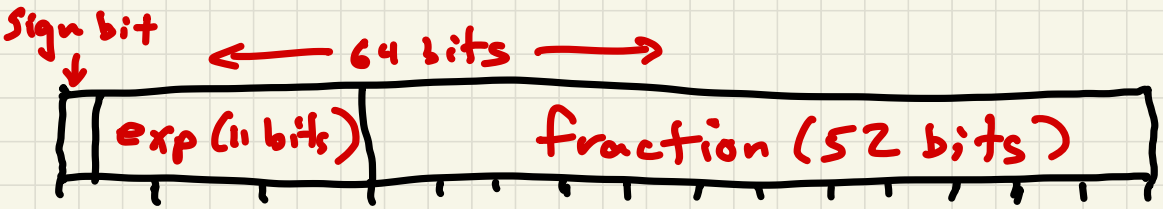
Properties of floating-point

CS 111

Nov. 10, 2020

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

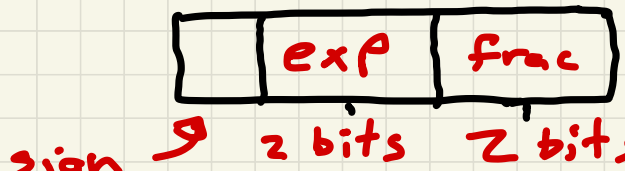
If $exp = 2047$ ($\equiv$ 7ff) and $frac = 0$,

Inf

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

NaN (not-a-number)

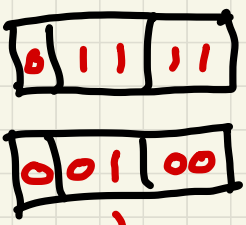
Toy 5-bit floating-point

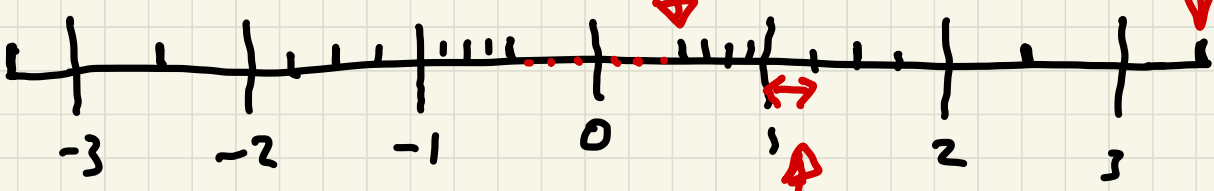


$1 \le exp \le 3: \quad \pm(1.frac) \times 2^{exp-2}$

max: $1.11 \times 2^{1} = 11.1 = 3\frac{1}{2}$

smallest normalized: $1.00 \times 2^{-1} = \frac{1}{2}$





machine epsilon = distance from 1 to the next larger floating-point number $= \frac{1}{4}$

Catastrophic Cancellation Example

$$(x-1)^7 = x^7 + 7x^6$$
$$+21x^5 - 35x^4$$
$$+35x^3 - 21x^2$$
$$+7x - 1.$$