CS & IT ENGINEERING

COMPUTER ORGANIZATION AND ARCHITECTURE

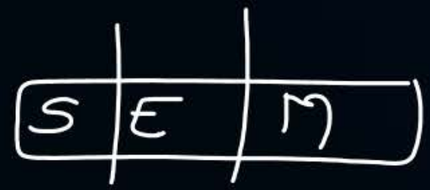
Floating Point Representation

Lecture No.- 02

By- Vishvadeep Gothi sir

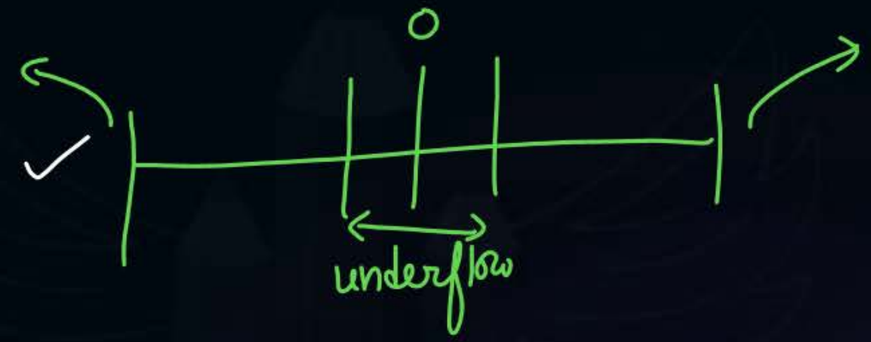
Recap of Previous Lecture

Topic — Floating-Point Numbers



Topic — Biased Exponent

Topic — Number Range



Topic — IEEE-754 Floating Point Representation

Topic — Denormalized Number

Topics to be Covered

- Topic: IEEE Floating-Point Representation
- Topic: Denormalized Number

Topic : IEEE-754 Floating Point Representation

$*2^{e}$ — e

e + bias ⇒ ∈

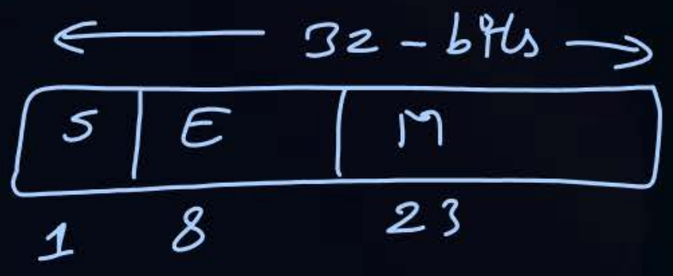
IEEE-754 Representation

Single Precision

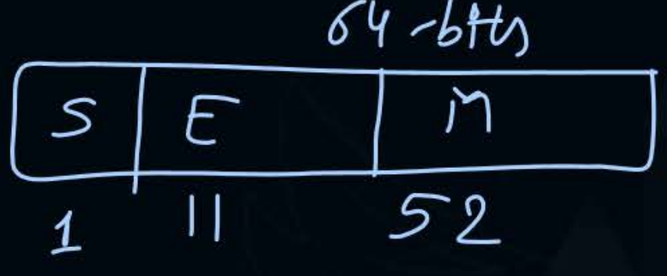
Double Precision

float variable of C-lang.

double - variable of C-lang.



bias = 127



bias = 1023

$E = 00\ldots\ldots0$
or
$E = 11\ldots\ldots11$
} special number

$E \neq 0\ldots\ldots0$
and
$E \neq 11\ldots\ldots1$
} normal number (Implicit normalization)

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	XX.....X	Implicitly normalized number

} 2 representations of zero

⇒ The number which can not be normalized.

Single precision IEEE-754

bias = 127

for a normalized number (implicitly)

Min	Max
$E \Rightarrow (00000001)_2$	$(11111110)_2$
$\Downarrow$	$\Downarrow$
$(1)_{10}$	$(254)_{10}$

Range of $E \Rightarrow 1$ to 254

Range of $e \Rightarrow -126$ to 127

$e \Rightarrow 1-127 = -126$

$254-127 = 127$

ex:- after normalizat^{n},

value = $1.01 * 2^{128}$

101000.....--- 0.0

NOT A Number

e = 128

E = 128 + 127

= $(255)_{10}$

= $(11111111)_{2}$

Denormalized Number

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

ex:- $0.0000\ldots\ldots0101$

$\Downarrow$

Implicitly normalizat^{n} $\longrightarrow$ $0.101 * 2^{-126}$

store it as denormalized form

$\downarrow$

$1.01 * 2^{-127}$

not allowed because

$e = -127 < -126$

S	E	M
0 or 1	00000000	10100....0

ex:- 0.0000.....11

⇓

normalize

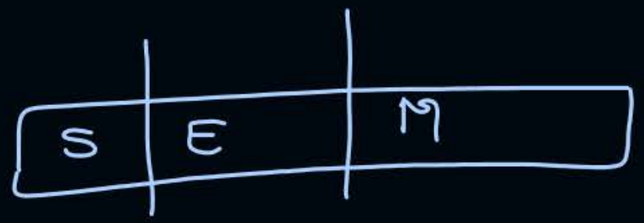
⇓

$0.00011 * 2^{-126}$

⇓

number is not implicitly normalized hence store it as denormalized number.

S	E	M
0 or 1	00000000	00011 00.....0



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

$$\text{Value (denormalized)} = (-1)^{S} * 0.M * 2^{-(\text{bias} - 1)}$$

$-(\text{bias} - 1) = -126$ ⇐ single precision

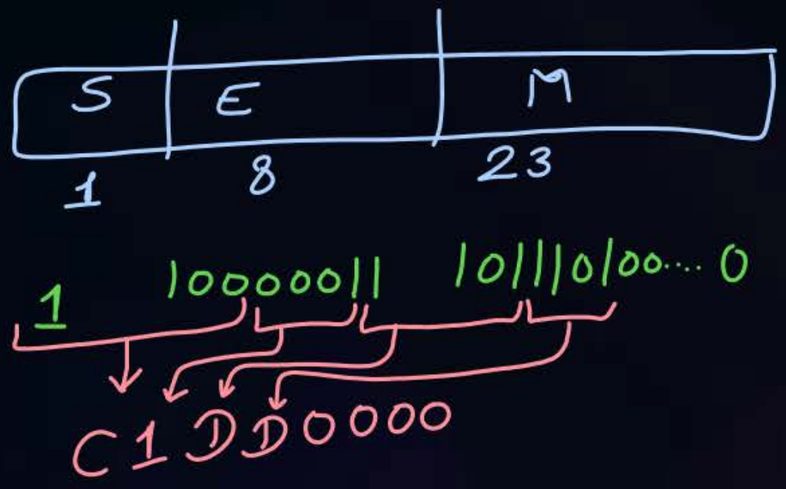
or

$(-1)^{S} * 0.M * 2^{-1022}$ ⇐ double precision

[NAT]

Ans = C1DD0000

#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?



$S = 1$

$(27.625)_{10} = (11011.101)_2$

⇓ Implicit normalizat^n

⇓

$1.1011101 * 2^4$

$e = 4$

$E = 4 + 127$

$= (131)_{10}$

$= (10000011)_2$

$M = 10111010000....0$

#Q. $-(37.25)_{10}$ how to represent it in IEEE-754 single precision representation ?

Ans = $(C2150000)_{16}$

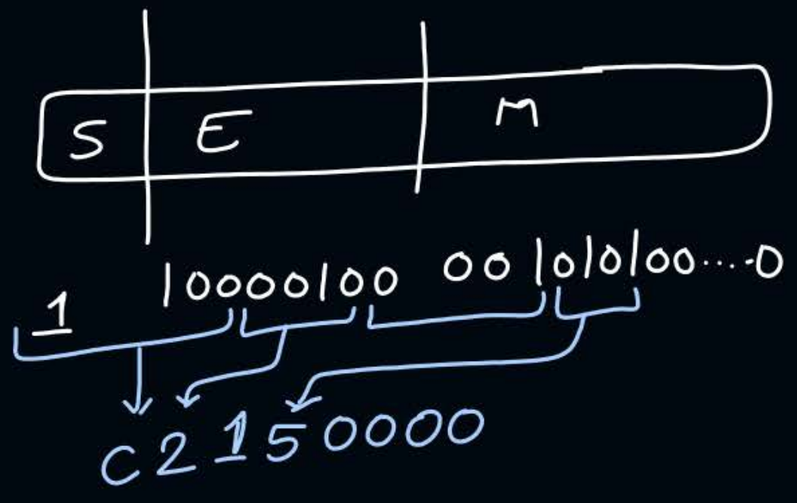
Soln

$(37.25)_{10} = (100101.01)_2$

↳ Implicit normalizat^n

$\Downarrow$

$1.00101101 * 2^5$



$M = 00101101$

$e = 5$

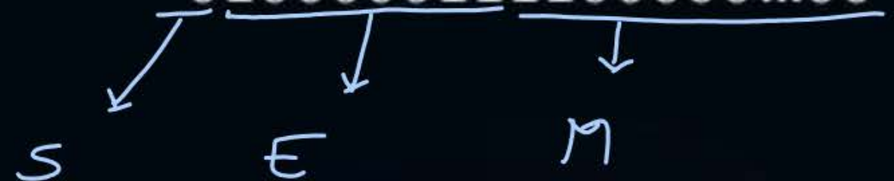
$E = 5 + 127 = (132)_{10} = (10000100)_2$

[NAT]

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

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

0100000111100000...00



$E = (10000011)_2$

$= (131)_{10}$

$M = 110\ldots0$

$E \neq 0\ldots0$ and $\neq 11\ldots1$ } Implicit normalized no.

$Value = +1.11 * 2^{131-127}$

$= 1.11 * 2^{4}$

$= (11100.0)_2$

$= +(28)_{10}$

#Q. 0 10000010 10101000000...0

S ↓ | E | M

Above 32-bits represents a float value in Single precision IEEE-754 format. The float value represented is ____ ?

Soln $S = 0 \Rightarrow +ve$

$E = (10000010)_2 = (130)_{10}$

Value $= +1.10101 * 2^{130-127}$

$= 1.10101 * 2^3$

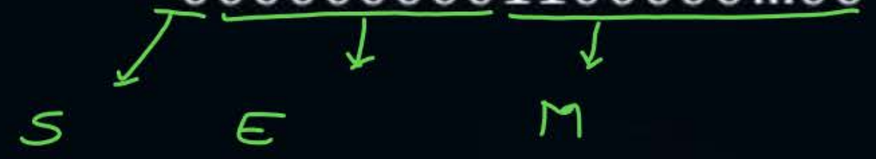
$= (1101.01)_2$

$= +(13.25)_{10}$ ⇐ Ans.

[NAT]

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

00000000011000000...00



E = 00...0 and M ≠ 00.....0 } denormalized

Value = + 0.11 * 2^{-126}

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

= 3 * 2^{-128}

[NAT]

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

$+\infty$

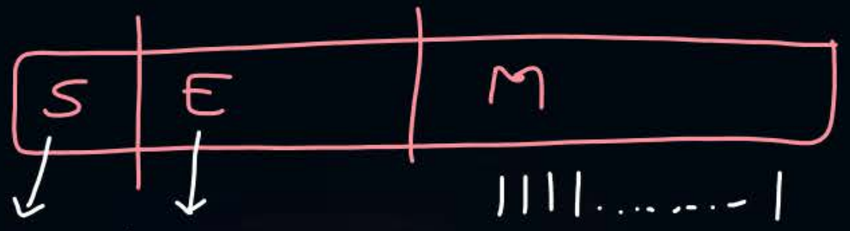
[NAT]

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

$-\infty$

[NAT]

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



0 for +ve

$(1111111_0)_2$

$\Downarrow$

$(254)_{10}$

$$\text{Value} = 1.11\ldots.1 * 2^{254-127}$$

$$= 1.111\ldots1 * 2^{127}$$

$$= 1111\ldots1.0 * 2^{-23} * 2^{127}$$

$$= (2^{24}-1) * 2^{104}$$

[NAT]

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

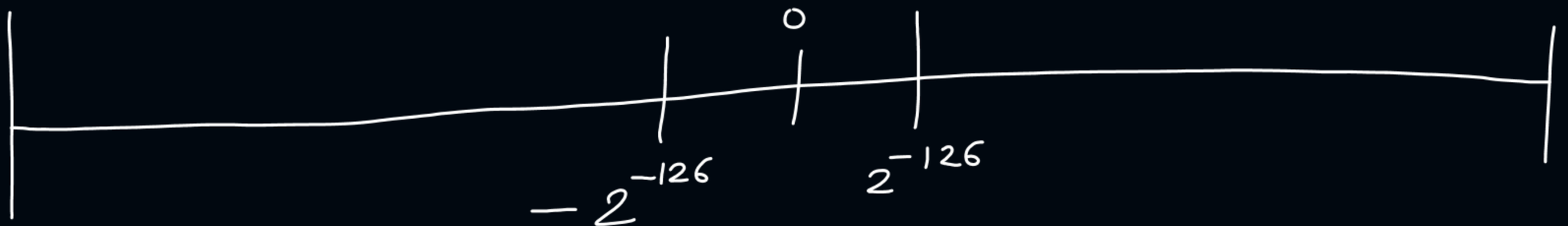
$$\text{Value} = 1.0 * 2^{1-127}$$

$$= 2^{-126}$$

normalized values range

$-(2^{24}-1)*2^{104}$

$+(2^{24}-1)*2^{104}$



[NAT]

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

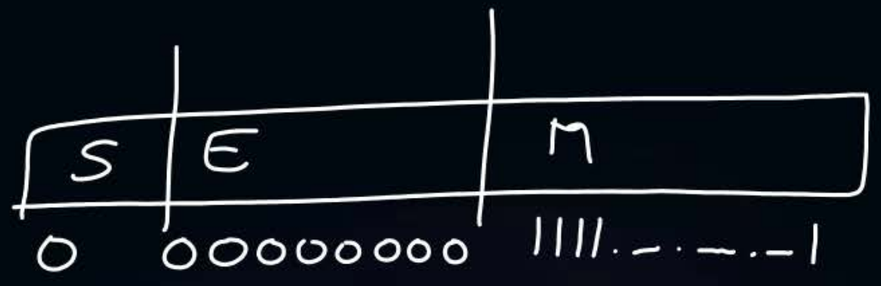
$$\text{value} = 0.0000\ldots01 * 2^{-126}$$

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

$$= +(2^{-149})$$

[NAT]

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



$$\text{Value} = 0.11\ldots-1 * 2^{-126}$$

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

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

[NAT]

Homework

#Q. How to represent $(+1)_{10}$ and $(-1)_{10}$ 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]

H. W.

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

S : I E : 10000001 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_____

[NAT]

H.W.

#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 IEEE Floating-Point Representation

Topic Denormalized Number

Happy Learning

THANK - YOU