

1) The three components are

- 1) sign bit S
- 2) Exponent (E)
- 3) fractional part (M)

Number is $F = (-1)^S M \times 2^E$

Two kinds of encoding done

- i) single precision (32 bits, E 8 bits, M 23 bits)
- ii) double precision (64 bits, E 11 bits, M 52 bits)

1 sign bit for each type

⇒ The precision depends on fractional part.

⇒ The number of significant digits depends on the number of bits in M

when $M = 24$ bits

we have x significant digits

where $2^{24} = 10^x \Rightarrow x = 7.2 \dots$

⇒ $x = 7 \Rightarrow$ up to seven significant

decimal places we can represent in binary floating point representation.

⇒ The range of number depends on the number of bits in exponent that is $2^8 - 1 = 255$

⇒ The exponent is encoded as a biased value

$E = \text{exp} + \text{bias}$ where $\text{bias} = 127 = (2^{8-1} - 1)$ for single precision

$\text{bias} = 1023 (2^{11-1} - 1)$ for double precision

→ when the number's ^{decimal part} lies between 0 to $\frac{1}{2^{52}}$ it can be represented accurately using double precision representation.

→ when number is

$$1.x$$

if $0 < x < \frac{1}{2^{23}}$

and it's represented in single precision, it has 100% accuracy

→ when number is

$$1.x$$

if $0 < x < \frac{1}{2^{52}}$

and it's represented in double precision, it has 100% accuracy.