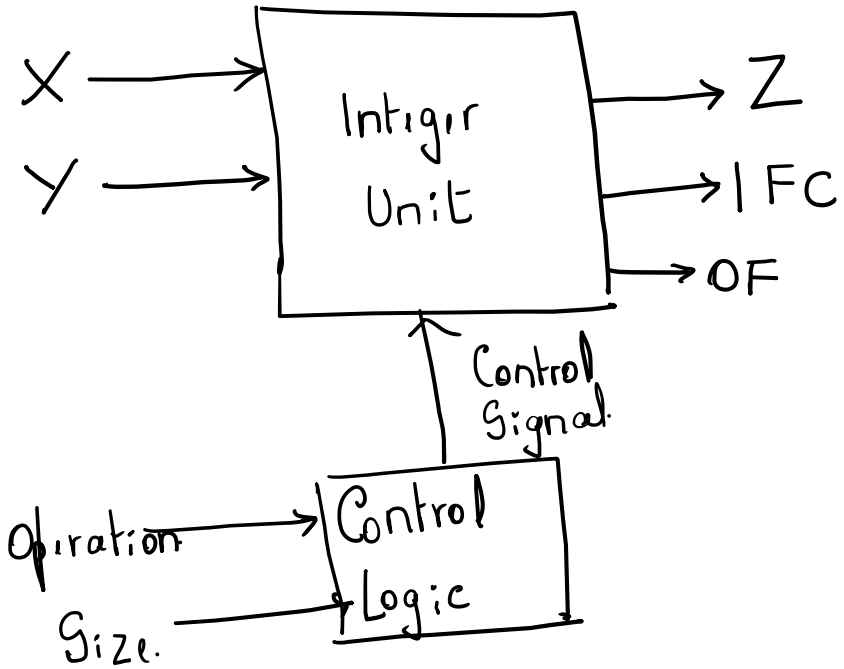
INTEGER UNIT

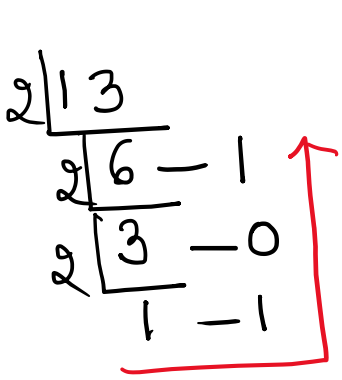


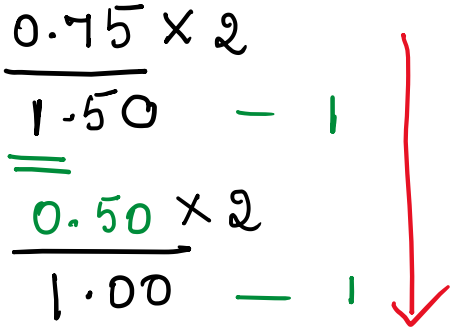
- Adder | Subtractor.
- Unsigned multiplier.
- Signed multiplier
- Unsigned divider
- Signed divider.

- Real valued data

Floating Point Numbers and Operations

Binary format





$13_{(10)} \rightarrow 1101_{(2)}$

$0.75_{(10)} \rightarrow 0.11_{(2)}$

Fixed Point Representation

- Assume there exists a binary point

Unsigned

$13 \longrightarrow 1101.$ ↑ Implicit binary point to the right of LSB.

$0.75 \rightarrow 0.11$

Signed fraction

Positive. $0.75 \longrightarrow 0.11$

Negative $-0.75 \xrightarrow[0]{} 2^1$'s complement of 0.11

$$\begin{array}{r} 1.00 \\ \quad\;\; 1 \\ \hline 1.01 \end{array}$$

$$-0.75_{(10)} \longrightarrow 1.01_{(2)}$$

Generalizing,

$$B = b_0 \cdot b_{-1} b_{-2} \cdots b_{-(n-1)}$$

Value of B is given by,

$$F(B) = -b_0 \times 2^0 + b_{-1} \times 2^{-1} + \cdots + b_{-(n-1)} 2^{-(n-1)}$$

Ex:

$0.75 \rightarrow 0.11$ $\quad F(B) = -0 \times 2^0 + 1 \times 2^{-1} + 1 \times 2^{-2}$

$$= 0 + \frac{1}{2} + \frac{1}{4}$$

$$= 0.5 + 0.25$$

$$= 0.75$$

$$-0.75_{(10)} \rightarrow 1.01_{(2)}$$

$$F(B) = -1 \times 2^{0} + 0 \times 2^{-1} + 1 \times 2^{-2}$$

$$= -1 + 0 + \frac{1}{4}$$

$$= -1 + 0.25$$

$$= \underline{-0.75}$$

In general → n-bits

- i bits can be used for whole part
- f bits can be used for fraction part.

$$\boxed{n = i + f}$$

$$13.75_{(10)} \longrightarrow \underbrace{1101}_{i=4} . \underbrace{11}_{f=2}$$

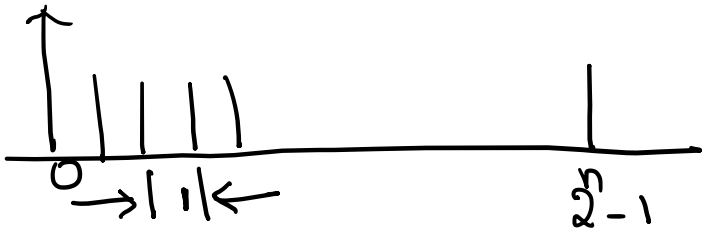
32-bit unsigned fixed-point format

$n = 32$

$i = 32 \quad f = 0$ } Binary point to the right of LSB

Range : 0 to $2^{n} - 1$

Resolution : 1 } Difference between successive numbers



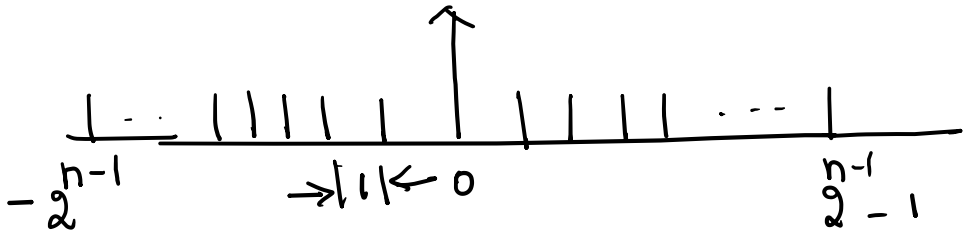
32 bit signed fixed-point

$n = 32$

(1) $i = 32$ $\quad f = 0$

Range: $-(2^{n-1})$ to $(2^{n-1} - 1)$

Resolution: 1



Range: -2^{31} to $2^{31} - 1$

0 to $\pm 2.15 \times 10^{9}$ } Looks like large number

Decimal

(2) $i = 0$ $\quad f = 32$: fraction

- Binary point is assumed to be → left of MSB

Decimal nos: $\pm 4.55 \times 10^{-10}$ to ± 1

Very small numbers

6.0247×10^{23} $mole^{-1}$ → Avogadro's number.

6.6254×10^{-27} erg-s → Planks constant

Range → 32 bit Signed fixed-point representation

↓

May not be sufficient

↓

Scientific computation.

Compiler → Position of Binary point → Variable

↓

Binary point is said to float

Floating point Numbers

↓

Representation.

Range and Resolution

n = bits $n = i + f$

Unsigned

Range : 0 to $2^{i} - 2^{-f}$

Resolution: 2^{-f}

$n = 3$

Binary

$i = 3$, $f = 0$ } Range : 0 to $2^{3} - 2^{-0}$

0 to $2^{3} - 1$

Resolution: $2^{-0} = 1$

$i = 2$, $f = 1$ } Range: 0 to $2^{2} - 2^{-1}$

Resolution 2^{-1}

$i = 1$, $f = 2$ } Range : 0 to $2^{1} - 2^{-2}$

Decimal

0 to $10^{3} - 10^{-0}$

0 to $10^{3} - 1$

0 to 999

$10^{-0} = 1$

0 to $10^{2} - 10^{-1}$

0 to $100 - 0.1$

0 to 99.9

10^{-1}

0 to $10^{1} - 10^{-2}$

0 to $10 - 0.01$

0 to 9.99

Resolution: 2^{-2} 10^{-2}

$i = 0$

$f = 3$

Range: 0 to $2^{0} - 2^{-3}$ 0 to $10^{0} - 10^{-3}$

0 to $1 - 0.001$

0 to 0.999

Resolution: 2^{-3} 10^{-3}

Scientific notation: Decimal

253.75×10^{0}

25.375×10^{1}

2.5375×10^{2}

0.25375×10^{3}

0.025375×10^{4}

Same number.

0.025375 → Mantissa. Significand.

4 → Exponent

-25.3×10^{-2}

Sign: − → Significand, Mantissa.

−2 → Exponent. Negative exponent

$13.75 \longrightarrow 1101.11 \times 2^{0}$

110.111×2^{1}

11.0111×2^{2}

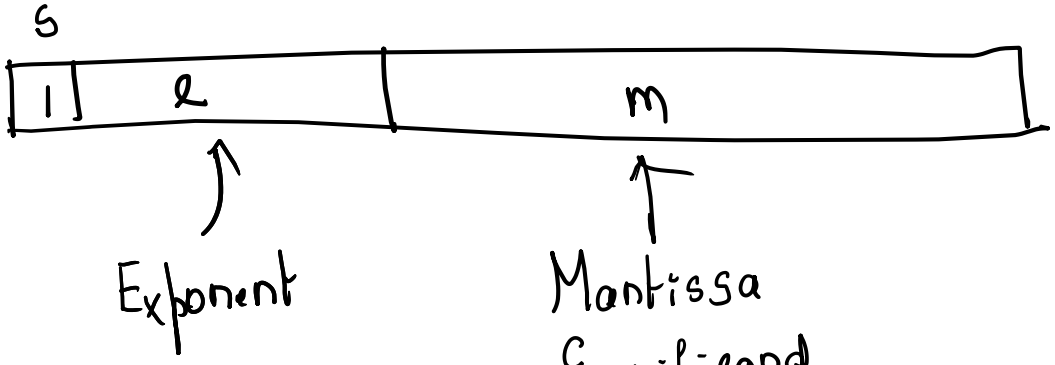
1.10111×2^{3}

0.110111×2^{4}

13.75

Exponent = -1

Significand



$$n = 1 + e + m$$

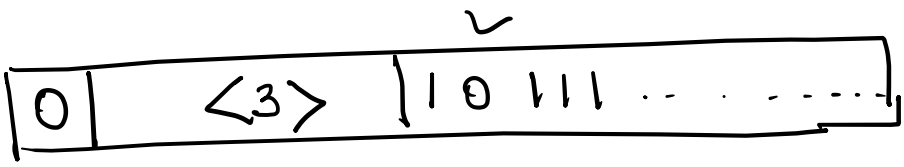
Normalized representation

— Ensure that there is only one bit to the left of the binary point and that bit is always 1

$13.75 \longrightarrow 1101.11 \times 2^{0}$

↓ Normalize.

1.101111×2^{3}



0.875 × 2

0.750 × 2 — 1

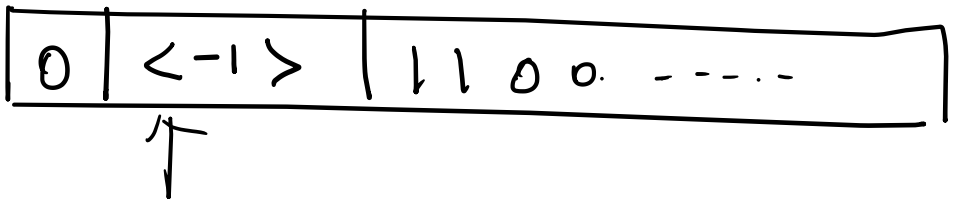
1.500 × 2 — 1

1.000 — 1

$0.875 \longrightarrow 0.111 \times 2^{0}$

Normalize.

1.11×2^{-1}



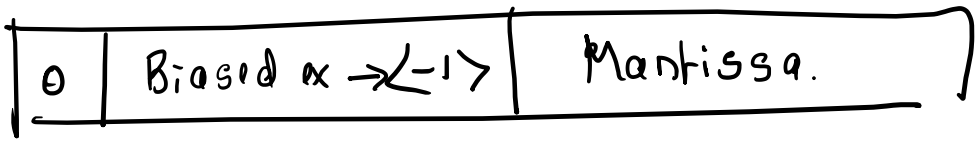
Signed number.

Biased exponent

↓

True exponent in Excess-K format

$1.11 \times 2^{(-1)}$ True exponent


0
Biased ex →<-1>
Mantissa.