3

Tuesday, May 7, 2024 12:18 PM

special cases

Number	Sign	Exponent	Fraction
0	Any	00000000	00000000000000000000000
∞	0	11111111	00000000000000000000000
- ∞	1	11111111	00000000000000000000000
NaN	Any	11111111	non-zero
denormals	Any	00000000	$F \neq 0$ $(-1)^S \times 2^{-126} \times 0.F$

ISEE 754 to base 10

$$(-1)^S \times 2^{(E-127)} \times 1.F$$

Double and Half precision Floating points

Precision	Num. Bits	Num. Sign Bits	Num. Exponent Bits	Num. Fraction Bits	Bias	Minimum Positive Normal Value	Maximum Value	Minimum Positive Subnormal Value
Single	32	1	8	23	127	2^{-126}	$(2-2^{-23}) \times 2^{127}$	2^{-149}
Double	64	1	11	52	1023	2^{-1022}	$(2-2^{-52}) \times 2^{1023}$	2^{-1074}
Half	16	1	5	10	15	2^{-14}	$(2-2^{-10}) \times 2^{15}$	2^{-24}

- **Denormals** - have a smaller magnitude than the smallest normalized number
  - We don't implicitly assume the leading bit to be 1.

    $(-1)^S \times 2^{(126)} \times 0.F$
    - have an exponent part which is 0 and Mantisa different from zero
    - the range of numbers that can be represented [2**-149, (1-2**-23) * 2**-126]

Rounding

round 1.100101 (1.578125) to only 3 fraction bits

Down (floor)

1.100 Up (ceiling) # Up: 1.101

Toward zero (truncation)

1.100

To nearest (default) – If equidistant, round towards the one with 0 in the least significant position of the fraction

1.101 (1.625 is closer to 1.578125 than 1.5 is)

- **Addition**
  1. write them in the formula that we use to convert them to decimal format

     $(-1)^S \times 2^{(126)} \times 0.F$
  2. adjust the decimal points using the exponents to make them easy to add
     - it's better to adjust the one with smallest exponent so that we don't lose the most significant bit in the process
  3. add like terms

     To subtract use the same technique making sure the sign of the second number is right
- **Multiplication**
  - follow the addition step except instead of addition use multiplication

    Use same method for division

    eg. 1011×11

    1011

    1011

    100001

    *remark- we only take one exponent as a common factor after adjustment
    - multiply the mantisas → $1.F \times 1.F'$
    - Consider the sign.
- **Binary coded decimal (BCD) system** - 4 bits are used to represent a decimal digit from 0 to 9**.**
  - **-**example, 37 is written as 0011_0111