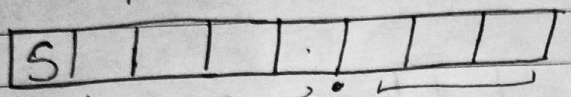
# Floating point Representation Lecture 12

Real No.
- Fixed point Notation
- Floating point Notation

* Fixed point Notation




$+(4.5)_{10}$ =

0 0 1 0 0 . 1 0 0 → in fixed point notation
2 4 H = in hex

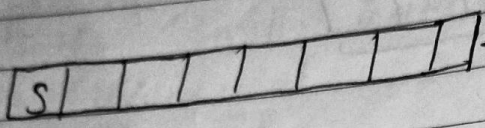
- $(6.25)_{10}$ = 0 0 1 1 0 . 0 1 0
  3 2 H = in hex
- $(-11.75)_{10}$ = 1 1 0 1 1 . 1 1 0
  D E H
- ~~Range~~ $(6.625)_{10}$ = 0 0 1 1 0 . 1 0 1
  3 5 H = in hex

- Range = 0 1111.1111 (+Ve) = $+(15.875)_{10}$
  1 1111.1111 (−Ve) = −15.875

We representing all real no. from −15.875 to 15.875 ✗

Stepup → 0.125

| | |
|---|---|
| .000 | .0 |
| .001 | .125 |
| .010 | .250 |
| .011 | .325 |
| .100 | .500 |
| ⋮ | ⋮ |




000... to 111...

$2^8 = 256$

+ve: 128 → 0 to 15.875

-ve: 128 → -0 to -15.875

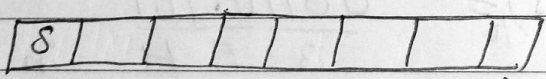
0000.000 → 111

⋮

1111.000 → 111

16 × 8 = 128

Disadvantage → Range will be very less




if we shift point this side range will decrease and accuracy will increase

vice ~~to~~ versa

# Floating Point Representation Part 2

* Floating point Notation

$363.4829 \times 10^0$

$36.34829 \times 10^1$

$3.634829 \times 10^2$

$.3634829 \times 10^3$ → Normalized notation

$363.4829$

$3634.829 \times 10^{-1}$

$36348.29 \times 10^{-2}$

Scientific notation

$a * b^e$ → exponent must be integer

a → real no.

b → base

Normalized notation

$m * b^e$ → Integer

b → base

$0.1 \le |m| < 1$

$3.6 \times 10^{-5}$ — Scientific

$.36 \times 10^{-4}$ — Normalized, Scientific

$4.98 \times 10^{-6}$ — Scientific

$23.4 \times 10^{4.7}$ → Neither normalized nor scientific

- Base 2

$101.11 \times 2^3$ → Scientific

$1.01 * 2^4$ → S

$.1 \times 2^3$ → N, S

$.01 \times 2^4$ → S

$.001 \times 2^{4.2}$ → X

* Normalized Binary
  .1 → it must be '1'

in other base
after point it not = to '0'

↓ ↓
0 ≠ 0

Ex:
- $.123 * 10^5$ = Normalize
- $.93 * 10^4$ = N —— "
- $.1 * 2^3$ = N —— "
- $.101 \times 2^5$ = N
- $.01 * 2^8$ = X

* Rule for Binary
  $a * b^e$
  ↓ ↳2

. ↓ should not be 0 after point

↓ Should be 0 before point

for mentisa

| S | E | M |

$.1 \leq |m| < 1$

while storing we'll not store the base because ∴ is obious i.e 2

format

| S | E | M |
|---|---|---|
| S | M | E |
| M | S | E |

| M | E | S |
|---|---|---|
| E | M | S |
| E | S | M |

| S | E | M |
|---|---|---|

⇒ 16 bit

1 — 8 — 7 bits

E → excess 128

Let number → $(3.5)_{10}$ → $(11.1)_2$

Convert this into normalized form

$(11.1)_2$ ⇒ $.111 * 2^2$

Let's store it according to sem

0 10000010 0 1110000

S — E — M

Sign(S) — 128+2 — 2 is exponent

excess 128 — $.111 \times 2^{②}$

Hex Representation

0100 0001 0111 0000

4170 H Ans

$\#(.25)_{10} = (.01$

$.1 * 2^{-1}$

0 01111101 0000000

3F80H Ans

$-(6.75)_{10}$ → $(110.11)_2$ ⇒ normalized $.11011 \times 2^3$

S — E — M

1 10000011 1011100

C1EEH Ans

- $-8.6 \rightarrow 1000.\overline{1001} \rightarrow 1.0001001\ldots \times 2^3$

  Mantissa

  S e M

  1 10000100 0001001

  C244H Ans

- 3FC0H

  0011 1111 1100 0000

  $+ \quad .1 * 2^{127-128} \Rightarrow .1 \times 2^{-1} = (.01)_2$

  $\Downarrow$

  $(.25)_{10}$ Ans

- 8CC0H

  1000 1001 1000 0000

  S E M

  $-ve \quad .1 * 2^{25-128} \Rightarrow -.1 * 2^{-103} = -1 * 2^{-104}$

  $\Downarrow$

  $-2^{104}$ Ans

* | S | M | E |

1 8 7 → excess 64

- $(1.5)_{10} \rightarrow (1.1)_2 \rightarrow .11 \times 2^1$

  0 1100000 0 1000001

  6041H 6041H Ans

* [E | S | M] → 22 bit = 32 bit
  ↓ ↓
  9 excess 256

o (13.875)10 → (1101.111)2 ⇒ .1101111 × 2^4

100000100 0 1101 111 000 -- | 32/4 = 8 digit

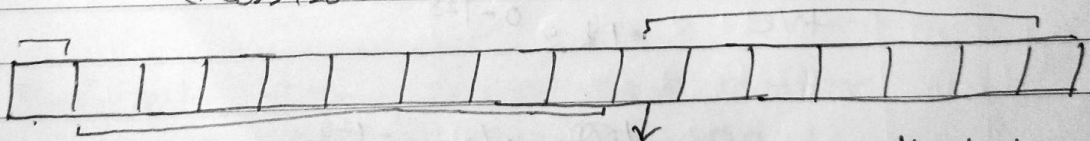
82378000H Ans
↓
5 + 3 = 8 digit

## * Floating point Representation Part 3

* Range
  - → Normalized → .1
  - → De normalize → .0

| S | E | M |
|---|---|---|
| 1 | 8 | 7 |

↘ excess 128



if 0 then Denormalized else Normaliz

*

P. T.O

* Normalized

- largest +ve number

0 11111111 . 11111111
S e M

$\cdot 1111111 * 2^{255-128}$

$\cdot 11111111 * 2^{127} \Rightarrow (1-2^{-7}) * 2^{127}$ Ans

$1111111 * 120\ 2^{120}$

$127 \times 2^{120}$ Ans

- Smallest −ve number

1 11111111 . 1111111
S e m

$- \cdot 1111111 * 2^{255-128}$

$- \cdot 1111111 * 2^{127} \Rightarrow (1-2^{-7}) * 2^{127}$

$- 127 * 120^{120}$

- Smallest +ve normalized number :−

0 00000000 . 1000000

↓

+ve $\cdot 1 * 2^{0-128}$

$\cdot 1 * 2^{-128}$

$1 * 2^{-129}$

$2^{-129}$ Ans

- largest -ve & normalized number

1 00000000 1000000

$-\ .1 * 2^{0-128}$

$-\ .1 * 2^{-128}$

$-\ 2^{-129}$ Ans

* Denormalized number
  - largest +ve denormalized number

0 11111111 0111111

$+\ .0111111 * 2^{255-128} \Rightarrow 111111 * 2^{126}$

$.0111111 * 2^{127}$

~~111111 * 1/2~~

$111111 * 2^{120}$

$63 * 2^{120}$ Ans

- Smallest -ve denormalized number

~~1.0001~~ $-63 *_2 120$ Ans

- Smallest +ve normalized number

0 00000000 0000001

S e M

$+\ .00000001 * 2^{0-128}$

$2^{-7} * 2^{-128}$

$2^{-135}$ Ans

$.1 \rightarrow 2^{-1}$

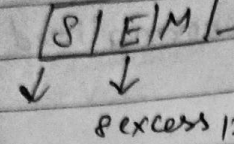
$.01 \rightarrow 2^{-2}$

- largest −ve denormalized number

$-2^{-135}$ Ans

Note - By default number will be in normalized form because there are very less chances of denormalized number

Ex




$(3.5)_{10}$ → Convert into normalized form

$11.1$ ⇒ Binary

$.111 * 2^2$ ⇒ Normalized number

Note Rule ⇒ After point number should be '0' in denormalized form

$.111 \times 2^2$ ⇒ Normalized number

many form of denormalized number

- $.0111 * 2^3$ ⇒ denormalized number
- $.00111 * 2^4$ —— "
- $.000111 * 2^5$ —— "
- $.0000111 * 2^6$ —— "

$.0111 * 2^3$ ⇒ 0 1000 0011 0111 0000

4 1 B 0 H

$\underset{S}{0}\ \underset{E}{1000\ 001}\ \underset{M}{0111\ 000} \Rightarrow 41B8H$

Mantissa's first bit is '0' so it is denormalized number

$\circ\ 0111 * 2^{128+3-128}$

$\circ\ \underbrace{0111 \times 2^{3}} \Rightarrow (11.1)_2 = (3.5)_{10}$

↑ Normalized this no. to get the ans

$\circ\ .111 * 2^2 \Rightarrow$ Normalized form

$\circ\ .00111 * 2^4$

$0111 . 00111 * 2^{128+4-128}$

$\underbrace{.00111 \times 2^{4}} = (0011.1)_2 = (3.5)_{10}$

↑ Normalized this number to get ans ↑

$.111 * 2^2 \Rightarrow$ Normalized form

* S | E | M → 20bit

↓ 16bit ↓ 11bit excess 1024

largest normalized number

Smallest $(1-2^{-20}) * 2^{1023}$

$0.1 * 2^{-127} \Rightarrow 2^{-128} \Rightarrow .5 \times 2^{-127}$

Largest positive de normalized number

~~(illegible)~~ $(1-2^{-19}) * 2^{1022}$

Smallest $= 2^{-7} * 2^{-127} \Rightarrow 2^{-134}$

- [S|E|M] → 21 bit (1 bit, 10 bit, excess 511)

+ve normalized no.

$(1-2^{-21}) * 2^{512}$

Smallest $\overset{+Ve}{0.1 * 2^{-511}} \rightarrow 2^{-512} \Rightarrow .5 * 2^{-511}$

Largest +Ve denormalized number

$1-2^{-20} * 2^{-511} \Rightarrow 2^{-532}$

Smallest

$2^{-21} * 2^{-511} \rightarrow 2^{-532}$

- [S|E|M] → 52 bit (1 bit, excess 1023 (11 bits))

* ~~[illegible]~~ ~~[illegible]~~

$(1-2^{-52}) * 2^{2047-1023}$

$.1 * 2^{-1023} \Rightarrow 2^{-1024} \rightarrow 0.5 * 2^{-1023}$

Largest +ve denormalized no.

$(1-2^{-51}) * 2^{1023}$

SPD → $2^{-52} * 2^{-1023} = 2^{-1075}$ Ans

- S | E | M → 13 bits
  - S: 1 bit
  - E: ex 32 (6 bits)
  - M: 6 bits

Largest +ve normalized no.

$1-2^{-32} * 2^{31}$

SPN → $.1 * 2^{-31} \Rightarrow .1 * 2^{-32} \Rightarrow 0.5 * 2^{-31}$

Smallest +ve denormalized number

$2^{-13} * 2^{-32}$

Largest:

$(1-2^{-12}) * 2^{30}$ Ans