Floating Point Representation

9.9×10^2
9.8×10^2
19.7

4.3420
10.340

4.342
10.340
14.682

4.342×10^2
$10.340 \times 10^{-2} \times 10^4$

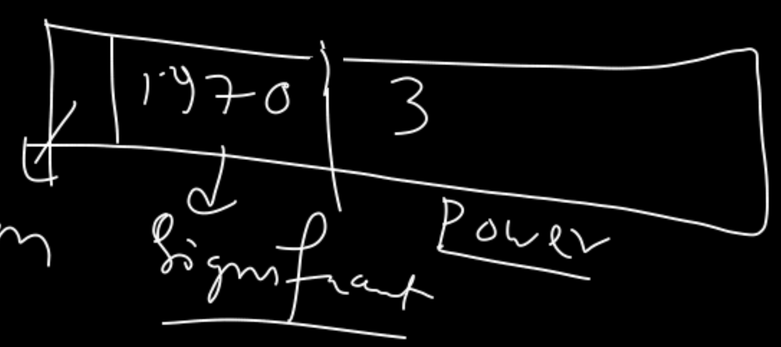
10^4

Converted to match Exp: 0.00010340×10^2
4.3420×10^2

4.34310×10^2 =

19.7×10^2
$- 4.34 \times 10^3$

1.97×10^3



-19.5

$(4.25)_{10}$

0.25

0.01

$= 10 \times 10^{-2}$

13.5 $= (100.01)_2$

1.0001×2^{2}

32-bit

0	1000 0001	00010000 ~ 0

sign (1-bit) ↓ power (8-bit) ↓ (23-bit) fraction / Mantissa / Significant

$2 + 127 = 129$

$\downarrow$

1000 0001

IEEE standard

bare - 127

① Convert it to binary
② Convert it to Normal form
③ Convert exponent to bare 127
④ put the sign + Exp + fraction

19 → 10011.1

100111 × 2^4

127 + 4 = 131

1	10000011	00 1110 0 . . 0

18-bit

254 —

32-bit = | 0 | 1111 1110 | 1111111 - - 1 |

8 exponent, 23-bits

$$1.11 + 0.01 = 10.00 = 1.00 \times 2^{1}$$

$$1.11111 + 0.00001 = 10.00000$$

$2 \quad 2^{-5}$

$$\text{Decimal value} = 1.\underbrace{111111 \ldots 11}_{23} \times 2^{254-127}$$

$$= (1.11\ldots + 2^{-23} - 2^{-23}) \times 2^{127}$$

$$= (2 - 2^{-23}) \times 2^{127}$$

$$\pm 1.f \times 2^{exp}$$

+∞ = | 0 | 1111 1111 | 0 0 0 0 0 . . all 0 |

−∞ = | 1 | 1111 1111 | 0 0 . . all 0 |

NaN = | 0 | 1111 1111 | ≠ 0 |

NaN = | 1 | 1111 1111 | ≠ 0 |

127 + actual value

Minimum →

$1.f \times 2^{exp}$

0	0000 0001	0 0 0 0 0 0 0 … 23
		0 0 0 — — 1

$1.0 \times 2^{1-127} = 2^{-126}$

$(0.000 - - 1)_2$ — 126

0	0000 0000	1 1 1 1 1 1 1 1 1 1 1

0 - 127

$1.1111 - - \times 2$ #

(2^{-23})

$(2^{-126} - 2^{-23})$