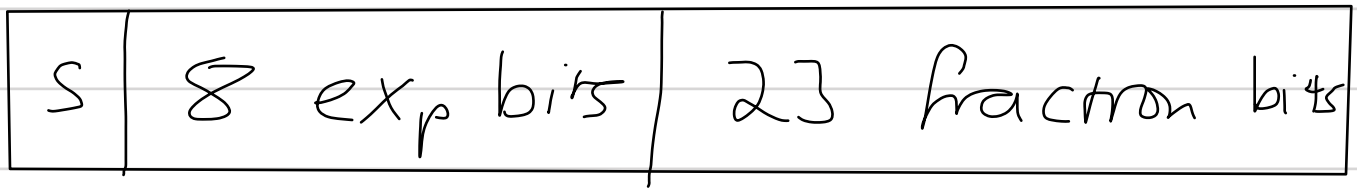
IEEE Floating point

First proposed in 1985; Now present in every computing platform. The format used to store fractional/real (IR) data on fixed/permanent storage.



1 Sign bit
8 Exponent bits (Q7.0 signed*)
23 Fractional bits (Q0.23 unsigned)

Exponent has a bias of -127_{10} so:

$E_{min} = 0 - 127 = -127$

$E_{max} = 254 - 127 = +127$

So offset of 127 is subtracted from the exponent value.

Exponents of 0x00 and 0xFF are used differently:

Exponent	Mantissa 0	Mantissa Non 0	Equation
0x00	+/- 0	sub-normal	$(-1)^S \times 2^{-126} \times 0.FB$
0x01→0xFE	Normalised		$(-1)^S \times 2^{E-127} \times 1.FB$
0xFF	+/- ∞	NaN	Quiet signalling.

$$= +\infty$$

$x : \infty \to 0$

$$\lim_{x:-\infty\to 0} \frac{1}{x} = -\infty$$

$$1_{10} = (-1)^0 \times 1.(00\ 0) \times 2^{E-127}$$

$E - 127 = 0$

$\therefore E = 127 = 01111111$

0 01111111 00000 ... 00

3 F 8 0 0 0 0 0

$\therefore 1_{10} = 3F800000$ IEEE fp.

C6A41010

1100 0110 1010 0100 0001 0000 0001 0000

$S = 1$

$E = 10001101 = 141$

$\therefore E - 127 = 14$

From definition.

1.01001000001000000001

$= 1 + 2^{-2} + 2^{-5} + 2^{-11} + 2^{-19}$

$= 1.281740189$

$(-1)^1 \times 2^{14} \times (1 + 2^{-2} + 2^{-5} + 2^{-11} + 2^{-19})$

$= -1 \times (2^{14} + 2^{12} + 2^{9} + 2^{3} + 2^{-5})$

$= -21000.03125$