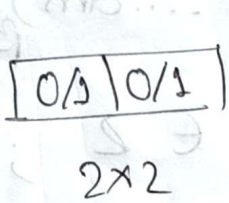
CH-1

Floating point arithmatic

Floats are more accurate and range is more.

In a 2bit system, without a sign bit there are 4 representations



∴ In a 64 bit system there are 2^{64} representations.

Fixed point numbers:

$$n = \pm (d_1 d_2 \ldots d_{k-1} \cdot d_k d_{k+1} \ldots d_n)_\beta$$

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

Floating Point Numbers:

$$F \subset R$$

Floating point numbers belong to subset of real numbers

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

where $\beta, d_i, e \in \mathbb{Z}$; $0 \leq d_i \leq \beta - 1$ and

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

we can now represent a large set of numbers but the numbers are not equally spaced. We can only represent a finite set of numbers

Conventions:

General / standard convention

1) $\pm (0.1 d_2 d_3 \ldots d_m)_\beta \times \beta^e$

In all the conventions the first number has to be non-zero so that the number can never be 0. This will ommit the confusion between +0 and -0

2) Normalized form:

$\pm (0.1 d_1 d_2 d_3 \ldots d_m)_\beta \times \beta^e$

3) Denormalized form:

$\pm (.1 d_1 d_2 d_3 \ldots d_m)_\beta \times \beta^e$

Example: If given, $m = 3$, $e_{min} = -1$ and $e_{max} = 2$

a) Find the highest possible number
b) Find the lowest possible number
c) Find the lowest possible non-negative number
d) Find the number of numbers that can be represented using the system
e) If the numbers are sequential or non-sequential.

a) Highest possible value $= +(0.111)_2 \times 2^2$

$= \left(2^{-1}+2^{-2}+2^{-3}\right) \times 4$

$= 3.5$

b) lowest possible value $= -(0.111)_2 \times 2^2$

$= -\left(2^{-1}+2^{-2}+2^{-3}\right) \times 4$

$= -3.5$

c) lowest possible non-negative value $= +(0.100)_2 \times 2^{-1}$

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

$= 0.25$

d) Total representation

$2 \times 2^2 \times 4 = 32$

↓ positive, negative

↓ 4 numbers in 2 slots

↳ possible e

e)

$0.111 \times 2^0 \longrightarrow 0.875$

$0.100 \times 2^1 \longrightarrow 1$ } difference 0.125

$0.101 \times 2^1 \longrightarrow 1.25$ } difference 0.25

So the numbers are non-sequential numbers

JEEE standard (1985) for double precision 64 bit

64bit distribution: 52 bits for fraction
11 bits for exponent
1 bit for signbit

$$\pm (0.1 d_1 d_2 d_3 \ldots d_{52})_2 \times 2^e$$

Now $2^{11} = 2048$. So possible values for e is from 0 to 2047.

But there is a problem to it.

In this state, lowest possible non-negative numbers is

$$(0.1000\ldots0)_2 \times 2^0 = (0.1)_2 = (0.5)_{10}$$

But we see we need numbers lower than this quite often. like 0.001 or 0.005 etc.

To achieve this the emin has to go lower. But that makes emax to go down. which decreases the range.

So JEEE found a suitable position where numbers can be of good range as well as of good precision

The process to achieve this is called exponent biasing
And for IEEE biasing is done by -1022.

$\therefore e_{min} = 0 - 1022 = -1022$

and $e_{max} = 2047 - 1022 = 1025$

But we need to reserve some special numbers.

like highest number $+(0.111\ldots1)_2 \times 2^{1025}$ is +infinity

again lowest number $-(0.111\ldots1)_2 \times 2^{1025}$ is -infinity

again non-negative lowest $+(0.1000\ldots0)_2 \times 2^{-1022}$ is zero.

So, actual ~~no~~ highest $= +(0.111\ldots1)_2 \times 2^{1024}$

and actual non-negative lowest $= +(0.1000\ldots0)_2 \times 2^{-1021}$

numbers greater than +infinity or lower than -infinity gives overflow.

number less than hypothetical zero and actual zero gives underflow.

$(19.25)_{10}$ convert to floating point number, where the general convention will be followed for $m=4$.

$$
\begin{array}{r|l}
2 & 19 \\
\hline
2 & 9 - 1 \\
\hline
2 & 4 - 1 \\
\hline
2 & 2 - 0 \\
\hline
2 & 1 - 0 \\
\hline
 & 0 - 1
\end{array}
\qquad
\begin{array}{r|l}
 & .25 \\
 & \times 2 \\
\hline
0 & .50 \\
 & \times 2 \\
\hline
1 & .00
\end{array}
$$

$19.25 \longrightarrow 10011.01$

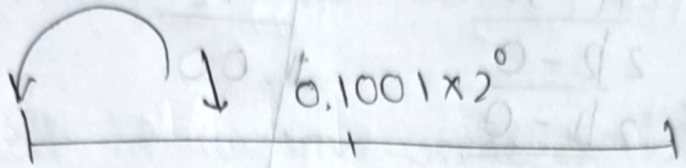
1st convention : 0.10011101×2^5

So here we need rounding.

Rounding:

Let, we have $(0.1000100)_2$

we have to store this in a system following standard convention for $m=3$. Where $e_{min} = -3$, $e_{max} = 3$.



0.100×2^0 0.101×2^0

$(2^{-1}) \times 1$ $(2^{-1}+2^{-3}) \times 2^0$

$= 0.5$ $= 5/8$

$$\text{mid} = \frac{0.5 + 5/8}{2} = \frac{9}{16} = \frac{8}{16} + \frac{1}{16}$$

$$= 2^{-1} + 2^{-4}$$

$$= (0.1001)_2$$

Since $0.1000100 < 0.1001$ so $fl(0.1000100)$

$= 0.100 \times 2^0$

if the number is lower than the mid then we round to the left.
if the number is greater than the mid then we round to the right.
if at the center then we go to the even number.
(the number that ends with 0 is an even number).

If we are given to calculate $fl(0.5 \times 0.7)$ then we actually need to calculate $fl(0.5) \times fl(0.7)$

Rounding error:

Let say we want to calculate the length of an object of 20cm. But we calculated 18cm.

$\therefore$ error = 2cm.

But this doesn't give the whole idea of error.

For example, if we were measuring a 2cm object but we calculated 0cm. Error is still 0cm but it is actually a massive error. This error is more significant than error found in the previous scenario.

for this we need to find scale invariant error δ,

$$= \frac{|\text{actual value} - \text{calculated value}|}{|\text{actual value}|}$$

So for rounding error we will consider scale invariant error.

$$\delta_1 = \frac{|2|}{|20|} \qquad \delta_2 = \frac{|2|}{|2|}$$

$$= 0.1 \qquad = 1$$

$\delta_2 > \delta_1$. So, Error_2 was higher

Machine Epsilon:

Highest possible δ is the Machine Epsilon, ϵ_M

for standard form,

$$\delta = \frac{|\text{actual value} - \text{calculated value}|}{|\text{actual value}|}$$

to maximise δ we need to increase the numerator and decrease denominator.

$\therefore x_{min}$, for $m = 3$,

$$= 0.100 \times 2^e$$

for $m = 4$; $x_{min} = 0.1000 \times 2^e$

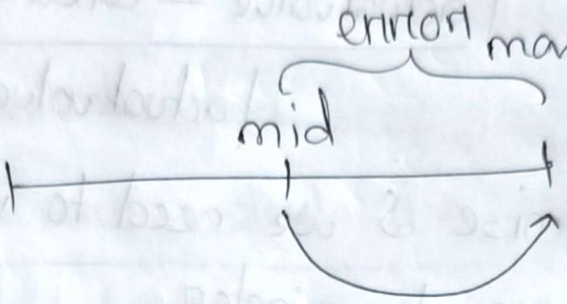
$\therefore$ for any m, $x_{min} = 0.1 \times 2^e$

$$= 2^{-1} \times 2^e$$

$$= \beta^{-1} \beta^e$$

Now maximum value for numerator comes when input is at the center and rounded to one of the ends

∴ we need to find the distance between midpoint and one of the ends.



for $m = 3$

left can be 0.100×2^e

and right " " 0.101×2^e

∴ difference = $0.001 \times 2^e = 2^{-3} \times 2^e$

for $m = 4$

left can be 0.1000×2^e

and right can be 0.1001×2^e

∴ difference = ~~0.000~~ $0.0001 \times 2^e = 2^{-4} \times 2^e$

∴ difference = $2^{-m} \times 2^e = \beta^{-m} \times \beta^e$

$\therefore$ numerator : $\frac{1}{2} \times \beta^{-m} \times \beta^{e}$ [$\because$ we actually need half the distance]

$$G_M = \delta_{max} = \frac{\frac{1}{2}\beta^{-m}\beta^{e}}{\beta^{-1}\beta^{e}} = \frac{1}{2}\beta^{1-m}$$

Normalized form,

For $m = 3$; $x_{min} = 0.1000 \times 2^{e}$

For $m = 4$; $x_{min} = 0.10000 \times 2^{e}$

$\therefore$ for any m; $x_{min} = 0.1 \times 2^{e}$

$= 2^{-1} \times 2^{e}$

$= \beta^{-1} \times \beta^{e}$

for $m = 3$

left can be 0.1000×2^{e}

right can be 0.1001×2^{e}

$\therefore$ difference $= 0.0001 \times 2^{e}$

$= 2^{-4} \times 2^{e}$

for $m = 4$

left can be 0.10000×2^{e}

right " " 0.10001×2^{e}

$\therefore$ difference $= 0.00001 \times 2^{e}$

$= 2^{-5} \times 2^{e}$

$\therefore$ numerator $= \frac{1}{2} \times 2^{-m-1} \times 2^{e} = \frac{1}{2} \times \beta^{-m-1} \times \beta^{e}$

$$\therefore G_M = \delta_{max} = \frac{\frac{1}{2}\beta^{-m-1} \times \beta^e}{\beta^{-1} \times \beta^e} = \frac{1}{2}\beta^{-m}$$

for denormalized form,

for $m=3$, $x_{min} = 1.000 \times 2^e$

for $m=4$, $x_{min} = 1.0000 \times 2^e$

$\therefore$ for any m, $x_{min} = 2^0 \times 2^e$

$= \beta^e$

for $m=3$

left can be 1.000×2^e

right can be 1.001×2^e

$\therefore$ difference $= 0.001 \times 2^e$

$= 2^{-3} \times 2^e$

for $m=4$

left can be 1.0000×2^e

right " " 1.0001×2^e

$\therefore$ difference $= 0.0001 \times 2^e$

$= 2^{-4} \times 2^e$

$\therefore$ for any m, numerator $= \frac{1}{2} \times 2^{-m} \times 2^e$

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

$$\therefore \epsilon_M = \delta_{max} = \frac{\frac{1}{2}\beta^{-m}\beta^e}{\beta^e} = \frac{1}{2}\beta^{-m}$$

loss of significance,

we know,

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

$$\therefore fl(y) = y + \delta_2 y$$

Now, $x \pm y = fl(x) \pm fl(y)$

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

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

$$= (x \pm y)\left(1 + \underbrace{\frac{x\delta_1 \pm y\delta_2}{x \pm y}}_{\text{scale invariant error}}\right)$$

for any two number $x \approx y$ if we try $x-y$ then the error becomes massive

because, for $x-y$; error $\left(1 + \frac{x\delta_1 - y\delta_2}{x-y}\right)$

and if $x-y \approx 0$; then error $\approx \infty$

So for subtraction of two very close number we get loss of significance.

example:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

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

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

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

Let take a system calculating upto 4 significant place,

$$\therefore x_1 = 28 + \sqrt{783} = 28 + 27.98 = 55.98$$

$$x_2 = 28 - \sqrt{783} = 28 - 27.98 = 0.0200$$

$$\therefore 0.02 \approx 0.01786$$

This happens because of loss of significance as we subtracted two very close number 28 and 27.98.

Solution:

We know $x^2 - (\alpha+\beta)x + \alpha\beta$

$$\alpha = x_1$$

$$\beta = \frac{1}{\alpha} = \frac{1}{x_1} = x_2$$

again, (Alternative solution)

equating $ax^2+bx+c = 0$,

$$x_1 \times x_2 = c/a$$

$$\therefore x_1 x_2 = 1$$

$$\therefore x_2 = \frac{1}{x_1}$$

Now $x_1 = 59.98$

$$\therefore x_2 = 0.1667.$$