

Ans

0 100 0011 0101 0100 0000 0000 0000 0000

↓ sign bit

↓ exp bits

↓ 23 bits fraction

MT2019S19
ROSHANALI

Sign single precision is considered

$$\begin{aligned} \text{exp. bias} &= 127 \\ \text{exp. value} &= 134 \\ \text{so, } e &= 134 - 127 = 7 \end{aligned}$$

$$m = 2^{-1} \times 1 + 2^{-2} \times 0 + 2^{-3} \times 1 + 2^{-4} \times 0 + 2^{-1} \times 1 + \dots$$

$$(0) \times (2^{-6} + \dots + 2^{-23})$$

$$= 0.65625$$

The decimal value

$$(-1)^0 \times (1 + 0.65625) \times 2^7 = 212$$

So, if the value of m changes or bits in fraction part change the decimal value is changing so,

Fraction part defines the precision.

As, even when we look at single & double precision,

single precision has 23 fraction bits which give

$$(\log_{10}(2^{23}) = 7) \quad 7 \text{ decimal digit of accuracy.}$$

double precision has 52 fraction bits gives

$$(\log_{10}(2^{52}) = 15.654) \quad \text{nearby } 16 \text{ decimal digits of accuracy.}$$

As, smallest change that can be represented in floating point is called precision so,

even from this fraction part defining the precision.

Discussion:

The decimal value

$$= 218 = 128 \times (200.0 + 1) \times 2^{-8}$$

So, if the value of α in the change of bits in fraction part is changing, so

fraction part defines the precision.

single iteration for 25 Fourier sine wave pairs

2nd normal and subnormal values go per IEEE standard

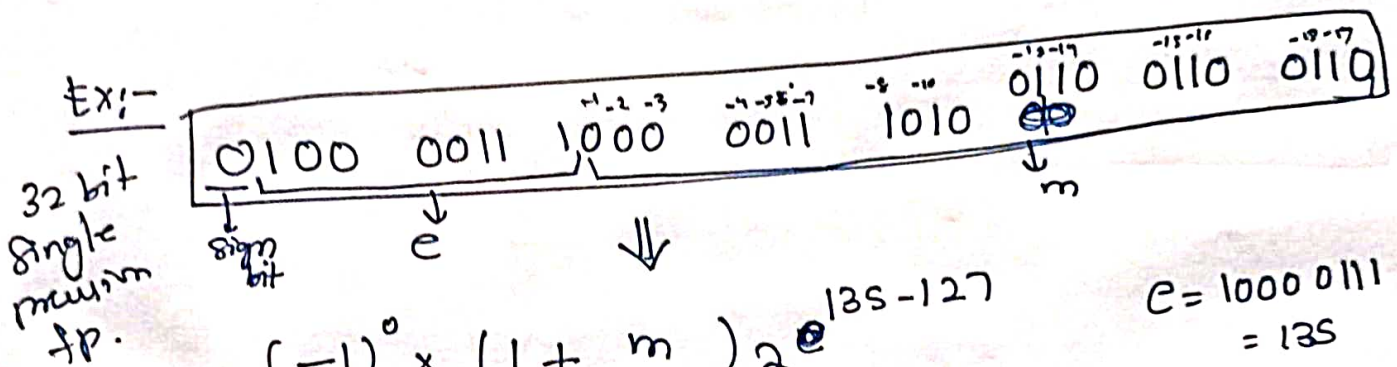
normal value

1. '1' before decimal point

ex:- single precision



$$F = (-1)^s \times (1 + m) \cdot 2^{e-127} \quad \left[\begin{array}{l} \text{bias} = 127 \\ e \rightarrow \text{range form} \\ 1 \rightarrow 255 \end{array} \right]$$



$$(-1)^0 \times (1 + m) \cdot 2^{135-127}$$

$$e = 10000111 = 135$$

$$(-1)^0 \times (1 + 0.2851) \cdot 2^8$$

$$m = 1 \times (2^{-5} + 2^{-6} + 2^{-7} + 2^{-8} + 2^{-9} + 2^{-10} + 2^{-11} + 2^{-12} + 2^{-13} + 2^{-14} + 2^{-15} + 2^{-16} + 2^{-17} + 2^{-18} + 2^{-19})$$

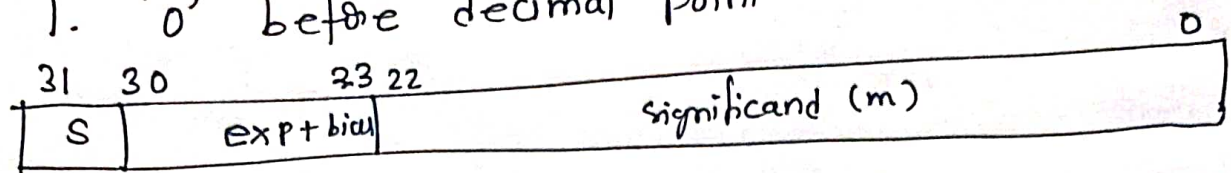
$$F = 263.3$$

$$\downarrow$$

1.000001110100110... $\times 2^8$

sub normal value

1. '0' before decimal point



$$F = (-1)^S \times (0 + m) \cdot 2^{-127+1}$$

$$= (-1)^S \times (m) \cdot 2^{-126}$$

~~exp = 0~~
exp = 0

$m \neq 0$
(significand)

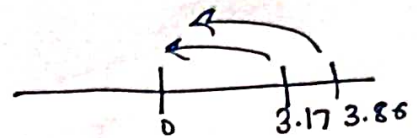
3rd

Five methods defined by IEEE for rounding floating point numbers

1. Rounding towards zero
2. Round down ($-\infty$)
3. Round up ($+\infty$)
4. Round to nearest
5. Round to even

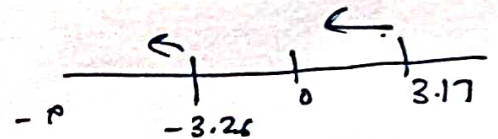
1. Round towards zero

$$\begin{aligned}\text{Ex:- } 3.17 &\Rightarrow 3 \\ 3.86 &\Rightarrow 3\end{aligned}$$



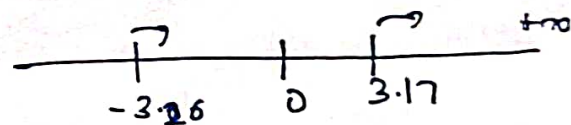
2. Round down ($-\infty$)

$$\begin{aligned}\text{Ex:- } 3.17 &\Rightarrow 3 \\ -3.26 &\Rightarrow -4\end{aligned}$$



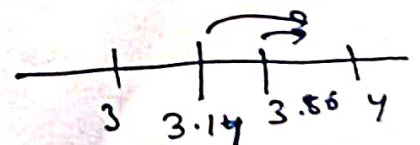
3. Round up ($+\infty$)

$$\begin{aligned}\text{Ex: } 3.17 &\Rightarrow 4 \\ -3.26 &\Rightarrow -3\end{aligned}$$



4. Round to even

$$\begin{aligned}\text{Ex:- } 3.17 &\Rightarrow 4 \\ 3.86 &\Rightarrow 4\end{aligned}$$

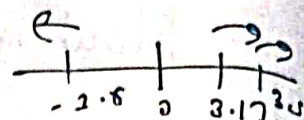


5. Round to nearest (ties away from zero)

$$3.17 \Rightarrow 3$$

$$3.5 \Rightarrow \text{don't know}$$

$$-3.6 \Rightarrow -4$$



1. Round to nearest
2. Round to even
3. Round to nearest
4. Round up (+)
5. Round down (-)



1. Round towards zero

$$\text{ex: } 3.17 \Rightarrow 3$$

$$3.88 \Rightarrow 3$$

2. Round down (-)

$$\text{ex: } 3.17 \Rightarrow 3$$

$$-3.88 \Rightarrow -4$$

3. Round up (+)

$$\text{ex: } 3.17 \Rightarrow 4$$

$$-3.88 \Rightarrow -3$$

4. Round to even

$$\text{ex: } 3.17 \Rightarrow 4$$

$$3.88 \Rightarrow 4$$

