Single precision IEEE floating point (32 bits)

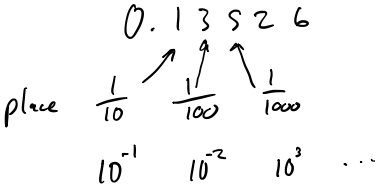
- First bit is the sign: 1 is negative, 0 is non-negative
- Next 8 bits represent the exponent in excess 127 notation (stored value is 127 larger than the exponent)
- Remaining 23 bits represent the fraction (mantissa), excluding the 1 to the left of the radix point
- Positive or negative infinity: exponent is all 1's and the fraction is 0
- Not-a-number (NaN): exponent is all 1's and fraction is anything non-zero

s - sign bit
f - fraction representation
e - exponent representation

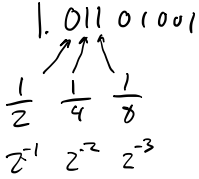
$$1.f \times 2^{e-127}$$

negative if s is 1

Decimal



Binary



- 62.125 to IEEE single-precision

$62_{10} = 111110_2$

$0.125_{10} = 0.001_2$

111110.001

Shift radix point 5 places → $1.11110001_2 \times 2_{10}^{5_{10}}$

Exponent is 5, add 127 → 132

$132_{10} = 10000100_2$

Fraction, drop the 1 to the left of radix point, pad with zeros on the right

1100 0010 0111 1000 1000 0000 0000 0000

0 x C 2 7 8 8 0 0 0

0.1

$$2^{-4} + 2^{-5} + 2^{-8} + 2^{-9} + 2^{-12} + 2^{-13} + \cdots$$

(0.0625)

0.0001100110011001100 ....

Exponent is -4, add 127 → 123

1111011

Fraction

10011001100110011001100