Normalised floating point Binary fractions.

Example:-

Converting Binary to decimal

1. $\overbrace{0111}^{M}\overbrace{0011}^{E}$ 4 bit mantissa

$0.111\ 0011$ 4 bit Exponent

$0.111 \times 2^{3} \rightarrow 0111.0 \rightarrow 7$

$\Rightarrow 01110011_2 = 7_{(10)}$

2. 01111110

$0.111 \times 2^{-2} \rightarrow 0.00111 \rightarrow 0.21875_{(10)}$

Converting Decimal to Binary

4 bit Mantissa and 4 bit Exponent

1. $0111.0 = 7.0$

$0.111 \times 2^{3} \rightarrow 0.111\ 0011$

$01110011 \rightarrow 7_{(10)}$

2. $0.25 \rightarrow 0.01$

$0.100 \times 2^{-1} \rightarrow 0.100\ 1111$

$01001111 \rightarrow 0.25_{(10)}$

Converting Negative numbers

$-6.0 \Rightarrow 1010.0$

$1.010 \times 2^{3} \rightarrow 10100011$

i) 01001111

0.100 1111

0.100×2^{-1}

0.0100 ⇒ 0.25

ii) 00010001

0.0010001

0.001×2^{1}

00.01

⇒ 0.25

iii) 00100000

0.010 0000

0.010×2^{0}

0.010

⇒ 0.25

By considering the point immediately to the right of MSB in a fixed sized of 'M' we get best range and precision.

→ For +ve values the normalised form start with a '0' followed by a '1'.

→ For -ve values the normalized form start with a '1' followed by a '0'.

00010001

0.001 0001

0.001×2^{1}

$000.1 \times 2^{1-2=-1}$

0.100×2^{-1}

0100 1111

Floating point binary addition

- → Make both numbers are normalized.
- → Same exponents.
- → Add Mantissas.
- → Normalise result.

Considering 6 bit 'M' ; 4 bit 'E'

4 ↖ 010000 0011 + 010010 0010 ↗ 2.25

$0.10000 \times 2^3 + 0.10010 \times 2^2$

$0.10000 \times 2^3 + 0.01001 \times 2^3$

0.11001×2^3

0110010011 → $6.25_{(10)}$

Truncation Error