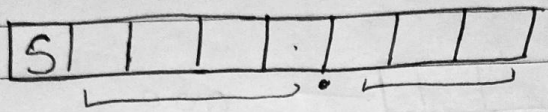
Floating point Representation Lecture 12

Real No.
- Fixed point Notation
- Floating point Notation

* Fixed point Notation



$+(4.5)_{10}$ =

0010 0.100 → in fixed point notation
2 4H = in hex

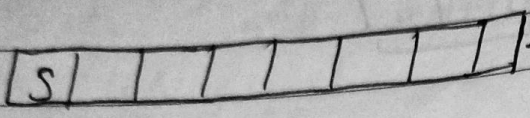
- $(6.25)_{10}$ = 00110.010
  3 2H = in hex
- $(-11.75)_{10}$ = 11011.110
  D EH
- ~~Range~~ $(6.625)_{10}$ = 00110.101
  3 5H = in hex

	+Ve	-Ve
Range =	0 1111.1111	11111.1111
	$+(15.875)_{10}$	-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 → 1111.111

⋮

$16 \times 8 = 128$

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

363.4829×10^0

36.34829×10^1

3.634829×10^2

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

363.4829

3634.829×10^{-1}

36348.29×10^{-2}

Scientific notation

$a * b^e$ → exponent must be integer

a → real no.

b → base

Normalized notation

$m * b^e$ → Integer

b → base

$0.1 \leq |m| < 1$

3.6×10^{-5} — Scientific

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

4.98×10^{-6} — Scientific

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

- Base 2

101.11×2^3 → 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

1 . 1 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 —— "
- $.01 * 2^3$ = X

* Rule for Binary

$a * b^e$

b_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

excess 128

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

Convert thin into normalized form

$(11.1)_2 \Rightarrow .111 * 2^2$

Let's store it according to sem

0 10000010 1110000

S E M

Sign(s) 128+2

excess 128 2 is exponent

$.111 \times 2^{②}$

$\#(.25)_{10} = (.01)_2$

$.1 * 2^{-1}$

0 01111111 10000000

3F C0H Ans

Hex Representation

0100 0001 0111 0000

4170H Ans

$-(6.75)_{10} \rightarrow (110.11)_2 \Rightarrow$ Normalized $.11011 \times 2^3$

S E M

1 10000011 10110000

C1 E CH Ans

- $-8.6 \rightarrow 1000.1001 \rightarrow 1.0001001 \times 2^3$

  Mantissa

  S e M

  1 10000100 0001001

  C244H Ans

- 3FC0H

  0011 1111 1100 0000

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

  $\Downarrow$

  $(0.25)_{10}$ Ans

- 8CE0H

  1000 1100 1110 0000

  S E M

  -Ve $\cdot 1 * 2^{25-128} \Rightarrow -\cdot 1 * 2^{-103} = -1 * 2^{-104}$

  $\Downarrow$

  -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 01000001

  ~~6041H~~ 6041H Ans

* [E | S | M] → 22 bit = 32 bit

↓ ↓

9 excess 256

- $(13.875)_{10} \rightarrow (1101.111)_2 \Rightarrow .1101111 \times 2^4$

100000100 0 1101 111 000 ... | $\frac{32}{4}$ = 8 digit

82378000H Ans

↓

5 + 3 = 8 digit