Floating Point Representation

32 Bit IEEE 754

$(-10.75)_{10} \longrightarrow$ Floating point respresentation (?)

Solution →

Step(I): Convert the number in binary form.

$(10.75)_{10} \longrightarrow$ Binary

$$\begin{array}{r|l l}
2 & 10 & \\
2 & 5 & 0 \\
2 & 2 & 1 \\
2 & 1 & 0 \\
 & 0 & 1
\end{array}$$

$0.75 \times 2 \rightarrow 1.50 \rightarrow 1$

$0.50 \times 2 \rightarrow 1.00 \rightarrow 1$

$(10.75)_{10} \longrightarrow (1010.11)_2$

Step II: Scientific form representation

$(1010.11)_2$ —— $(1.\underbrace{01011.....}_{\text{Mantissa}}) \times 2^3$

Step III: Exponent calculation

$x = 3$

Exponent $= 127 + x$

$= 127 + 3 = 130$

$130 \xrightarrow{\text{Binary}} (10000010)_2$

Step IV

1 bit	8 bits	23 bits
1	1 0 0 0 0 0 1 0	0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Sign	Exponent	Mantissa

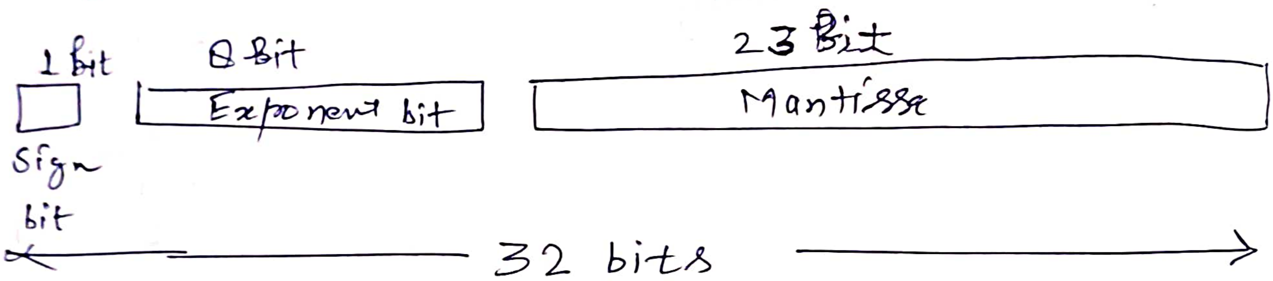
← 32 bits →

If given number is positive then sign bit = 0

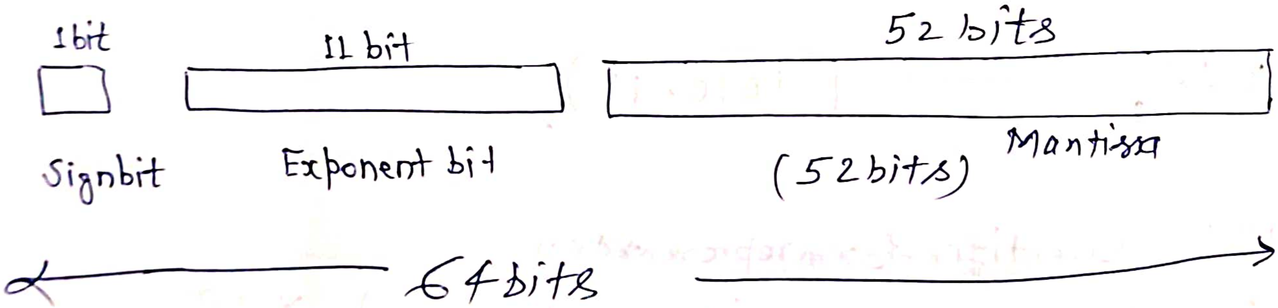
If given number is negative then sign bit = 1

* Combining sign bit, exponent bits and mantissa bits will be the floating point representation [IEEE 754 format].

* Single precision [32bit IEEE 754]



* Double precision [64bit IEEE 754]



Floating point ⟶ Decimal number (32 bit IEEE 754)

Q: 1 10000010 01011000000000000000000

Solution: Step(I)

1 $\underbrace{10000010}$ 0101100.........

↓

Number will be negative

$$(10000010)_2 \xrightarrow{\text{Decimal}} (130)_{10}$$

$$127 + x = 130$$

$$x = 3$$

Step (II) ∴ Scientific representation ⇒

$$[1.xxxx\ldots\ldots] \times 2^x$$

$$[1.\text{Mantissa}\ldots\ldots] \times 2^x$$

$$[1.01011000000000\ldots\ldots] \times 2^3$$

$$= 1010.1100000000000\ldots\ldots$$

$$= (10.75)_{10}$$ Decimal value

Step III final answer $= (-10.75)_{10}$

Floating point addition : (32 bit IEEE 754)

$$\begin{array}{r} (100)_{10} \\ +(0.25)_{10} \\ \hline ? \\ \hline \end{array}$$

$(100)_{10} \rightarrow 1100100 \longrightarrow 1.100100 \times 2^{6}$

$(0.25)_{10} \rightarrow 0.0100 \longrightarrow [1.0000.....] \times 2^{-2}$

IEEE representation ⇒

0	1000 0101	1001 0000 0000 0000 0000 000 ; $(100)_{10}$
0	0111 1101	0000 0000 0000 0000 0000 000 ; $(0.25)_{10}$

Note

Exponent of 0.25 is $(01111101)_2$ and exponent of 100 is $(10000101)_2$

We have to make both exponent same. Because exponent of 0.25 is lesser than exponent of 100 so we will increase the exponent of 0.25.

Hidden bit

Initially	0111 1101	1	0000 0000 0000 0000 0000 000
	0111 1110	0	1000 0000 0000.0000 0000 000
	0111 1111	0	0100 0000 0000 0000 0000 000
	1000 0000	0	0010 0000 0000 0000 0000 000
	1000 0001	0	0001 0000 0000 0000 0000 000
	1000 0010	0	0000 1000 0000 0000 0000 000
	1000 0011	0	0000 0100 0000 0000 0000 000
	1000 0100	0	0000 0010 0000 0000 0000 000
	1000 0101	0	0000 0001 0000 0000 0000 000

Hiddenbit

0	1000 0101	1	1001 0000	0000	0000	0000 000
0	1000 0101	0	0000 0001	0000	0000	0000 000
0	1000 0101	1	1001 0001	0000	0000	0000 000

(Add the mantissa)

Remove this

* Remove the hidden bit and final answer is given as:-

<u>Ans</u> 0 1000 0101 10010001 0000 0000 0000 000

<u>Note</u>

$$0 \quad \underbrace{1000\,0101}_{\downarrow\,(133)_{10}} \quad \underbrace{1001\,0001\;0000\,0000\;0000\,000}_{\text{Mantissa}}$$

$127 + x = 133$

$\therefore x = 6$

$[1.\text{Mantissa}] \times 2^x$

$[1.100100010000....] \times 2^6$

$= 1100100.010000......$

$= (100.25)_{10}$; (Decimal), <u>Ans</u>

Floating point Subtraction (32 bit IEEE 754)

$(100)_{10} - (0.25)_{10} = ?$

Solution

		Hidden Bit		
0	1000 0101	1	1001 0000 0000 0000 0000 000	← $(100)_{10}$
1	0111 1101	1	0000 0000 0000 0000 0000 000	← $(0.25)_{10}$

0	1000 0101	1	1001 0000 0000 0000 0000 000	
1	1000 0101	0	0000 0001 0000 0000 0000 000	
0	1000 0101	1	1000 1111 0000 0000 0000 000	Difference of both the Mantissa

→ Remove this

* Remove the hidden bit and final answer is given as :-

Ans 0 1000 0101 1000 1111 0000 0000 0000 000

$\underbrace{10000101}_{(133)_{10}}$ $\underbrace{1000\ 1111\ 0000\ 0000\ 0000\ 000}_{\text{Mantissa}}$

$127 + x = 133 \quad \therefore x = 6$

Ans = $[1.\text{Mantissa}] \times 2^{x}$

$[1.1000\,1111\,0000 \ldots] \times 2^{6}$

$= 1100011.110000\ldots$

$= (99.75)_{10}$ [Answer in decimal form]

Floating point multiplication (32 bit IEEE 754)

$(100)_{10}$
$\times (0.25)_{10}$
25

1100100
$\times$ .01
11001.00

IEEE 754 representation:

$100 \Rightarrow 1.100100 \times 2^6$
$0.25 \Rightarrow 1.0000\ldots \times 2^{-2}$
} Scientific form

100 ⇒	0	1000 0101	1001 0000 0000 0000 0000 000
0.25 ⇒	0	0111 1101	0000 0000 0000 0000 0000 000

10000 0010 (Add the exponent)

− 00111 1111 ⇐ $(127)_{10}$

1000 0001

**
127 + 6
127 + (−2)
127 + 127 + 4

* Include hidden bit for both the mantissa and then multiply.

Hidden bit
↓
1 . 1001 0000 ← Mantissa →
1 . 0000 0000 ← Mantissa →

1.1001 0000

multiply the above number

Remove this

* Remove the hidden bit and your final answer will be as ⇒

0 1000 0001 1001 0000 . . . .

1000 0001 ⟶ 131

$127 + x = 131$

$x = 4$

$$[1.\text{Mantissa}] \times 2^{x}$$

$$[1.10010000\ldots] \times 2^{4}$$

$$= 11001.0000$$

$$= 11001$$

$$= (25)_{10}$$ (Decimal Value)

Note :

* If multiplication output (result) is not in scientific form then we have to make multiplication result in scientific form. Accordingly we have to adjust the exponent.

Floating point division
(32bit IEEE 754)

Divisor) Dividend (Quotient
Remainder

$$\frac{(10.35)_{10}}{(2.25)_{10}} = \frac{1010.0101100\,1100110\,0\,1100\,1100}{10.01}$$

2.25) 10.35 (4.6

$(10.35)_{10} \Longrightarrow [1.010010\ 1100\ 1100\ 1100\ 1100\ 1100] \times 2^3$

$(2.25)_{10} \Longrightarrow [1.001] \times 2^1$

23rd bit

IEEE representation :-

10.35 ;	0	10000010	0100 1011 0011 0011 0011 001
2.25 ;	0	10000000	0010 0000 0000 0000 0000 000
		00000010	(Take the difference of exponent)

* Include hidden bit for both mantissa and then divide.

$$\frac{1.\underbrace{0100\ 1011\ 0011\ 0011\ 0011\ 001}_{\text{mantissa}}}{1.\underbrace{0010\ 0000\ 0000\ 0000\ 0000\ 000}_{\text{Mantissa}}}$$

$$= \frac{1.0100\ 1011\ 0011\ 0011\ 0011\ 011}{1.001}$$

```
        1.0010011 0011 0011 00 11
1001 ) 1010 0101 1001100 11 0011 001
       1001
       0001010
          1001
       0001110
          1001
          01010
           1001
           0001
```

Ans ⇒ (IEEE format)

0 0000 0010 001001 ....

$\underbrace{0000\ 0010}_{\text{Decimal value} = 2}$ Mantissa

| 127+3 |
| 127+1 |
| 2 |
| $\boxed{\therefore x = 2}$ |

Ans

$$[1.\ \text{Mantissa}] \times 2^x$$

$$= [1.0010011001100110011... \times 2^2$$

$$= 100.10011001100110011$$

$$= 4 + \frac{1}{2^1} + \frac{1}{2^4} + \frac{1}{2^5} + ...$$

$$= 4.59375 .....$$

Answer

Decimal

Decimal value

Note: If quotient is not in scientific format then we have to make quotient in scientific format. Accordingly we have to adjust the exponent.

1. 0010011 0011 0011 0011

1001] 1010.010 1100 1100 1100 11001

1001

0001 010

1001

0 001 110

1001

x1010

1001

xxx1 110

1001

1010

1001

xxx1 110

1001

x1010

1001

0001 110

1001

01010

1001

00011