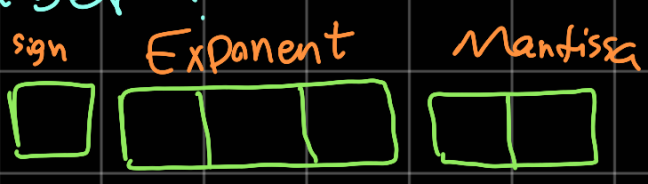
Remember our 6 bit Floating point number:



all the possible values was:

					Em
2^{-2}	0	0.0625	0.125	0.1875	
2^{-2}	0.25	0.3125	0.375	0.4375	$\rightarrow \Delta h = (2^{-2}) \times (2^{-2}) = \frac{1}{16}$
2^{-1}	0.5	0.625	0.75	0.875	$\rightarrow \Delta h = (2^{-2}) \times (2^{-1}) = \frac{1}{8}$
2^{0}	1	1.25	1.5	1.75	$\rightarrow \Delta h = (2^{-2}) \times (2^{0}) = \frac{1}{4}$
2^{1}	2	2.5	3	3.5	$\rightarrow \Delta h = (2^{-2}) \times (2^{1}) = \frac{1}{2}$
2^{2}	4	5	6	7	$\rightarrow \Delta h = (2^{-2}) \times (2^{2}) = 1$
2^{3}	8	10	12	14	$\rightarrow \Delta h = (2^{-2}) \times (2^{3}) = 2$

As you can see every $2^{n} \rightarrow 2^{n+1}$ sections are devided to 3 equal parts

I call each $[2^{n}, 2^{n+1})$ interval set(n)

So set(0) is $[1, 2)$

as you can see the Δh in each set

different, Δh_n is the Δh for set (n)

$$\Delta h_n = 2^{-t} \times 2^n$$

t: is the number of bits for mantissa (t = 2 for this example)

It is obvious that ∀

$$\left.\begin{array}{l} x \in \text{set}(n) \\ y < \Delta h_n \end{array}\right\} \rightarrow x + y = x$$

For example

$$\begin{cases} x = 8 \in \text{set}(3) \\ y = 1 < \Delta h_n = 2 \end{cases} \Rightarrow 8 + 1 = 8$$

In order <u>not to</u> have numerical extinction, we need to have $y > \Delta h_n = 2^{-t} \times 2^n$

Since $x \in \text{Set}(n)$, so $x = (1.\alpha_1 \cdots \alpha_t)_2 2^n$

$$\Rightarrow \frac{y}{x} > \frac{2^{-t} \times 2^y}{(1.\alpha_1 \cdots \alpha_t) 2^y} < \frac{2^{-t}}{(1.00\cdots0)_2} = 2^{-t}$$

Important :

if we have $\frac{y}{x} < 2^{-t} = \varepsilon_m$, then we will have $x + y = x$ (round to the lower number is assumed)

What Real happens ?

$8 \rightarrow (1.00)_2 \times 2^3$

$0.5 \rightarrow (1.00) \times 2^{-1}$ $\Rightarrow$ $8 + 0.5 = (1.00) \times 2^3 + (1.00) \times 2^{-1}$

$= \left(1.00 + 1.00 \times 2^{-4}\right) \times 2^3$

$1.00 \xrightarrow{\times 2^{-4}} 0.0001$

$1.00 + 1.00$

1.00

$= (1.00) \times 2^3 = 8$

$8 + 1 = 8$

$8 + 3 = 10$

$8 = (1.00)_2 \times 2^3$

$3 = (1.11)_2 \times 2^1$

$\rightarrow 8 + 3 = (1.00) \times 2^3 + (1.01) \times 2^1$

$= \left(1.00 + 1.01 \times 2^{-2}\right) \times 2^3$

$$\begin{array}{r} 1.00 \\ +\ 0.01 \\ \hline 1.01 \end{array} \quad = (1.01) \times 2^3 = 10$$

So in general :

$$x = (1.\alpha_1\alpha_2 \ldots \alpha_t) \times 2^{b_x} =$$

$$y : (1.\beta_1\beta_2 \ldots \beta_t) \times 2^{b_y}$$

$$\Rightarrow x + y = \left(1.\alpha_1\alpha_2 \ldots \alpha_t + \underbrace{1.\beta_1\beta_2 \ldots \beta_t \times 2^{b_y - b_x}}_{} \right) \times 2^{b_x}$$

if $|b_y - b_x| < t$

then we will loose all bits of $1.\beta_1 \ldots \beta_t$ because of the shift (devision)

$\Rightarrow$ So if $|b_y - b_x| < t$ then $x + y = x$

$$\frac{y}{x} = \frac{1.\beta_1 \cdots \beta_t \times 2^{b_y}}{1.\alpha_1 \cdots \alpha_t \times 2^{b_x}} < \frac{2^{b_y}}{2^{b_x}} = 2^{b_y - b_x}$$

$$\xrightarrow{|b_y - b_x| < -t} \quad \frac{y}{x} < 2^{b_y - b_x} < 2^{-t}$$

↓

So this expersssion is equivalent to $|b_y - b_x| < -t$.

$$\frac{y}{x} < 2^{-t} = \varepsilon_m \Rightarrow x + y = x$$