Problem-1

1. Does any of the

1.The Sign Bit
2.Exponent
3.Fractional Part

three components play a role in the defining the Precession of the number ? If so which are the component or Components which play the role in defining precession and how ? Explain this with example in your own words

Answer:

The smallest change that can be represented in floating point representation is called as precision. The Fractional part plays a role in defining precision of the number. The fractional part of the single precision has 23 bits for resolution. Whereas, in double precession we have 52bits.

The Sign bit is used to know whether the number is positive or negative and Exponent is used the define the range.

Example -

Let's consider a number 0.11874245533534 which is

0	**01111011**	**11100110010111100111111**
Sign	Exponent	Fraction

In single precession and in double precision it is

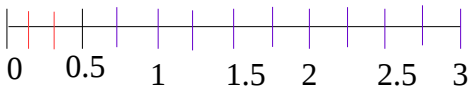
0	**01111111011**	**1110011001011110011111010010010011111110000011110000**
Sign	Exponent	Fraction

Hence, we see the value to be more accurate when we use more fractional bits. Because we get more bits to represent the number. Whereas, Exponent only increases the range.

2. What is Normal and Subnormal Values as per IEEE 754 standards explain this with the help of number line

Answer:

The normal values are the one which has no leading zeros in the fractional part. All the leading zeros are moved to exponent. For example – 0.001 is written as 1x10^-3. Whereas, in Subnormal values as per IEEE 754 standards, when we move the leading zeros to exponent it results in values exponent that is below the minimum exponent. Because, exponent has limited range it can represent.



All the lines marked in blue in number line can be represented as normal values and lines in red in number line is represented as subnormal numbers as per IEEE 754 standard.

3. IEEE 754vv defines standards for rounding floating points numbers to a represent able value. There are five methods defines by IEEE for this – Take time and understand what these five methods and explain it in your words using diagrams, illustrations of your own.

Answer:

- Round to nearest: Here, the value is rounded to closest value of the two possible values based on the accuracy needed. If the value is half way between two possible value then the least significant bit of fraction is made zero to make it even.

  Example:

  1.23 = 1.2 (rounded to one decimal point)

  1.27 = 1.3 (rounded to one decimal point)

  1.25 = 1.2 (rounded to one decimal point)

  1.35 = 1.4 (rounded to one decimal point)

- Round up: Here the value is changed to the larger of two possible values.

  Example:

  1.23 = 1.3(rounded to one decimal point)

  1.27 = 1.3 (rounded to one decimal point)

  1.25 = 1.3 (rounded to one decimal point)

- Round down: Here the valueis changed to lower of the two possible values.

  Example:

  1.23 = 1.2 (rounded to one decimal point)

  1.27 = 1.2 (rounded to one decimal point)

  1.25 = 1.2 (rounded to one decimal point)

- Round to zero: Here the value changed such that value is closer to zero.

  Example:

  1.23 = 1.2(rounded to one decimal point)

  - 1.23 = - 1.2 (rounded to one decimal point)

  1.25 = 1.2 (rounded to one decimal point)

References:

1. http://www.binaryconvert.com/result_double.html?decimal=048046049049056055052050052053053051051053051052

2. https://www.h-schmidt.net/FloatConverter/IEEE754.html

3. http://www.ias.ac.in/article/fulltext/reso/021/01/0011-0030

4. https://en.wikipedia.org/wiki/Denormal_number

5. http://www.keil.com/support/man/docs/armlib/armlib_chr1358938950865.htm