

In Floating point representation, we have three components

- 1.The Sign Bit
- 2.Exponent
- 3.Fractional Part

Precision is one the prime attribute of any Floating Point Representation,

1.Does any of the above three components play a role in the defining the Precision of the number? If so which are the component or Components which play the role in defining precision and how? Explain this with example in your own words

Ans: Precision means the smallest change that can be represented in Floating point representation. Here fractional part will determine the precision of Floating point number. Fractional part is called Mantissa in floating point representation. For example, the number 2 can be represented in 4 bits as 0.002×10^3 or 0.200×10^1 or 2.000×10^0 .

Among these three types of decimal floating point representation, the last is most precised since there are three zeros to the right of 2, it says that if any extra error in actual result like 2.0009×10^0 will lead to only 0.018% error.

2. What is Normal and Subnormal Values as per IEEE 754 standards explain this with the help of number line.

Ans: Normal representation doesn't have many leading zero's, but subnormal representation has minimum number in its exponent value which leads to more zero's in its mantissa. For example: 0.05 decimal value can be represented in binary by 2 ways:

- (i) Normal Representation: 1.01×2^{-6}
- (ii) Subnormal Representation: 0.00101×2^{-3}

3. IEEE 754vv defines standards for rounding floating points numbers to a representable value. There are five methods defined by IEEE for this – Take time and understand what these five methods and explain it in your words using diagrams, illustrations of your own.

Ans: IEEE754 standard defines five rounding rules:

- (i). Rounding to nearest, ties to even: In this method, real number is rounded off to the nearest even number.

For example: 5.5 is rounded off to 6.0

6.5 is rounded off to 6.0

-5.6 is rounded off to -6.0

- (ii). Rounding to nearest, ties away from zero: In this method, real number is rounded off to the nearest integer number. If a real number falls in the middle of two integers, it is rounded to the nearest value above (for positive numbers) or below (for negative numbers).

For example: 5.5 is rounded off to 6.0

5.9 is rounded off to 6.0

-5.6 is rounded off to -6.0

- (iii). Round towards zero: In this method real number is truncated to the nearest integer while going towards zero.

For example: 5.5 is rounded off to 5.0

5.9 is rounded off to 5.0

-5.6 is rounded off to -5.0

- (iv). Round toward $+\infty$: In this method real number is truncated to the nearest integer while going towards to $+\infty$.

For example: 5.5 is rounded off to 6.0

5.9 is rounded off to 6.0

-5.6 is rounded off to -5.0

- (v). Round toward $-\infty$: In this method real number is truncated to the nearest integer while going towards to zero.

For example: 5.5 is rounded off to 5.0

5.9 is rounded off to 5.0

-5.6 is rounded off to -6.0