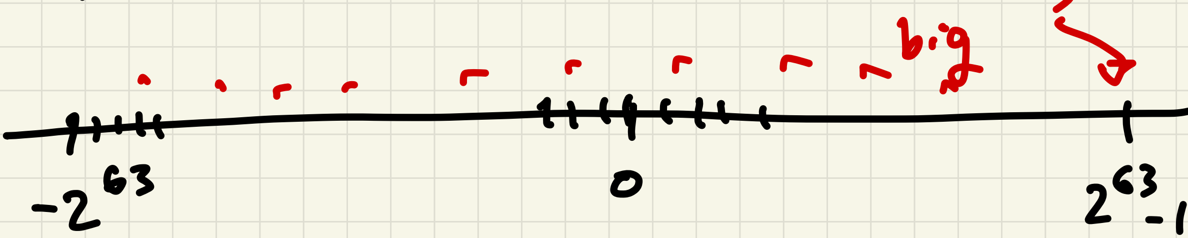
Floating-point Representation

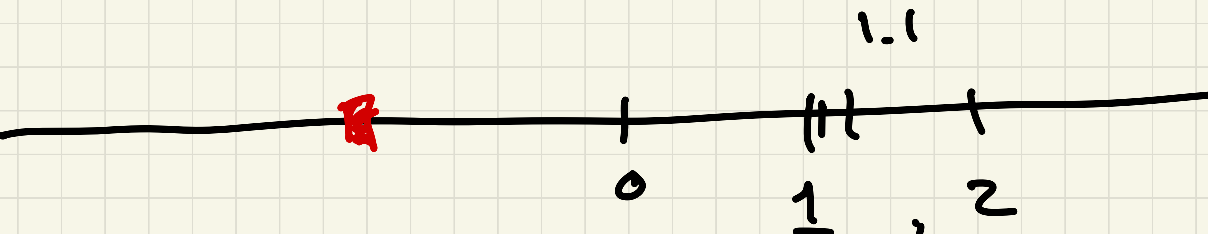
CS 111

Nov. 10, 2020

INTEGERS



REAL NUMBERS



1.01
1.001
1.0001
1.000001
⋮

Infinitely many reals, in <u>every</u> interval, no matter how small.

Scientific notation (base 10)

$$\pm\ \underbrace{6.02214}_{\substack{\text{mantissa}\\\text{(fixed \# digits)}}} \times 10^{29}$$

sign

base

exponent

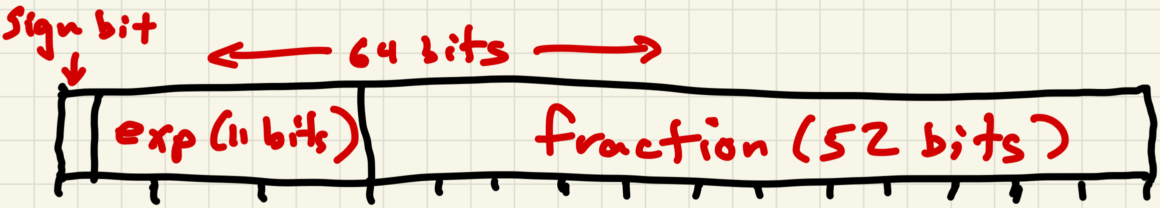
NORMALIZED $\equiv$ one digit before the decimal point

$$+602.214 \times 10^{26}$$

same number

not normalized.

IEEE Standard Float64



REPRESENTS THE NUMBER:

If $1 \le exp \le 2046$,

$$\pm(1.frac) \times 2^{exp-1023}$$

If $exp = 0$ and $frac = 0$,

$$0$$

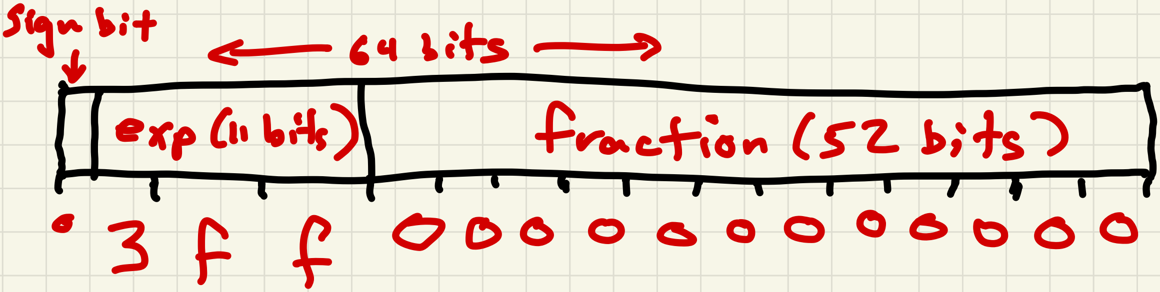
If $exp = 2047$ ($\equiv 7ff$) and $frac = 0$,

Inf

If $exp = 2047$ and $frac \neq 0$,

NaN (not-a-number)

1 : 3ff000--0



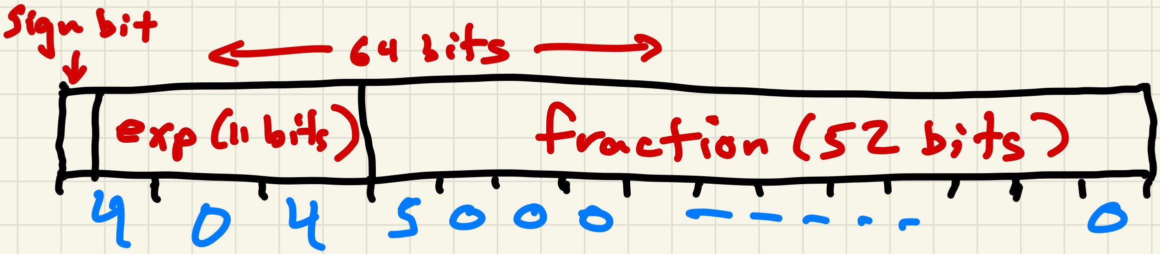
If $1 \le exp \le 2046$,

$$\pm(1.\underline{frac}) \times 2^{exp-1023}$$

$$+\ 1.\underline{000} \times 2^{1023-1023}$$

3ff

$$42 = 4045000\cdots0$$



If $1 \le exp \le 2046$,

$$\pm(1.\underline{frac}) \times 2^{exp-1023}$$

$$\begin{array}{cccccc} 32 & 16 & 8 & 4 & 2 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \end{array}$$

$$42_{10} = 101010_2 = +(1.\underline{01010}) \times 2^5$$

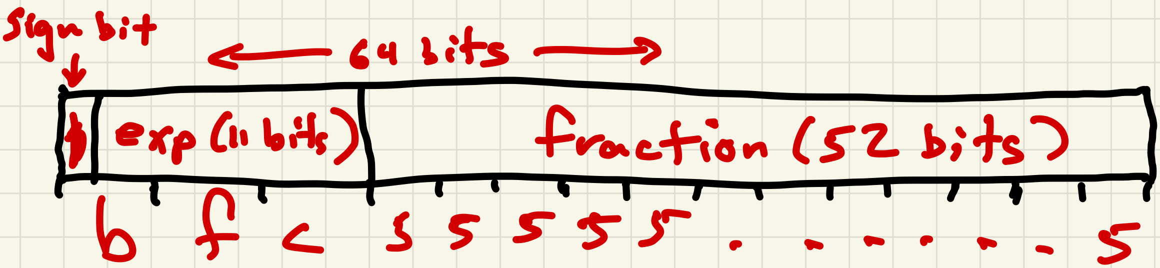
$$2^5 = 2^{1028-1023}$$

$$exp = 1028 = 404_{16}$$

$$frac = \underbrace{0101}_{5}\ \underbrace{0000}_{0}\ \underbrace{0000}_{0} \cdots$$

$$-1/3 = bfd555\cdots5$$

$$800 + 3fd = bfd$$



If $1 \le exp \le 2046$,

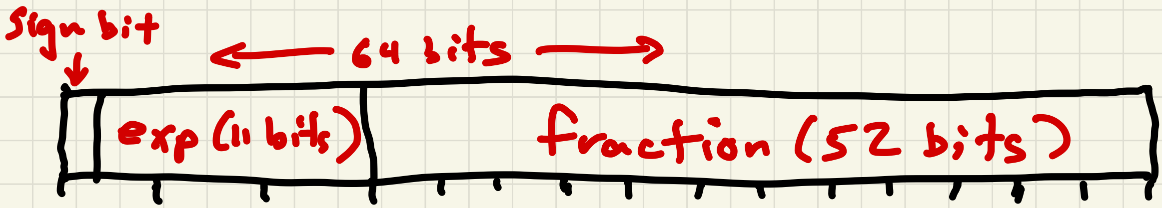
$$\pm(1.\underline{frac}) \times 2^{\underline{exp-1023}}$$

$$1/3 = 0.01010101\ldots$$

$$= 1.\underbrace{0101}_{5}\underbrace{0101}_{5}\underbrace{\cdots}_{5} \times 2^{-2}$$

$$-2 = 1021 - 1023$$

$$1021_{10} = \underbrace{3fd}_{exp}$$


Sign bit
64 bits
exp (11 bits)
fraction (52 bits)