Floating Point Representation :

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

fraction/mantisa/significand

β : Base, e : exponent.

The fraction bits are commonly referred to as 'mantisa'.

Convention 1 (Standard/General form)

$\pm (0.d_1 d_2 d_3 \ldots d_m)_\beta \times \beta^e$

$d_1 = 1$ always

Convention 2 (Normalized Form)

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

d_1 can be either 0 or 1.

Convention 3 (Denormalized Form)

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

d_1 can be either 1 or 0.

Example : $\beta = 2$, $e_{min} = -1$, $e_{max} = 2$, $m = 3$

Questions : ① Find largest possible/highest number for Convention 1, 2 and 3?

Convention 01 : $(0.111)_2 \times 2^2$

Convention 02 : $(1.111)_2 \times 2^2$

(Normalized)

Convention 03 : $(0.1111)_2 \times 2^2$

(Denormalized)

(ii) Find smallest possible (non-negative) number for Convention 01, 02 and 03?

Convention 01 : $(0.100)_2 \times 2^{-1}$

Convention 02 : $(1.000)_2 \times 2^{-1}$

(Normalized)

Convention 03 : $(0.1000)_2 \times 2^{-1}$

(Denormalized)

(iii) Find smallest possible number for convention 01, 02 and 03? [Always remember to consider sign bit].

Convention 01 : $-(0.111)_2 \times 2^{+2}$

Convention 02 : $-(1.111)_2 \times 2^{+2}$

(Normalized)

Convention 03 : $-(0.1111)_2 \times 2^{+2}$

(Denormalized)

** Trick is to find the highest number and give negative sign.

(iv) Possible combinations of numbers for the three conventions?

e = -1, 0, 1, 2

Convention 01 : 16 possible combinations.

$(0.1\_\_)_2 \times 2^e$

$d_2\ d_3 \longrightarrow$ 4 possible combinations: 00, 01, 10, 11

Convention 02 : 32 possible combinations.

$(1.\_\_\_)_2 \times 2^e$

$d_1\ d_2\ d_3 \longrightarrow$ 8 possible combinations: 000, 001, 010, 011, 101, 110, 100, 111

Convention 03 : 32 possible combinations

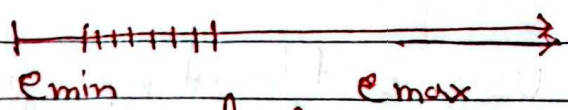
$(0.1\_\_\_)_2 \times 2^e$

$d_1\ d_2\ d_3 \longrightarrow$ 8 possible combinations

IEEE Standard (1985) for double precision (64-bit arithmetic)

$\beta = 2$, 52 bits for the fraction → precision
11 bits for the exponent → (bigger range) → e_{max}, e_{min}
1-bit for the sign.

If we have more bits in our fractional part, then we will have more numbers nearby (in the number line) so precision gets better.



0.110
0.111 → 0.1101

$(1.d_1d_2 \ldots d_{52})_2 \times 2^e$ We use binary numbers in computer and (normalized) form.

maximum value?
$e_{max} = 2^{11} = 2048$ [0, 2047]

minimum value?
$e_{min} = 0$.

largest possible number?

$(1.11\ldots1) \times 2^{2047}$

smallest possible number?
$(1.00\ldots0) \times 2^0 \approx 1$

We need smaller representation of number?

(0-1)

For this we need smaller values of e for which we need exponent biasing.

$(1.d_1d_2 \ldots d_{52})_2 \, 2^{e-1023}$ → exponent bias.

IEEE standard uses denormalized form

$= (0.1d_1d_2 \ldots d_{52})_2 \, 2^{e-1022}$

↳ $e \in [0, 2047]$ - previous range

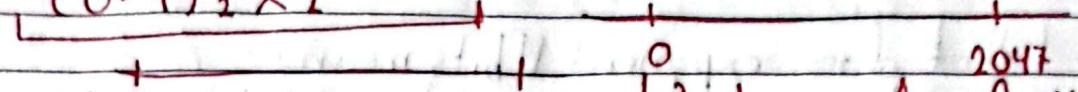
↳ biasing $(e - 1022) \in [-1022, 1025]$.

largest possible number

$(0.11\ldots1)_2 \times 2^{1025}$.

Smallest possible number

$(0.1)_2 \times 2^{-1022}$



We have reduced the range of largest possible number but we have incorporated very small range of numbers.

Exponent biasing is very important because we need numbers in the smaller range rather than in the larger range.

$2^{1025} \rightarrow \pm\infty$

$2^{-1022} \rightarrow \pm 0$ → 1025 is reserved for infinity.

largest number: $(0.11\ldots1)_2 \times 2^{1024} \approx 1.798 \times 10^{308}$.

smallest number: $(0.1)_2 \times 2^{-1022} \approx 2.225 \times 10^{-308}$.

underflow → 0

overflow → exception.

underflow is mapped to 0.

overflow is exception.

Rounding Error :

When we round a number we need to draw a number line. We need to do rounding because computer can read upto certain bits or significant figures.

m = 3 (3 bits after the decimal place)

$(0.1001)_2$

$(0.100)_2$ | $(0.101)_2$

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

$= \frac{1}{2}$ | $= \frac{1}{2} + \frac{1}{8} = \frac{8+2}{16}$

| $= \frac{10}{16}$, $\frac{5}{8}$

$\left(\frac{1}{2} + \frac{5}{8}\right) / 2 = \left(\frac{8+(2\times5)}{16}\right) \times 2$

$= \frac{18}{16} \times 2 = \frac{18}{8}$, $\frac{9}{8}$, $\frac{8}{16} + \frac{1}{16}$

$\frac{1}{2} + \frac{1}{2^4}$

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

$(0.1001)_2$

If the last bit is 0 of a binary number then it is even and if the last bit is 1 of a binary number then it is odd.

Suppose a number is given $(0.1000100)_2$

$= (0.100)_2$ x $(0.100)_2$ y $(0.101)_2$

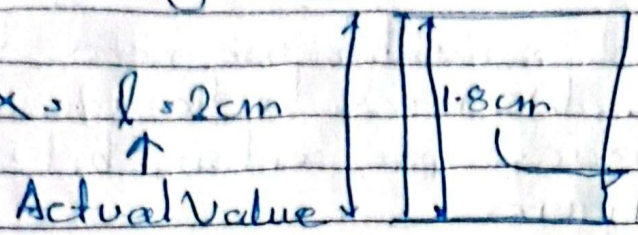
$(0.1000100)_2 = (0.100)_2$

$(0.100101000)_2 = (0.101)_2$

$(0.1001)_2 = (0.100)_2$

Since it is exactly the middle value it will get rounded to the [illegible]

Rounding Error :



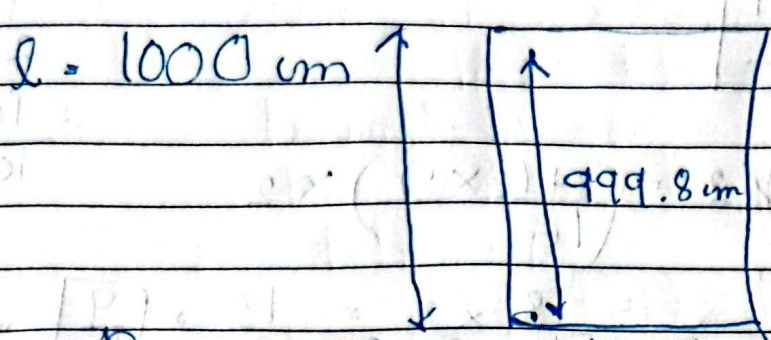
fl(x) Error/ Rounded Value

Error = (2 - 1.8) cm = 0.2 cm

x : Actual value

Error : $|fl(x) - x|$ fl(x) : Rounded value.

$= 0.2$ cm



Error : $|fl(x) - x| = |999.8 - 1000|$

$= 0.2$ cm.

It is a bit to difficult to interpret the impact using only Error : $|fl(x) - x|$ so we need to do Error : $\frac{|fl(x) - x|}{|x|}$.

Scale invariant rounding error : $\delta = \frac{|fl(x) - x|}{|x|}$

We deal with maximum scale invariant rounding error, Machine epsilon, ϵ

~~Max~~

δ_{max}/ϵ

Scale invariant Rounding Error $\delta = \frac{|fl(x)-x|}{|x|}_{max}$ $\frac{}{}_{min}$

Convention 01: $(0.d_1d_2d_3)_B \times B^e$

δ δ_{max}

$(0.100)_2 \times 2^e$ $(0.101)_2 \times 2^e$

$d = (0.001)_2 \times 2^e$

$m = 3$ $(0.001)_2 \times 2^e$

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

$B^{-m} B^e$ $(1 \times 2^{-m}) \times 2^e$

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

Convention (1)

$(0.100)_2 \times 2^e$

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

$|x|_m = B^{-1} B^e$

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

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

Two types of questions: $m = ?$ (Given) $B = ?$

Calculate Machine Epsilon value for mantissa convention 1?

Answer: $\epsilon = \frac{1}{2} B^{1-m}$

Minimum value of x for convention 01?

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

There is no exponent e in machine Epsilon ϵ

So the value of machine epsilon won't be affected by the value of exponent for convention 1

Convention 02 (Normalized form) m=3

$(1.d_1d_2d_3)_\beta \times \beta^e$

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

$|x|_{min} = (1.000)_2 \times 2^e$

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

$= (1 \times 2^0) \times 2^e$

$|x|_{min} = 2^e \times 2^0$

$|x|_{min} = \beta^0 \beta^e$ $\boxed{\beta^0 = 1}$

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

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

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

Convention 03 (Denormalized Form)

$(0.1d_1d_2d_3)_\beta \times \beta^e$

$(0.1d_1d_2d_3)_2 \times 2^e$

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

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

An extra bit 1 is added in this convention.

$\frac{1}{2}\beta^{-(m+1)}\beta^e$

$|x|_{min} = \frac{1}{2} B^{-m-1} B^e \Big/ B^{-1} B^e = \frac{1}{2} B^{-m}$

$(0.1)_2 \times 2^e$

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

$B^{-1} \times B^e$

Question: Base and mantisa will be given, calculate Machine Epsilon

Part-5 $B = 2$, $m = 3$, Calculate the value of $fl(x * y)$ (float)

We need to round the output.

$x = \frac{5}{8}$ $y = \frac{7}{8}$

① Convert to binary

$\frac{7}{8} = \frac{6}{8} + \frac{1}{8}$

$\frac{5}{8} = \frac{4}{8} + \frac{1}{8}$ $\quad \frac{4}{8} + \frac{2}{8} + \frac{1}{8}$

$= \frac{1}{2} + \frac{1}{2^3}$ $\quad \frac{1}{2} + \frac{1}{4} + \frac{1}{8}$

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

$= (0.101)_2$ $\quad = (0.111)_2$

actual value: $(x * y) = (0.101)_2 * (0.111)_2$

$= \frac{5}{8} \times \frac{7}{8} = \frac{35}{64}$ $\quad = \frac{32}{64} + \frac{2}{64} + \frac{1}{64}$

$= \frac{1}{2} + \frac{1}{32} + \frac{1}{64}$

We need to round the value since $m = 3$,

$= 2^{-1} + \frac{1}{2^5} + \frac{1}{2^6}$

$= 2^{-1} + 2^{-5} + 2^{-6}$

$(0.100)_2$ $\quad 0.1000$ $\quad (0.1001)_2$ $\quad 0.1001$

$(0.101)_2$ $\quad = (0.100011)_2$

Value In the left range $(0.100011)_2 = (0.100)_2$

rounded value $= \frac{1}{2}$.

$\frac{35}{64}$ / $\frac{1}{2}$. rounding error

loss of significance x will be rounded to fl(x)
y " " " " fl(y)

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

fl(x) / x (may not be equal)
fl(y) / y (")

rounded value — actual rounding error value is being incorporated.

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

$x \pm y$

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

$= x + \delta_1 x + y + \delta_2 y$

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

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

$x - y$

$\frac{x\delta_1 - y\delta_2}{x-y}$ ← approximately zero

$x = y$

→ rounded value will be ∞ infinity

Scale invariant error becomes infinitely large when we subtract two numbers. This phenoma is loss of significance. When I subtract two numbers and both the numbers are closer then in case of rounding or scale invariant error this denominator will be very small or approximately close to zero, the scale invariant error will be very high.

scale invariant error

Example-1:

$x^2 - 56x + 1 = 0$ $\quad -b \pm \sqrt{b^2 - 4ac} \over 2a$

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

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

→ 4sf - computer can only process.

$\sqrt{783} = 27.98$

$x_1 = 28 + 27.98 = 55.98$

$x_2 = 28 - 27.98 = 0.02000$ (4sf)

$0.01786 \neq 0.02000$ → loss of significance.

As x_2 numbers are very small, denominator will be very small so rounding error will be very high.

Solution: We will not subtract the numbers.

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

$\alpha = x_1$

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

5.01 5.02

average = 5.015 → 3sf

$fl\left(\frac{5.01 + 5.02}{2}\right) = fl\left(\frac{10.03}{2}\right)$

$= fl\left(\frac{10.0}{2}\right)$

= 5, Rounding error.