ARM Research Assignment - IMT2017507

Saturday, 17 October 2020 5:37 PM

1st Ques) Floating Point Representation:

There are various architectures (ARM, INTEL) and each Architecture has their own representation.

To avoid ambiguity, All these processor builders, (building companies) should follow Standerds.

One of the famous standerds for floating point Representation is IEEE 754.

SINGLE PRECISION:

- It is a 32-bit Representation

S	E	M
1 bit	8 bits	23 bits

DOUBLE PRECISION:

- It is a 64-bit Representation

S	E	M
1 bit	11 bits	52 bits

STANDERD CONVERSION:

SP DP

$$(-1)^S \times (1.M) \times B^{(E-127)} \text{ or } (E-1023)$$

(to decimal)

Where,

S → Sign

B → Base For Binary = 2

M → Mantissa

BY Seeing this expression we Can understand that we Can even Store smaller numbers.

Mantissa Plays an important role in defining Precision of a Number in Floating Point Representation.

Eg: $\frac{2}{3} = 0.66666666$ ----- (here 0.6666..67 is more Precise than 0.66)

This above example Can have more precision if number of mantissa bits are more. Single Precision has 23 bits fractional part and double Precision has 52 bits fractional parts. More number of fractional bits will result in avoiding loss of information of the number. So in double precision, precision is mattered more.

Eg: $2/3 = 0.6666---67$

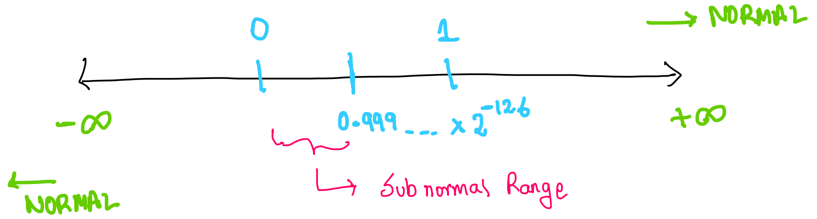
SINGLE PRECISION:

S	E	M
0	01111110	01010101010101010101010
1 bit	8 bits	23 bits
↓	↓	↓
+ve Num	Exponent field = 126 − 127 = −1	M = 0.3333333

DOUBLE PRECISION: (MORE PRECISION)

S	E	M
0	01111111110	0101010101 - - - - - - - - 0101
1 bit	11 bits	52 bits
↓	↓	↓
+ve Num	Exponent field = 1022 − 1023 = −1	M = 0.33333333 --- 3

2 Ques) NUMBER LINE:



NORMAL NUMBERS:

Format of Normal Numbers: $(-1)^S \times (1.M) \times 2^{E-127}$

Here the bit before the decimal point is always 1.

SUBNORMAL NUMBERS:

In Short, Numbers Smaller than NORMAL NUMBERS ARE SUBNORMAL NUMBERS.

Format of Sub Normal Numbers: $(-1)^S \times (0.M) \times 2^{E-127}$

Here the bit before the decimal point is 0.

- From the Number line

SUBNORMAL NUMBER < Smallest Normal Number.

3 Ques) According to the IEEE 754vv Standerds, The five rounding methods are:

- Round to nearest even
- Round to nearest away
- Round Up (Round to infinity)
- Round down (Round to -infinity)
- Round to Zero.

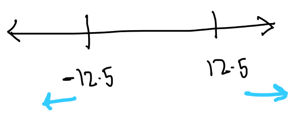
1. Round to nearest even: Rounded to the nearest possible value s.t if the number is the middle, moves to the even least Significant bit

Eg: 1.5 → 1.0 , -1.5 → -2

2. Round to nearest away: Rounded to the nearest possible value s.t it is rounded away from zero. (nearest Value above the num)

Eg: 12.5 → 13

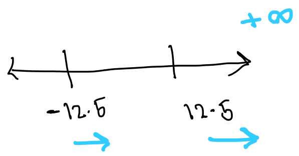
-12.5 → -13



3. Round Up: Rounded to the number larger than itself i.e number is rounded towards + ∞.

Eg: 12.5 → 13.0

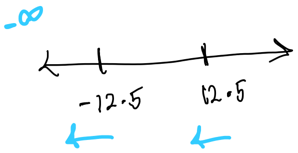
-12.5 → -12.0



4. Round down: Rounded to the number smaller than itself i.e number is rounded towards -∞

Eg: 12.5 → 12.0

-12.5 → -13.0



5. Round to zero: The number is rounded to zero.

Eg: 12.5 → 12.0

-12.5 → -12.0