Course Name: A Level (1st Sem) **Subject: CO**

Topic: Number Representation Contd. **Date: 23-04-20**

So far, we have learnt representation of Fixed Point numbers.
Now, we are going to learn representation of **Floating Point numbers**.

Floating Point Numbers: A Floating Point number is the number that has a decimal point placed somewhere in between of the number. These numbers are called Floating Point because the decimal point floats freely in the number. When we multiply the number by 10, the point floats one digit right. Similarly when we divide that number by 10 the point floats one digit left.
We have the following hierarchy of representations for Floating Point numbers:

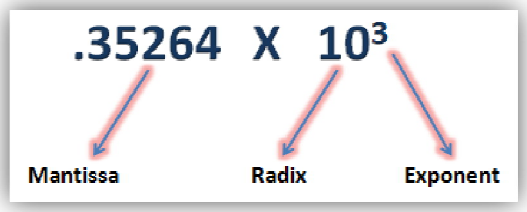
1. Positive Numbers
   - IEEE 754 32-bit representation
2. Negative Numbers
   - IEEE 754 32-bit representation

Here in the case of Floating Point numbers, there is only one method of representation i.e. IEEE-754. The same method is used for both positive and negative numbers. Only the sign bit changes.

IEEE-754 32 bit representation: A floating-point number is typically expressed in the scientific notation, with a mantissa [m], and an exponent [e] of a certain radix [r], in the form of [m x r^e].

Decimal numbers use radix of 10 [m × 10^e]; while binary numbers use radix of 2 [m × 2^e].

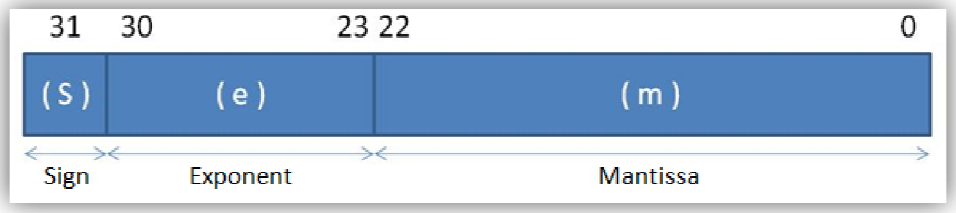
For ex. 352.64 can be represented as-



{ Here, since decimal point is placed 3 digits left hence it produces a power of 3 to 10 }

Normally, IEEE-754 standard is used for representation of Floating Point Numbers in 32 bits. In this floating-point representation:

1. The most significant bit is the sign bit (S), with 0 for positive numbers and 1 for negative numbers.
2. The following 8 bits represent exponent (e).
3. The remaining 23 bits represent mantissa (m).



NOTE: The example of this representation will be discussed tomorrow. Till then observe the rules and theory of this method.