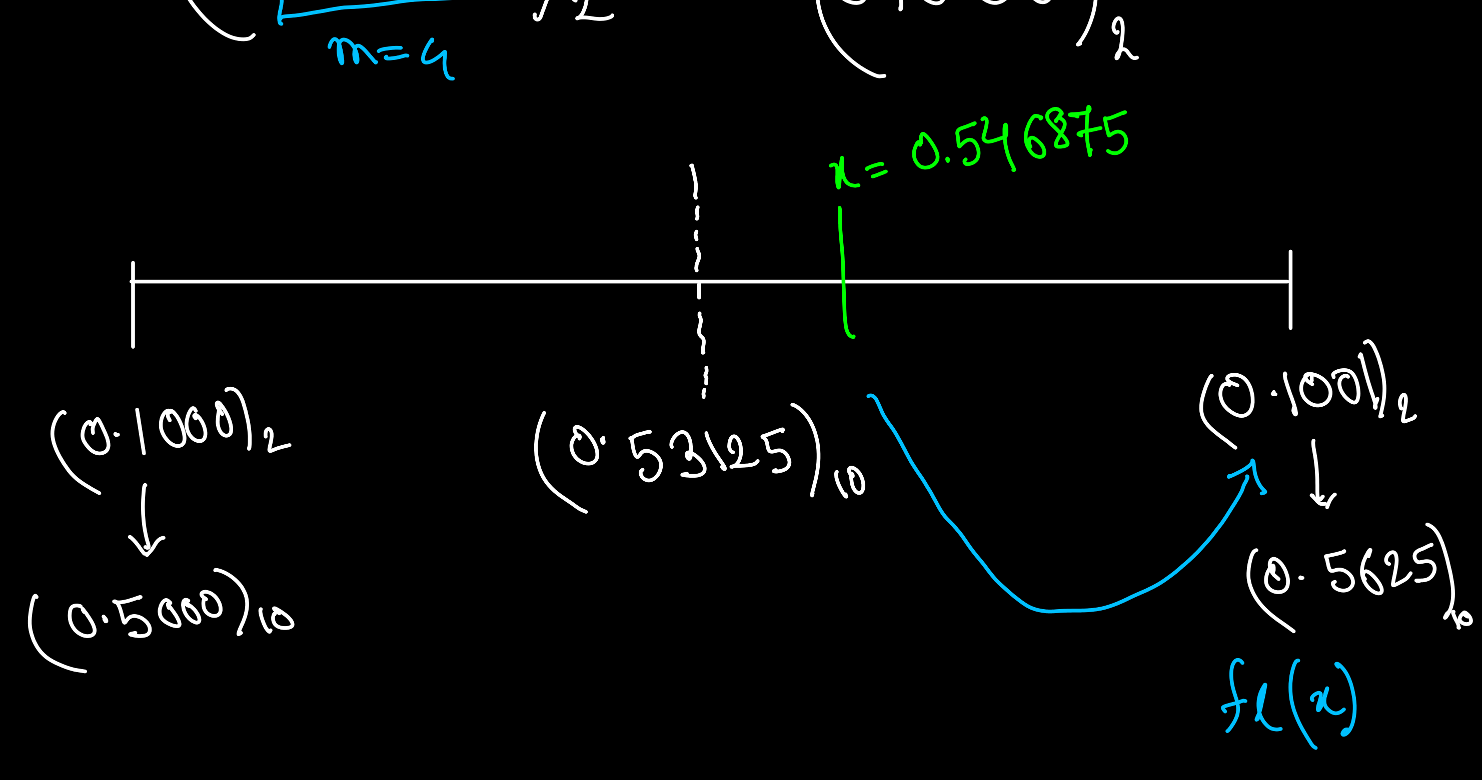
Absolute Rounding Error = $|fl(x) - (x)|$

Scale Invariant Rounding Error (Relative rounding Error), δ = $\dfrac{|fl(x) - (x)|}{|x|}$

Ex 3. Let's say we have x = $5/8$ and y = $7/8$. Using Convention 1, m = 4 and $\beta=2$. Find Fl (x.y). And find the relative rounding error.

$$x \cdot y = \frac{5}{8} \times \frac{7}{8} = \frac{35}{64} = (0.546875)_{10}$$
$$= (0.100011)_2$$

$(0.\underbrace{1000}_{m=4}11)_2 \rightarrow (0.1000)_2$



$$(0.100011)_2 \rightarrow (0.546875)_{10}$$

F.P.R $= (0.1001)_2$

$$\delta = \frac{|0.5625 - 0.546875|}{|0.546875|} \times 100$$

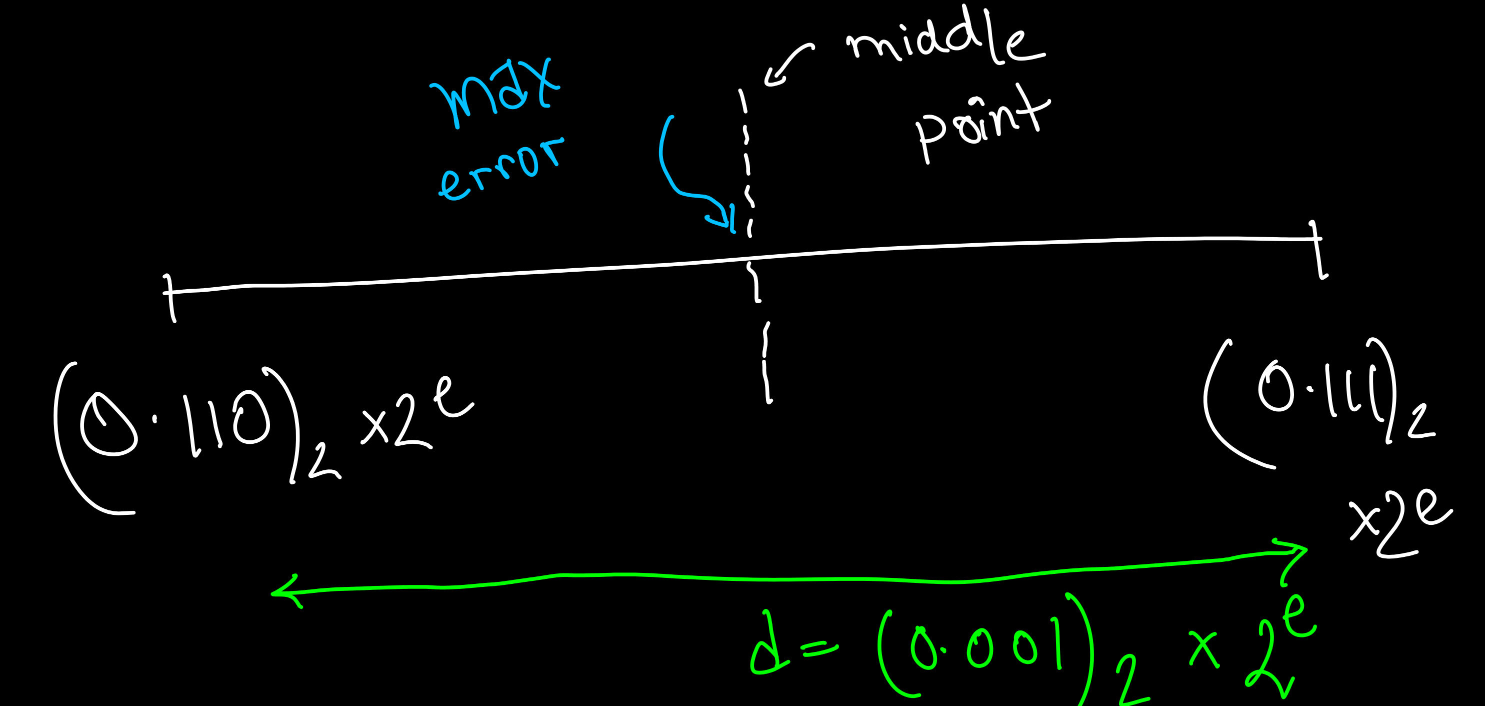
$$= 2.86\ \%$$

Machine Epsilon = Maximize scale invariant rounding error

$$\epsilon = \delta_{max} = \frac{|fl(x) - x| \uparrow}{|x| \downarrow}$$

Convention 1 $(0.d_1 d_2 d_3 \ldots)_\beta \times \beta^e$

$d_1 \ne 0$ ...



$m=3$: $1 \times 2^{-3} \times 2^e$

$m=4$ $[0.1100 \to 0.1101]$: $0.0001 \times 2^e = 1 \times 2^{-4} \times 2^e$

$m=2$ $[0.10 \to 0.11]$: $0.01 \times 2^e = 1 \times 2^{-2} \times 2^e$

$\beta^{-m}\ \beta^e$

$$\text{max error} = |fl(x) - x|_{max} = \frac{1}{2} \times \beta^{-m} \times \beta^e$$

Let's take $m=3$,

$$|x|_{min} = 0.100 \times \beta^e = 1 \times 10^{-1} \times \beta^e$$

irrespective of the value of m, $|x|_{min}$ will be same.

$$|x|_{min} = \beta^{-1} \beta^e$$

$$\epsilon = \frac{\frac{1}{2} \beta^{-m} \cancel{\beta^e}}{\beta^{-1} \cancel{\beta^e}}$$

$$\epsilon = \frac{1}{2} \beta^{1-m} \quad [\text{for convention } 1]$$

Convention 2 (Normalized) $(1.d_1 d_2 d_3)_\beta \times \beta^e$

$$\text{max error} = |fl(x) - x|_{max} = \frac{1}{2} \beta^{-m} \beta^e$$

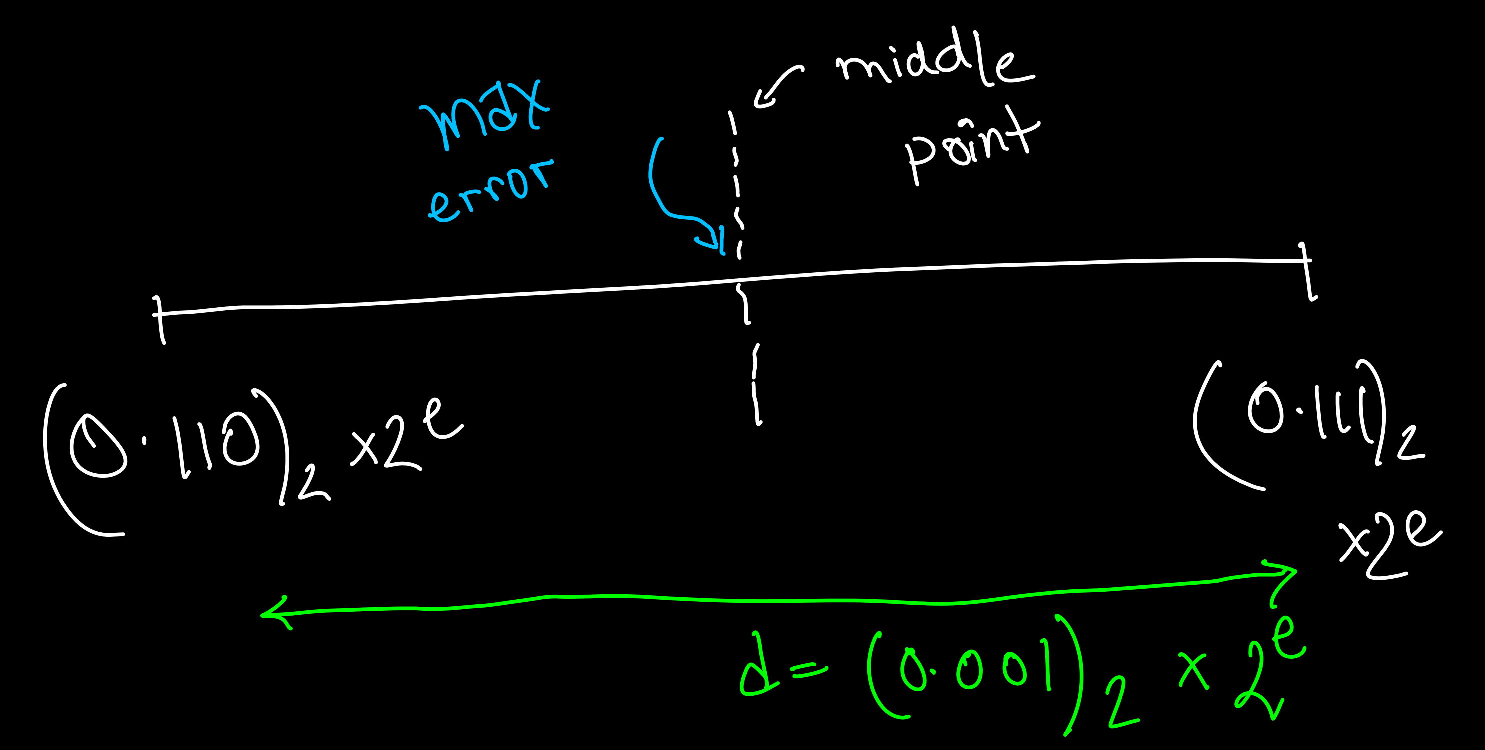
$$|x|_{min} = (1.0)_2 \times 2^e = 1.0 \times 2^0 \times 2^e$$

$$|x|_{min} = \beta^0 \beta^e = \beta^e$$

$$\epsilon = \frac{\frac{1}{2} \beta^{-m} \cancel{\beta^e}}{\cancel{\beta^e}}$$

$$\epsilon = \frac{1}{2} \beta^{-m} \quad [\text{for Normalized}]$$

Convention 3 (Denormalized) $(0.1 d_1 d_2 d_3)_\beta \times \beta^e$



$m=2$: $1 \times 2^{-3} \times 2^e$

$m=3$ $[0.1100 \to 0.1101]$: $0.0001 \times 2^e = 1 \times 2^{-4} \times 2^e$

$m=1$ $[0.10 \to 0.11]$: $0.01 \times 2^e = 1 \times 2^{-2} \times 2^e$

$$\text{max error} = |fl(x) - x|_{max} = \frac{1}{2} \times \beta^{-m-1} \times \beta^e$$

$$|x|_{min} = (0.10)_e \times 2^e = 1 \times 2^{-1} \times 2^e$$

$$|x|_{min} = \beta^{-1} \times \beta^e$$

$$\epsilon = \frac{\frac{1}{2} \times \beta^{-m-1} \times \cancel{\beta^e}}{\beta^{-1} \cancel{\beta^e}} = \frac{1}{2} \beta^{-m}$$

Machine Epsilon Formulae (Very Important)

Convention 1: $\frac{1}{2}\beta^{1-m}$

Convention 2(Normalized): $\frac{1}{2}\beta^{-m}$

Convention 3(DeNormalized): $\frac{1}{2}\beta^{-m}$

Loss Of Significance

$$fl(x) \neq x \qquad fl(y) \neq y$$

$$fl(x) = x + \delta_1 x$$

$$fl(y) = y + \delta_1 y$$

$$x \pm y$$

$$fl(x) \pm fl(y)$$

$$= x(1+\delta_1) \pm y(1+\delta_2)$$

$$= (x \pm y)\left(1 + \boxed{\frac{x\delta_1 \pm y\delta_2}{x \pm y}}\right)$$

scale invariant error

Ex 1.

$$x^2 - 56x + 1 = 0$$

$$x_1 = 28 + \sqrt{783} = 55.98$$

$$x_2 = 28 - \sqrt{783} = 0.01786$$

4 significant figures

$$\sqrt{783} = 27.98$$

$$x_1 = 28 + 27.98 = 55.98$$

$$x_1 = 28 - 27.98 = 0.0\underbrace{2000}_{4 \text{ s.f.}}$$