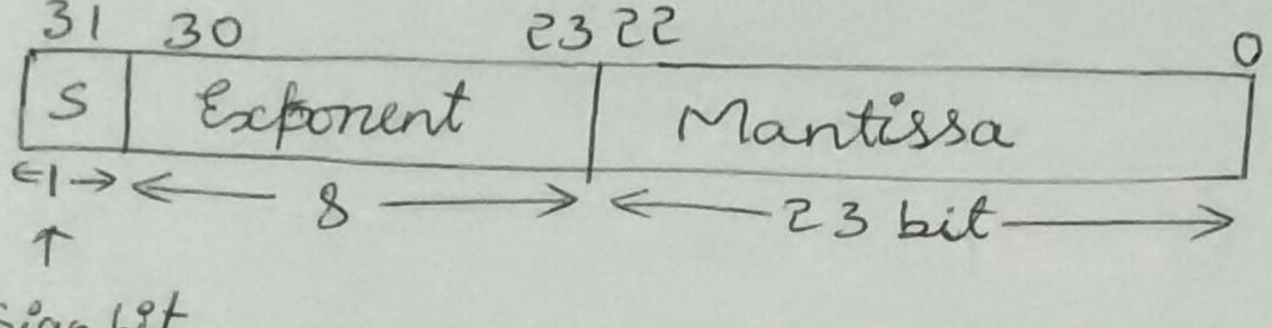
Floating point data type

- Float (32 bit)
- Double (64 bit)

Float 32 bit Format (Single Precision)

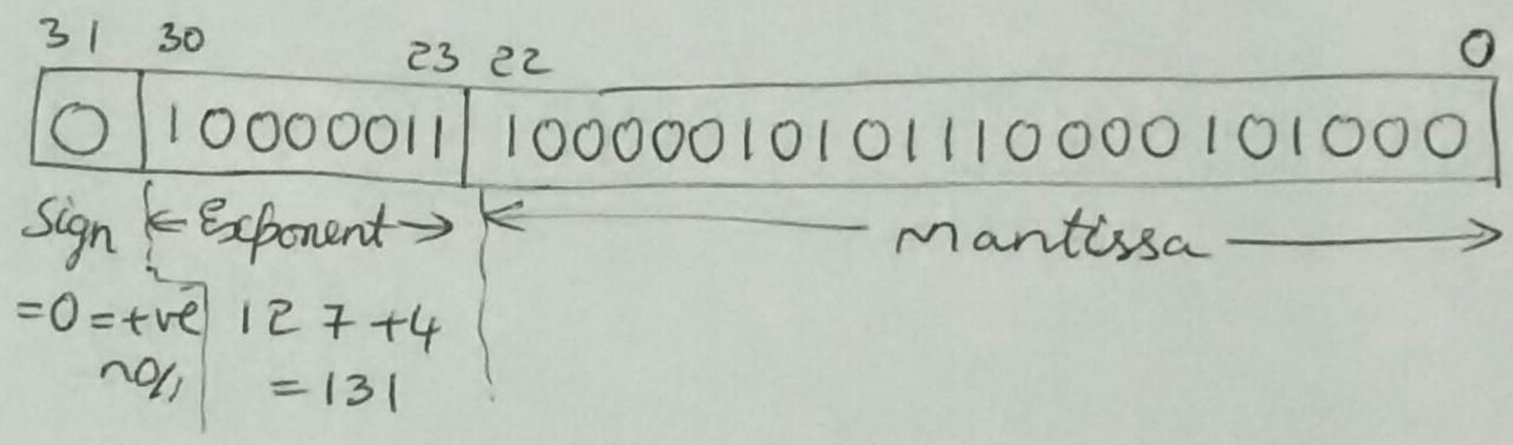


Sign bit
1 → -ve number
0 → +ve nuber

Ex: float a;
a = 24.17;

+ 24.17 =

For -ve no's
Make S=1



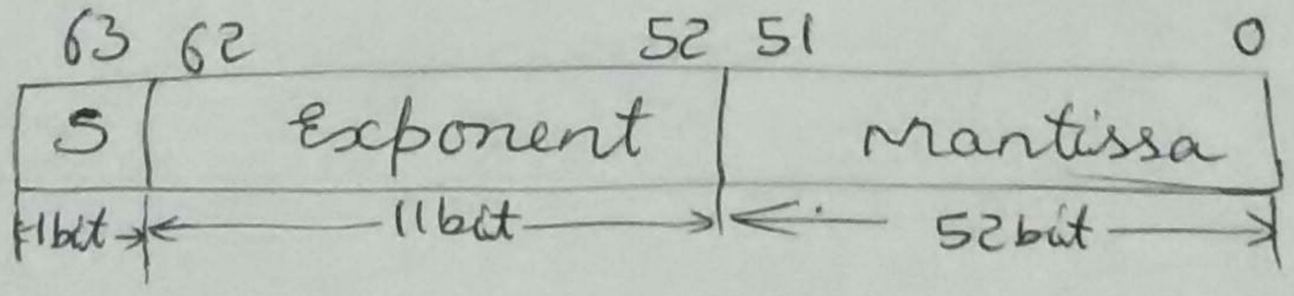
24.17

Excess = 127

11000.00101011110000101000
1.10000010101111000010100 0 $\times 2^4$ $\Rightarrow \begin{pmatrix}127\\+4\end{pmatrix}$ = Exponent
= 131

mantissa

Double 64 bit Format (Double Precision)



S=0 → +ve no's
S=1 → -ve no's

Exponent = (1023 + power of 2) Excess = 1023

Mantissa is the fraction part