NOTES: Floating Point

notes/floating-point.md

Floating Point

- The advantage of floating point is that it can represent a much larger range of numbers than fixed point—bits can be "allocated" to the integer and fractional parts as needed.
- IEEE 754 is the standard for floating point representation.

Representation

- Floating point is similar to scientific notation: ex. $1.001_2 \times 2^{10}$1.0012 × 210
- 32 bits allocated to a floating point number
  - 1 bit: sign bit (+/-)
  - 8 bits: exponent
  - 23 bits: mantissa
- A float is in **normalized representation** iff $exp \neq 0$exp □= 0 and $exp \neq 255$exp □= 255.

$$(-1)^s \times 1.man \times 2^{exp-127}$$

(−1)s × 1.man × 2exp−127

- Sign bit: negative if 1, positive if 0
- Mantissa: mantissa is the number that goes after the decimal point of the 1 in scientific notation
- Exponent: the power of 2 that the number is multiplied by
  - The exponent is **biased** by $2^{e-1} - 1$2e−1 − 1 (127 for 8 bits)
  - Subtract the bias from the exponent to get the actual value
    - 127 represents 0
    - 128 and above represent positive exponents
    - 126 and below represent negative exponents

Exceptions

- When exponent in bits is 255:
  - If mantissa is all 0s, then the number is ±∞±∞
  - If mantissa is not all 0s, then the number is NaN (Not a Number)
- When exponent in bits is 0:
  - If mantissa is all 0s, then the number is ±0±0
  - If mantissa is not all 0s, then the number is a denormalized number

Denormalized Representation

$$(-1)^s \times 2^{1-127} \times 0.man$$

(−1)s × 21−127 × 0.man

- Smallest number that can be represented in normalized form is 2^{-126}2−126
- Smallest number that can be represented in denormalized form is $2^{-126} \times 2^{-23} = 2^{-149}$ 2−126 × 2−23 = 2−149

Doubles

- Doubles are 64 bits
  - 1 bit: sign bit
  - 11 bits: exponent
  - 52 bits: mantissa