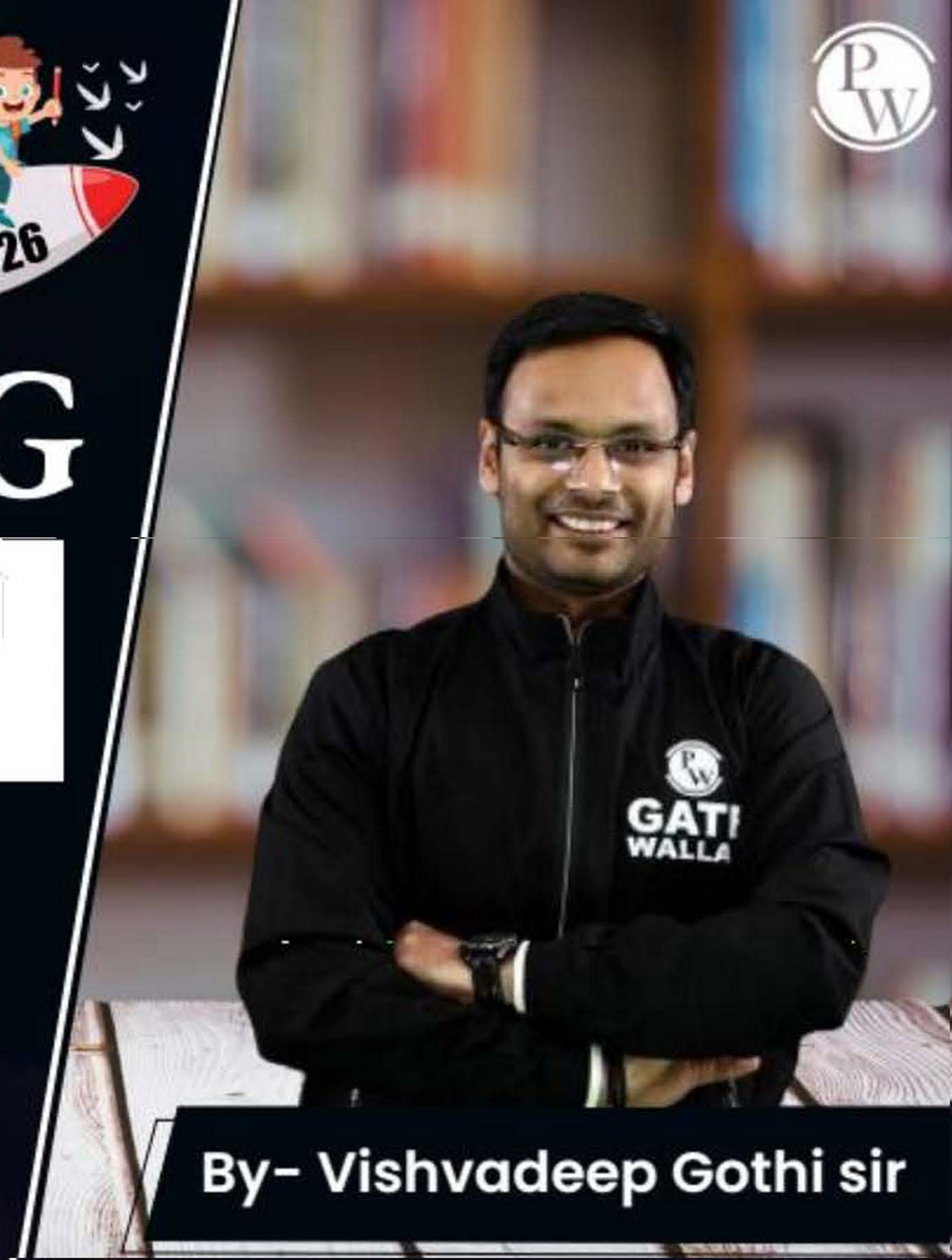
CS & IT ENGINEERING

COMPUTER ORGANIZATION AND ARCHITECTURE

Floating Point Representation

One Shot



By- Vishvadeep Gothi sir

Recap of Previous Lecture

Topic Control Unit

Topic RISC vs CISC

Topic Floating Point Representation

Topics to be Covered

Topic Floating Point Representation

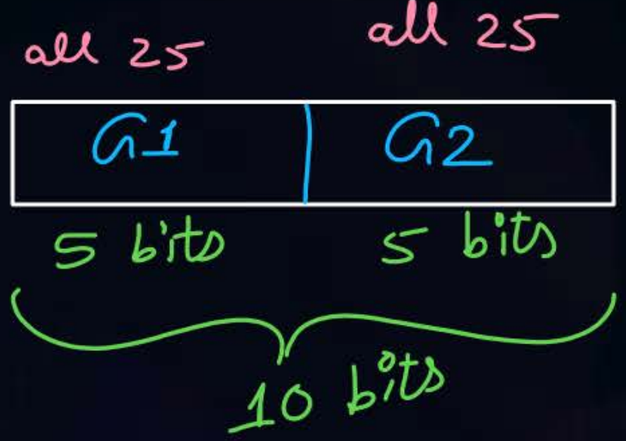
Topic IEEE-754 Floating Point Representation

Topic Booth Algorithm

[NAT]

Ans = 10

#Q. A micro-programmed control unit is required to generate a total of 25 control signals. Assume that during any microinstruction at most 2 control signals are active. Minimum number of bits required in the control word to generate the required control signals will be?



S1, S2, S3, ..., S25

example:-

microprog. control unit

No. of inst^{ns} supported = 16

each inst^{n} execution needs, 8 microinst^{ns}

Control word	MUX select	Add.
70 bits	4 bits	7

81 bits

Control mem. size = ——— ?

⇓

128 * 81 bits

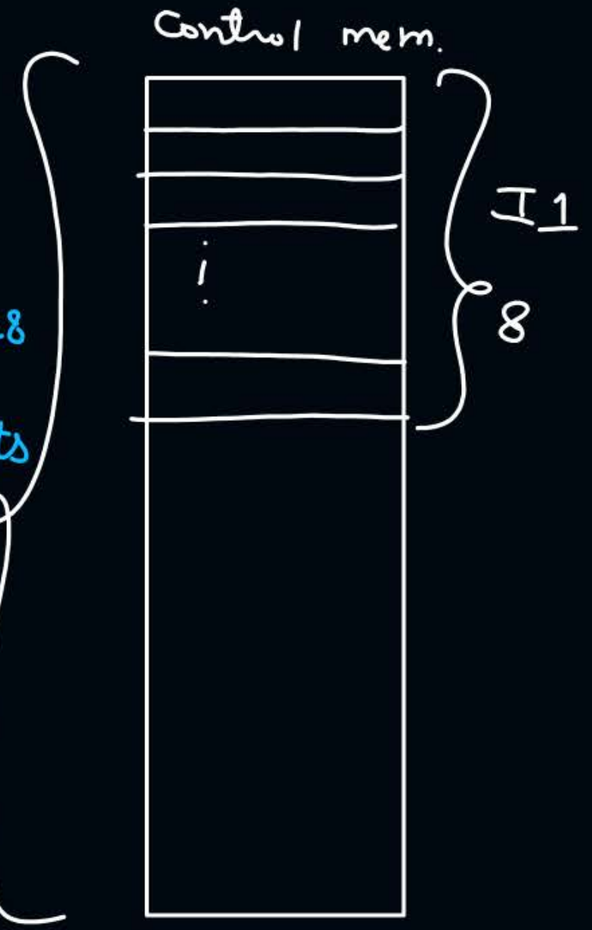
no. of microinst^{ns}

= 16 * 8

= 128

Size of Control memory

= 128 * 81 bits



[NAT]

#Q. Consider a microprogrammed control unit which has to support 64 number of instructions. For each instruction execution control unit generates a sequence of 64 control words. Each microinstruction contains 3 fields: 122 control signals to support horizontal control unit, a MUX select field to select one of 7 inputs, and a next address field. The size of control memory needed is?

no. of microinstns = 64 * 64
= 2^{12}
⇓
add. = 12 bits

Control mem. size = 2^{12} * 137 bits
= 548 k bits

[NAT]

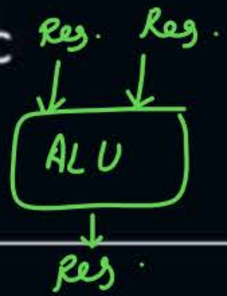
H.W.

#Q. Design of a vertical microprogrammed control unit requires to generate 50 signals. Out of first 43 those only 4 signals can be active at a time. And for remaining 7, anyone can be active anytime. The microinstruction of the control unit stores control signal information along with 3-bit mux select and 10-bits address field. The size of control memory required is?

Topic : RISC vs CISC

S. No.	RISC (Reduced Instruction-Set Computer)	CISC (Complex Instruction-Set Computer)
1.	Less Number of Instructions Supported	More Number of Instructions
2.	Fixed Length Instructions	Variable Length Instructions
3.	Simple Instructions	Complex Instructions
4.	Simple and less number of addressing Modes	Complex and More number of addressing Modes
5.	Easy to implement using hardwired control unit	Difficult to implement using hardwired control unit

Topic : RISC vs CISC

S. No.	RISC (Reduced Instruction-Set Computer)	CISC (Complex Instruction-Set Computer)
6.	One Cycle per instruction	More than one cycle per instruction
7.	Register-to-Register arithmetic operation only 	Register-to-Memory & Memory-to-Register arithmetic operations possible
8.	More Number of Registers	Less Number of Registers

[MCQ]

GATE- 2018

#Q. Consider the following processor design characteristics.

I. Register-to register arithmetic operations only

II. Fixed-length instruction format

III. Hardwired control unit

Which of the characteristics above are used in the design of a RISC processor?

A I and II only

B II and III only

C I and III only

D I, II and III

Topic : Floating-Point Numbers

motive ⇒ larger range of numbers with lesser bits.

Topic : Floating-Point Numbers

- The number is represented in format:

sign Exponent Mantissa / fraction

S	E	M

1 bit

- Mantissa is signed normalized (implicit/explicit) fraction number
- Exponent is stored in biased form. (original exponent + bias) ⇒ stored exponent

$$S = \begin{cases} 0 & +ve \\ 1 & -ve \end{cases}$$

Topic : Biased Exponent

if E represented in k bits

$$\text{bias} = 2^{k-1}$$

S	E	M

5 bits

$(-16 \text{ to } +15) \xrightarrow{\text{Transform}} 0 \text{ to } 31$

original (e)	stored (E) (excess −16)
−16	0
−15	1
−14	2
⋮	⋮
0	16
⋮	⋮
15	31

bias = 16

$E = e + 16$

Topic : Mantissa

101.01

→ Explicit normalizat^n $\Rightarrow 0.10101 * 2^3$

(1 should be '1')

$e = 3$

$E = 3 + bias$

$m =$ number after point

$= 10101$

→ Implicit normalizat^n $\Rightarrow 1.0101 * 2^2$

(1 should be 1)

$e = 2$

$E = 2 + bias$

$m = 0101$

Topic : Value Formula

$(-1)^S$ $\quad S=0 \Rightarrow (-1)^0 = 1$ $\quad S=1 \quad (-1)^1 = -1$

S	E	M

$$Value_{(explicit)} = (-1)^S * 0.M * 2^{E-bias}$$

$$Value_{(Implicit)} = (-1)^S * 1.M * 2^{E-bias}$$

[MCQ]

H.W.

#Q. A certain well-known computer family represents the exponents of its floating-point numbers as "excess-64" integers; i.e., a typical exponent $e_6e_5e_4e_3e_2e_1e_0$ represents the number:

A $e = -64 + \sum_{i=0}^{6} 2^i e_i$

B $e = -64 + \sum_{i=0}^{6} 2e_i$

C $e = 64 - \sum_{i=0}^{6} 2^i e_i$

D $e = 64 - \sum_{i=0}^{6} 2e_i$

[NAT]

explicit

#Q. Consider a 16-bit register used to store floating point numbers. The mantissa is normalized signed fraction number. Exponent is represented in excess-32 form. What is the 16-bit value for $+(23.5)_{10}$ in this register?

$bias = 32 = 2^{k-1} \Rightarrow 2^5 = 2^{k-1} \Rightarrow 5 = k-1 \Rightarrow k = 6$

+ve $\Rightarrow s = 0$

16

s	E	m
1	6	9
0	100101	101111000

$(23.5)_{10} = (10111.1)_2$

$\Downarrow$

Explicit normalize

$\Downarrow$

$0.101111 * 2^5$

$e = 5 \Rightarrow E = 5 + 32 = 37 = 100101$

$m = 101111$

[NAT]

#Q. What is the 4-digit hexadecimal value for $+(23.5)_{10}$ in above question's register?

0 100 101 1011 1000

4B78

[NAT]

#Q. What is the 4-digit hexadecimal value for $+(39.75)_{10}$ in above question's register?

S	E	M
0	100110	10011110

4D3E Ans.

$(100111.11)_2$

⇓

explicit norm.

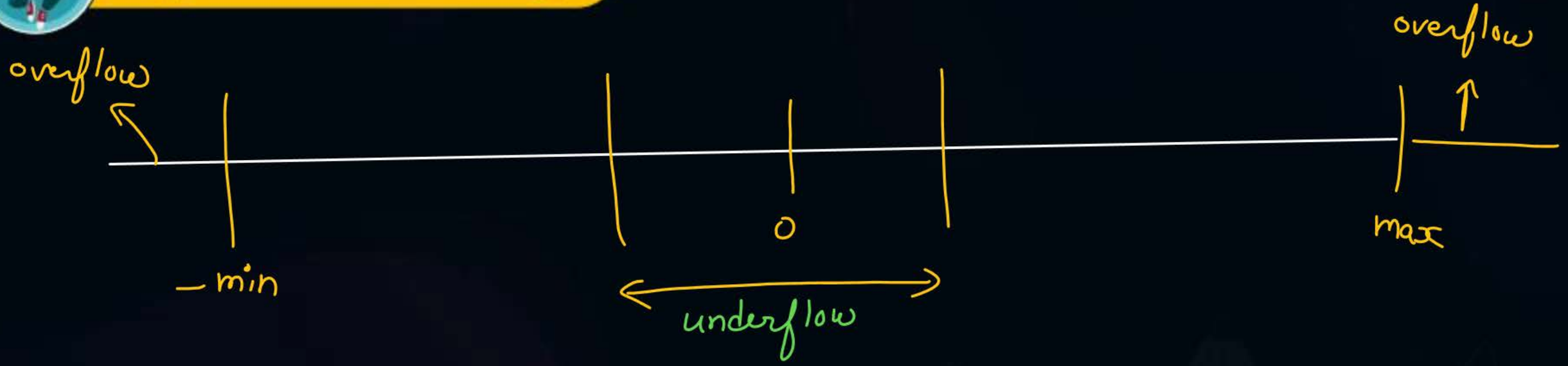
⇓

$0.10011111 * 2^6$

$e = 6 \Rightarrow E = 6 + 32 = 38 = (100110)_2$

$m = 10011111$

Topic : Number Range



underflow example :-

bias = 32

E ⇒ 0 to 63

⇓

e ⇒ −32 to 31

value to store

⇒ 0.0000----011

⇓

explicit normalize

⇓

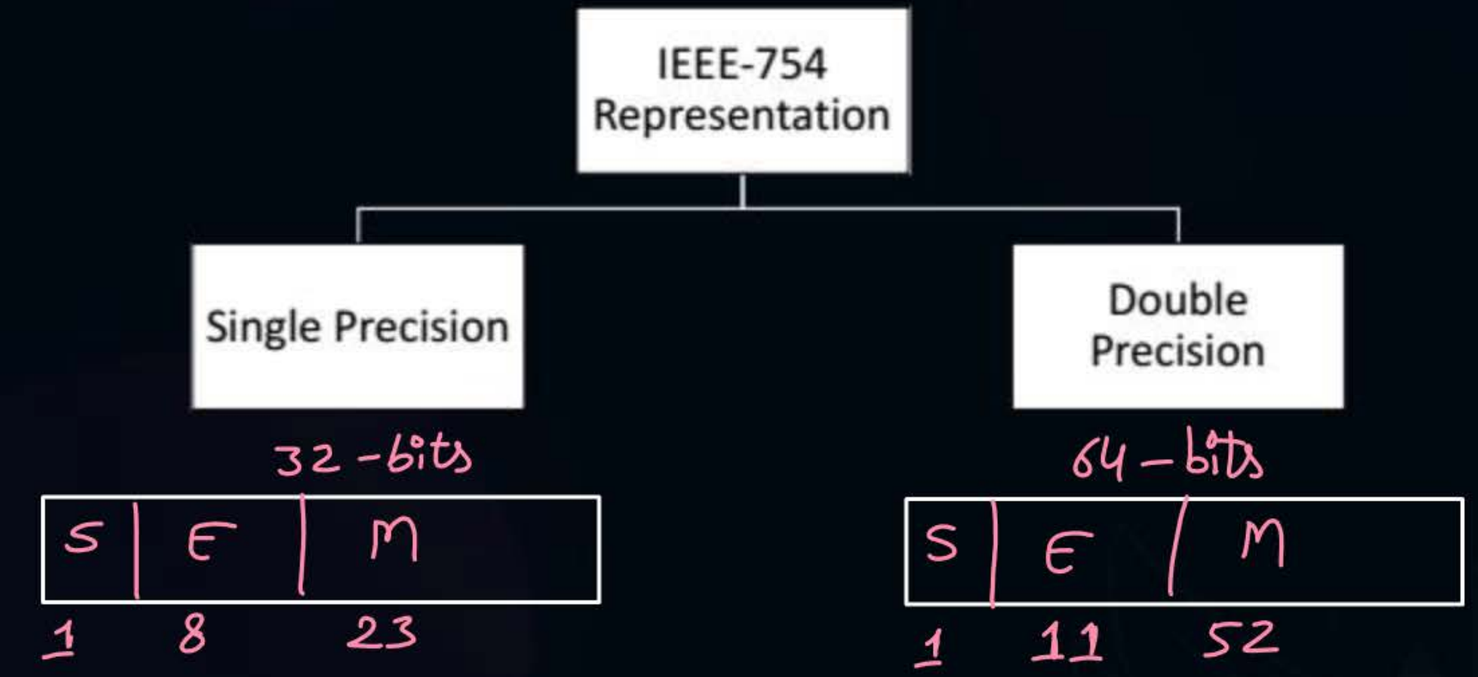
$0.11 * 2^{-33}$ ⇒ number can not be stored

not allowed

Topic : Disadvantages of Conventional Representation

1. Can not store zero
2. Has underflow

Topic : IEEE-754 Floating Point Representation



$$E = \left\{ \begin{matrix} 00\ldots\ldots0 \\ \text{or} \\ 11\ldots\ldots1 \end{matrix} \right\} \text{ special case}$$

Topic : IEEE-754 Floating Point Representation

S	E	M	Number
0	0 ~ ~ ... ~ 0	0 ... 0	+0
1	0 ... 0	0 ... 0	-0
0	11 ... ~ ... 1	0 ... 0	$+\infty$
1	11 ... ~ 1	0 ... 0	$-\infty$
0 or 1	11 ... 1	$M \neq 00 \ldots 0$	N.A.N. (Not A Number)
0 or 1	00 ... 0	$M \neq 0 \ldots 0$	Denormalized number
0 or 1	$E \neq 0 \ldots 0$ and $E \neq 11 \ldots 1$	$m = 0 \ldots 0$ to $11 \ldots 1$	Implicit normalized number

Denormalized numbers:-

A very-very small number which can not be stored as normalized number.

for single precision ⇒

bias = 127

⇓

for Normalized number

E ⇒ 00000001 to 11111110 } 1 to 254

$E_{min} = 1$

$e_{min} = 1 - 127 = -126$

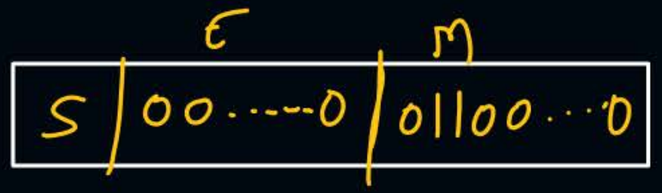
ex:- number ⇒ 0.00000....011

⇓

normalize ⇒ $0.011 * 2^{-126}$

⇓

Can not be normalized hence store it as denormalized number



single

double precision

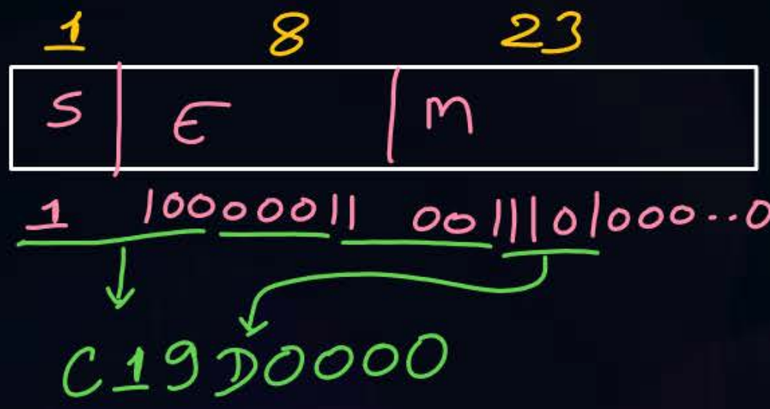
$$\text{Value}_{(\text{denormalized})} = (-1)^S * 0.M * 2^{-126} \text{ or } -1022$$

E − bias

$$\text{Value}_{(\text{implicit})} = (-1)^S * 1.M * 2$$

[NAT]

#Q. The value of a float type variable is represented using the single- precision 32-bit floating point format IEEE-754 standard that uses 1bit for sign, 8 bits for biased exponent and 23 bits for mantissa. A float type variable X is assigned the decimal value of −19.625. The representation of X in hexadecimal notation is?



S = 1

$(10011.101)_2$

⇓

Implicit normalize

⇓

$1.0011101 * 2^4$

$e = 4 \Rightarrow E = 4 + 127 = 131 = (10000011)_2$

$m = 00111010000\cdots0$

[NAT]

Ans = $+(26)_{10}$

#Q. The value represented by the following 32-bits in IEEE-754 representation is?

01000001110100000...00

S | E | M

$E \neq 0\ldots0$ and $\neq 11\ldots1$ } Implicit normalized

$E = (10000011)_2$

$= (131)_{10}$

$Value = +1.101 * 2^{131-127}$

$= +1.101 * 2^{4}$

$= +(11010)_2$

$= +(26)_{10}$

[NAT]

#Q. The value represented by the following 32-bits in IEEE-754 representation is?

00000000011000000...00

E = 0---0 and m ≠ 0---0 } denormalized

value = + 0.11 * 2^{-126}

= + 11.0 * 2^{-2} * 2^{-126}

= + (3 * 2^{-128})

[MCQ] GATE- 2014 H.W.

#Q. The value of a float type variable is represented using the single- precision 32-bit floating point format IEEE-754 standard that uses 1bit for sign, 8 bits for biased exponent and 23 bits for mantissa. A float type variable X is assigned the decimal value of −14.25. The representation of X in hexadecimal notation is

A C1640000H

B 416C0000H

C 41640000H

D C16C0000H

[NAT] GATE - 2021

#Q. Consider the following representation of a number in IEEE 754 single-precision floating point format with a bias of 127.

S : 1 E : 10000001 $=(129)_{10}$ F: 11110000000000000000000

Here S, E and F denote the sign, exponent and fraction components of the floating-point representation.

The decimal value corresponding to the above representation (rounded to 2 decimal places) is -7.75

$$\text{Value} = -1.1111 * 2^{129-127}$$

$$= -1.1111 * 2^{2}$$

$$= -111.11$$

$$= -(7.75)_{10}$$

[MSQ] GATE-2025

H.W.

Q.49	Three floating point numbers X, Y, and Z are stored in three registers R_X, R_Y, and R_Z, respectively in IEEE 754 single precision format as given below in hexadecimal: R_X = 0xC1100000, R_Y = 0x40C00000, and R_Z = 0x41400000 Which of the following option(s) is/are CORRECT?
(A)	$4(X + Y) + Z = 0$
(B)	$2Y - Z = 0$
(C)	$4X + 3Z = 0$
(D)	$X + Y + Z = 0$

2 mins Summary

Topic Control Unit

Topic RISC vs CISC

Topic Floating Point Representation

Happy Learning

THANK - YOU