Fixed points or floating points are how numbers are stored / represented in a computer

Fixed point Representation

$$X = \pm (d_1 d_2 \cdots d_{k-1} \cdot d_k \cdots d_n)_\beta$$

where $d_1 \cdots d_n \in \{0, 1 \ldots \beta - 1\}$

Example

$X = +(10.1)_2$

$X = -(123.12)_{10}$

Evaluating fixed point numbers in base 10:

$$(10.1)_2 \rightarrow 1 \times 2^1 + 0 \times 2^0 + 1 \times 2^{-1}$$
$$= 2 + 0 + \frac{1}{2}$$
$$= (2.5)_{10}$$

Floating point Representation:

$$F = \{\pm \underbrace{(0 . d_1 d_2 \ldots d_n)}_{\text{fraction / mantissa}} {}_\beta \times \beta^{e}$$

$e \rightarrow$ exponent, $\beta \rightarrow$ base

where $\beta, d_i, e \in \mathbb{Z} \rightarrow$ integers

$0 \leq d_i \leq \beta - 1$

$e_{min} \leq e \leq e_{max}$

Scanned with CamScanner

Evaluating floating point numbers ~~in base 10~~

Examples

123.45

$= 12.345 \times 10^{1}$

$= 1.2345 \times 10^{2}$

$= 0.12345 \times 10^{3}$

1001.11

$= \underbrace{0.100111}_{\text{Fraction}} \times \underbrace{2}_{\text{Base}}^{4}$ — exponent

Conventions :

① $\pm (0.\boxed{d_1}\, d_2 \cdots d_m)_\beta\ \beta^{e}$

$\Downarrow$

$d_1 = 1$ always.

Example:

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

$m = 3 \quad e_{max} = 2$

$\Rightarrow$ highest possible FP number $= (0.111)_2 \times 2^{2}$.

② DeNormalized Form:

$\pm (1.d_1\, d_2 \cdots d_m)_\beta\ \beta^{e}$

$\downarrow$

$\left.\begin{matrix} d_1 = 1 \\ d_1 \neq 1 \end{matrix}\right\}$ both correct

Example:

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

$m = 3 \quad e_{max} = 2$

$\Rightarrow$ highest possible FP number $= (1.111)_2 \times 2^{2}$

Normalized

~~Denormalized~~ normalized Form

$$\pm (0.1\, d_1 d_2 \ldots d_m)_\beta\, \beta^e$$

Example

$\beta = 2$ $e_{min} = -1$

$m = 3$ $e_{max} = 2$

$$(0.1111)_2 \times 2^2$$

$\beta = 2$ $e_{min} = -1$ convention 1 [Standard form]

$m = 3$ $e_{max} = 2$

Ignoring the $\pm$ sign

Find the smallest and largest non-negative number

$$\underbrace{0.1}_{\text{Fixed}}\ \square\ \square \quad (d_1\ d_2\ d_3)$$

d_2	d_3
0	0
0	1
1	0
1	1

} 4 possible #

e: $-1, 0, 1, 2$ } 4 possible #

$\therefore$ Total possible # that can be represented $= 4 \times 4 = 16$

smallest #

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

$$= (1 \times 2^{-1}) \times 2^{-1}$$

$$= \frac{1}{4}$$

largest #

$$(0.111)_2 \times 2^2 \leftarrow e_{max}$$

$$= (1 \times 2^{-1} + 1 \times 2^{-2} + 1 \times 2^{-3}) \times 2^2$$

$$= \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8}\right) \times 2^2$$

$$= \frac{7}{2}$$

Considering sign bit

Smallest possible # $= -\frac{7}{2}$

$\boxed{e=-1}$

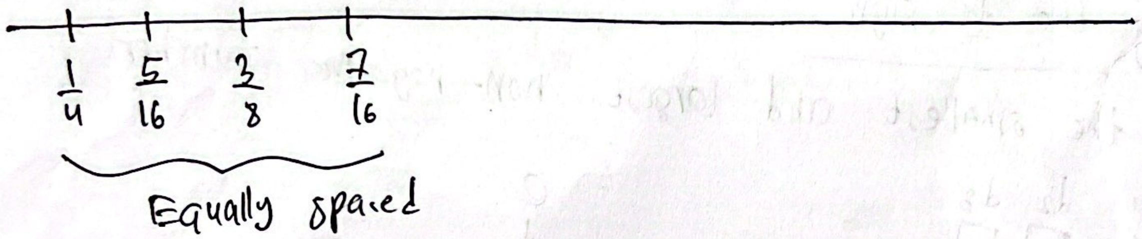
1^{st} smallest non-negative number $= (0.100) \times 2^{-1} = \frac{1}{4}$

2^{nd} " " " $= (0.101) \times 2^{-1} = \left(\frac{1}{4} + \frac{1}{8}\right) \times 2^{-1} = \frac{5}{16}$

3^{rd} " " " $= (0.110) \times 2^{-1} = \frac{3}{8}$

4^{th} " " " $= (0.111) \times 2^{-1} = \frac{7}{16}$

e = -1



$\frac{5}{16} - \frac{1}{4} = \frac{3}{8} - \frac{5}{16} = \frac{7}{16} - \frac{3}{8} = \frac{1}{16}$ [exponent constant = equally spaced]

$\boxed{e=0}$

$(0.100) \times 2^{0} = \frac{1}{2}$

$(0.101) \times 2^{0} = \frac{5}{8}$

⋮

$\boxed{e=1}$

$(0.100) \times 2^{1} = 1$

⋮

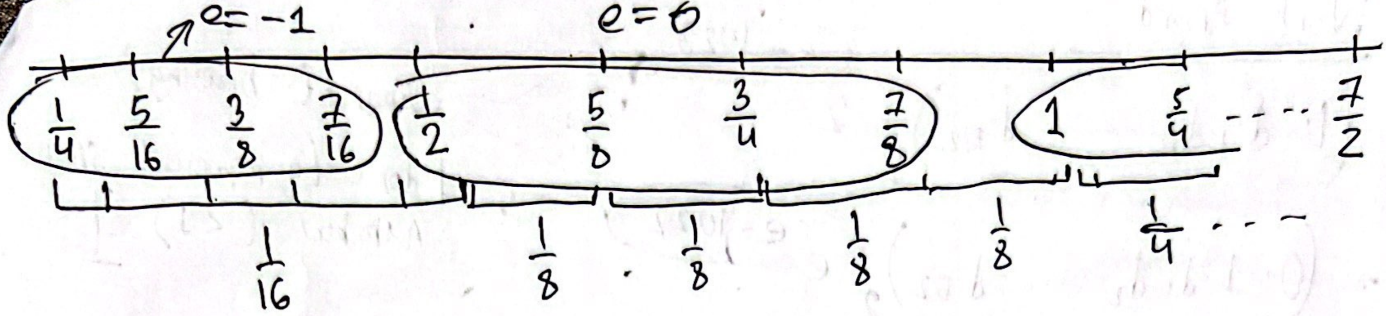
$\boxed{e=2}$

$(0.100) \times 2^{2} = 2$

⋮

$(0.111) \times 2^{2} = \frac{7}{2}$

-negative number line:



Important

If convention 1/denormalized form is used, we dont have 0 in our number system. Bcz $(0.1\,00\dots)$

we have 1 fixed here.

But 0 can be approximated using, for example, $(0.1000\dots)\,2^{-50} \approx 0$

IEEE standard for double precision

$\beta = 2$ 52-bits for fraction/mantissa

11-bits for exponent

1-bit for sign

Starting with Normalized Form:

$$(1.d_1d_2\cdots d_{52}) \times 2^{e}$$

min value of $e = 0$ [all 11 bits 0]

max value of $e = 2^{11}-1 = 2047$ [all 11 bits 1]

$e_{min} = 0$

$e_{max} = 2047$

largest possible non-neg num $= (1.11\cdots1)_2 \times 2^{2047}$

smallest " " " $= (1.0\cdots0)_2 \times 2^{0} = 1$ (not that small)

cannot express 0.001 for example.

Work Around

$$(1 \cdot d_1 d_2 \cdots d_{52})_2 \, 2^{e-1023}$$

$$= (0 \cdot 1 \, d_1 d_2 \cdots d_{52})_2 \, e^{e-1022}$$

exponent biasing.
[done to represent small numbers (<1)]

previously $e \in [0, 2047]$

Now, with exponent biasing $[-1022, 1025]$

Now, highest possible num $= (0 \cdot 1 \underbrace{11 \cdots 1}_{52 \text{ 1s}}) \times 2^{1025} \approx \infty$

smallest possible number $= (0 \cdot 1 0 \cdots 0) \times 2^{-1022} \approx 0$

In IEEE standard, 2 bits from exponent is reserved for ∞ and 0.

Highest possible exponent is used to store infinity

$$2^{1025} \longrightarrow \infty$$

smallest possible exponent is used to store 0

$$2^{-1022} \longrightarrow 0$$

Now, highest possible num, except inf $= (0.1\ 11 \cdots 1)_2 \times 2^{1024}$

$\approx 1.798 \times 10^{308}$

lowest " ", except 0 $= (0.1)_2 \times 2^{-1021}$

$\approx 2.225 \times 10^{-308}$.