CS & IT ENGINEERING

COMPUTER ORGANIZATION AND ARCHITECTURE

Floating Point Representation

Lecture No.- 02

By- Vishvadeep Gothi sir

Recap of Previous Lecture

Topic Floating-Point Numbers

Topic Biased Exponent

Topic Normalization: Implicit & Explicit

Topics to be Covered

Topic Floating-Point Numbers

Topic Biased Exponent

Topic Number Range

Topic IEEE-754 Floating Point Representation

Topic Denormalized Number

Topic : Floating-Point Numbers

- The number is represented in format:

S	E	M

- Mantissa is signed normalized (implicit/explicit) fraction number
- Exponent is stored in biased form.

[NAT]

explicitly

#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 $+(11.5)_{10}$ in this register?

bias = 32

16 – bits

S	E	M
1	6	9

$$2^{k-1} = 32$$

$$2^{k-1} = 2^5$$

$$k-1 = 5$$

$$k = 6$$

$+(11.5)_{10} \Rightarrow$ positive $\Rightarrow S = 0$

$$(11.5)_{10} = (1011.1)_2$$

explicit Normalizat^n

$$\Downarrow$$

$$0.10111 * 2^4$$

$$M = 101110000$$

$$e = 4$$

$$E = 4 + 32 = 36 = (100100)_2$$

S	E	M
0	100100	101110000

[NAT]

#Q. What is the 4-digit hexadecimal value for $+(11.5)_{10}$ in above question's register?

$\underline{0 \quad 100}\underline{100} \; \underline{1011}\underline{10000}$

4 9 7 0

$(4970)_{16}$

0x4970

4970H

[NAT]

Ans = 4D2E

#Q. What is the 4-digit hexadecimal value for $+(37.75)_{10}$ in above question's register?

$$(37.75)_{10} = (100101.11)_2$$

$\Downarrow$

Explicit normalization

$\Downarrow$

$$0.10010111 * 2^6$$

$M = 100101110$

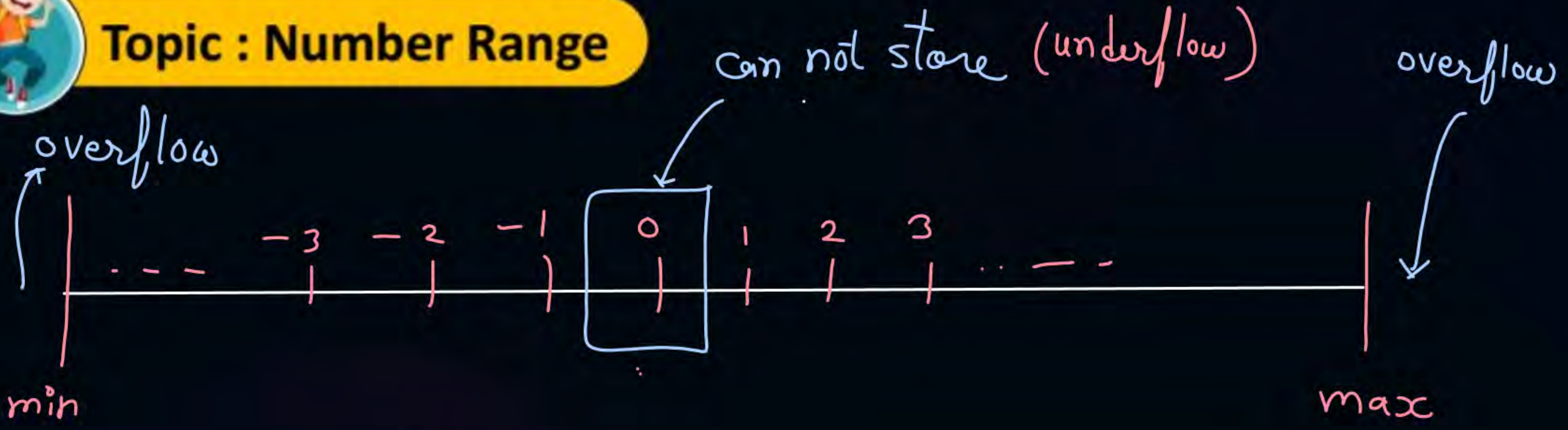
$e = 6$

$E = 6 + 32 = 38 = (100110)_2$

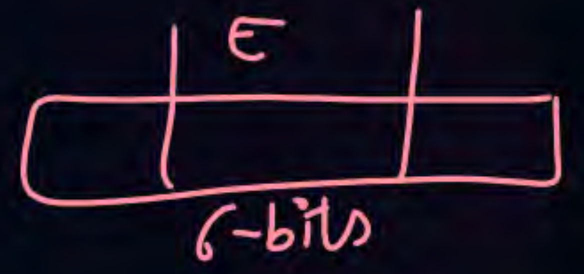
1	6	9
S	E	M
0	100110	100101110

4 D 2 E

Topic : Number Range



example :-



$E_{min} = 0$

$e_{min} = 0 - 32$

$= -32$

if any value $\Rightarrow$ 0.0000...... 11

$\Downarrow$

explicit normalizat^n

$\Downarrow$

$0.11 * 2^{-33}$

$e = -33 \Leftarrow$ cannot store

Topic : Bits in E and M

More no. of bits in E $\Rightarrow$ larger range of numbers

—— " —— M $\Rightarrow$ Better precision or Accuracy

Topic : Disadvantages of Conventional Representation

→ can not represent zero.

→ it can not store very small numbers around zero.
(has underflow)

Topic : IEEE-754 Floating Point Representation

IEEE-754 Representation

- Single Precision
  - 32 -bits
- Double Precision
  - 64-bits

Single Precision (32 bits)

S	E	M
1	8	23

bias = 127

Double Precision (64 bits)

S	E	M
1	11	52

bias = 1023

$E = 00\ldots\ldots 0$

or

$E = 11\ldots\ldots 1$

special number (exceptions)

$E \neq 00\ldots\ldots 0$

and

$E \neq 11\ldots\ldots 1$

normal number (Implicitly normalized)

Topic : IEEE-754 Floating Point Representation

S	E	M	Number
0	00.......0	0.......0	+0
1	00.......0	0.......0	−0
0	11.......1	0.......0	+∞
1	11.......1	0.......0	−∞
0 or 1	11.......1	M ≠ 0.......0	N.A.N. (Not A Number)
0 or 1	00.......0	M ≠ 0.......0	Denormalized number
0 or 1	E ≠ 0.......0 and E ≠ 11.......1	xxxxxx.......x	Implicitly normalize

Topic : Denormalized Number

A very-very small number which can not be implicitly normalized.

single precision

S	E	M
1	8	23

bias = 127

E_{min} for normalized number = 1

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

ex:-
Value = 0.00000........11
⇓
Implicit normalizat^{n}
⇓
$1.1 * 2^{-128}$ } not allowed beyond 2^{-126}
⇓
if can not be normalized till 2^{-126} then store the number as denormalized number

$0.000\cdots\cdots11$

$\downarrow$

try to normalized till 2^{-126}

$\downarrow$

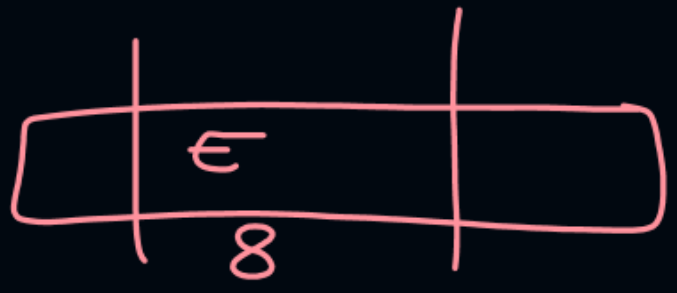
$0.\underbrace{011}_{M} * 2^{-126}$

	E	M
S	0 0	01100...0

value = $0.M * 2^{-126}$

$$\underset{\text{(Implicit)}}{\text{Value}} = (-1)^S * 1.M * 2^{E-bias}$$

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



max value $e = +127$

max value $E = 127 + \text{bias}$

$= 127 + 127$

$= (254)_{10}$

$= (11111110)_2$

that's why $E = 111....1$ is preserved.

[NAT]

Ans = $(C1DD0000)_{16}$

#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 −27.625. The representation of X in hexadecimal notation is?

$(27.625)_{10} = (11011.101)_2 \Rightarrow$ Implicit normalizatⁿ

$\Downarrow$

$1.1011101 * 2^4$

$s = 1$

S	E	M

$1 \; 10000011 \; 10111010100 \cdots 0$

C 1 D D 0000

$M = 1011101 00 \cdots 0$

$e = 4$

$E = 4 + 127 = (131)_{10} = (10000011)_2$

[NAT]

Ans = $+(28)_{10}$

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

$\underbrace{0}_{S}\underbrace{10000011}_{E}\underbrace{1100000...00}_{M}$

$S = 0 \Rightarrow +ve$

$E = 10000011 = (131)_{10}$

$M = 1100\cdots0$

$E \neq 0\cdots0$ and $E \neq 11\cdots1$ } Implicit normalized

$$\text{value} = 1.1100\cdots0 * 2^{131-127}$$

$$= 1.110\cdots0 * 2^{4}$$

$$= (11100.0)_2$$

$$= +(28)_{10}$$

[NAT]

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

$\underbrace{0}_{S=0}\underbrace{00000000}_{E=00000000}\underbrace{1100000...00}_{M=1100\cdots0}$

$\left.\begin{array}{l} E = 0\cdots0 \\ \text{and} \\ M \neq 0\cdots0 \end{array}\right\}$ denormalized

$$\text{Value} = 0.11 * 2^{-126}$$
$$= 11.0 * 2^{-2} * 2^{-126}$$
$$= +3 * 2^{-128}$$

[NAT]

#Q. Maximum value represented in IEEE-754 single precision?

$+\infty$

[NAT]

#Q. Minimum value represented in IEEE-754 single precision?

$-\infty$

[NAT]

#Q. ~~Minimum~~ Max positive normalized value represented in IEEE-754 single precision?

S	E	M
0	$\underbrace{11111110}_{\Downarrow\ 254}$	111.........1

$$\text{Value} = +\,1.11\ldots1 * 2^{254-127}$$

$$= 111\ldots1.0 * 2^{-23} * 2^{127}$$

$$= +\left(2^{24}-1\right) * 2^{104}$$

min possible normalized value.

1	11111110	111...1

$$\text{Value} = -\left(2^{24}-1\right) * 2^{104}$$

[NAT]

#Q. Minimum positive normalized value represented in IEEE-754 single precision?

S	E	M
0	$(00000001)_2$	$00\cdots0$

$$(00000001)_2 \Downarrow (1)_{10}$$

$$\text{value} = +\ 1.00\cdots0 * 2^{1-127}$$

$$= +\ 2^{-126}$$

[NAT]

Maximum

#Q. ~~Minimum~~ positive denormalized value represented in IEEE-754 single precision?

S	E	M
0	00000000	11......1

$$\text{Value} = 0.11\ldots\ldots1 * 2^{-126}$$

$$= 111\ldots\ldots1.0 * 2^{-23} * 2^{-126}$$

$$= +(2^{23}-1) * 2^{-149}$$

[NAT]

Minimum

#Q. ~~Maximum~~ positive denormalized value represented in IEEE-754 single precision?

S	E	M
0	00000000	0000....01

$$\text{Value} = +\,0.000\cdots01 * 2^{-126}$$

$$= +1.0 * 2^{-23} * 2^{-126}$$

$$= 2^{-149}$$

[NAT]

H.W.

#Q. How to represent +1 and -1 in IEEE-754 single precision floating point number?

[NAT]

H. W.

#Q. How to represent +0.0000101 in IEEE-754 single precision floating point number?

[MCQ]

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-PYQ Ans. $-(7.75)$

#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__________

$$\left.\begin{array}{l} E \neq 0\ldots0 \\ \text{and} \\ E \neq 11\ldots1 \end{array}\right\} \text{Implicit normalized}$$

$$\begin{aligned} \text{Value} &= -1.1111 * 2^{129-127} \\ &= -1.1111 * 2^{2} \\ &= -(111.11)_2 \\ &= -(7.75)_{10} \end{aligned}$$

[~~NAT~~] MCQ GATE- PYQ

#Q. The format of the single-precision floating-point representation of a real number as per the IEEE 754 standard is as follows:

Sign	Exponent	mantissa

Which one of the following choices is correct with respect to the smallest normalized positive number represented using the standard?

A. exponent = 00000001 and mantissa = 00000000000000000000001

B. exponent = 00000001 and mantissa = 00000000000000000000000

C. exponent = 00000000 and mantissa = 00000000000000000000000

D. exponent = 00000000 and mantissa = 00000000000000000000001

2 mins Summary

Topic Biased Exponent

Topic Normalized Mantissa

Topic Explicit vs Implicit Normalization

Topic IEEE-754 Floating Point Representation

Happy Learning

THANK - YOU