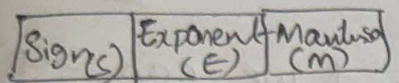
7) Floating point representation :-

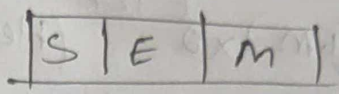
There are various architecures and each of them has their own representation. To avoid confusion all should follow a single representation. One of the famous representation is IEE754.

Single Precision (32 bit representation)



1 bit 8 bits (23 bits)

Double Precision



1 bit 11 bits 52 bits

Standard Conversion:-

$$(-1)^{S} \times (1+M) \times 2^{E\text{-bias}}$$

S - Sign
E - Exponent
M - Mantissa.
bias = 127 (for Single precision)
bias = 1023 (for double precision)

from this conversion we can say that Mantissa plays a vital role.

Eg:- $\frac{1}{3} = 0.333333$

S	E	M
0	01111101 (1021)	01010101010101010101011

→ 3EAAAAAB.

→ $(-1)^{0} \times (1+0.333322) \times 2^{1021-1023} = 0.333333$ (for S.P)

for double precision ⇒ 0.3333333333.

→ precision increases with increase in mantissa. So double precision is more precise than single precision.