Advanced ARM Architecture – Assignment 3

Aparna Nair
IMT2015504

Question

In Floating point representation we have three components
1.The Sign Bit
2.Exponent
3.Fractional Part
Precession is one the prime attribute of any Floating point Representation,

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

[For 1 to 3 Do not Copy paste from any published sources, including diagrams]

Answer

1.

The precision of a number is defined as the number of digits present in any number. [1]
As such, the only part of the floating point representation that deals with the digits present in the original number is the mantissa, or the fractional part. [2]
For eg.) 0.55 is more precise than 0.5 since 0.55 has two digits and 0.5 has only one.

2.

Normal numbers: Normal numbers are basically numbers that are non-zero floating point numbers that follow the below mentioned format, i.e., the bit before the decimal point is always a 1. [3,4]
For IEEE-754-32 bit floating point format: [4]

$$(-1)^{s} \times 2^{(e-127)} \text{ x } 1.f$$

As such, these are able to make complete use of the full precision available for a given format.
Eg.) 0 10000100 10100001110000000000000 = 52.21875

Subnormal numbers: These are numbers that are smaller than normal number. They follow the below mentioned format, i.e., the bit before the decimal point is always a 0. [3,4]
For IEEE-754-32 bit floating point format: [4]

$$(-1)^{s} \times 2^{(e-127)} \text{ x } 0.f$$

Clearly in this case one more bit is being used thus reducing the precision. Further, all the exponent bits are zero.
Eg.) 0 00000000 00000000000000000000001 = 2^(-149)

The number line is shown below:

```
(-inf)------------|------------||-|----------|------------(+inf)
                 -1           0 y        1
```

where, $y=0.999999988 \times 2^{-126}$
Key:
Subnormal – YELLOW
Normal - RED

3. [5,6]
We round off a number in order to squeeze it such that it fits into the specified amount of space or conforms to a particular format.
We can round any given number using five different methods. Round to:
i) Nearest Even: The default method that all of us uses. It's rounded off to the closest floating point number that's of maximum precision in a given format. In the event that the number is equidistant from both the candidates, the number with an even least significant digit is returned.
ii) Nearest Away: It's rounded off to the closest floating point number that's of maximum precision in a given format. In the event that the number is equidistant from both the candidates, the number with a larger magnitude is returned.
iii) Up: This returns the number that is larger than the given number to the highest possible precision for that format. (+inf)
iv) Down: This returns the number that is smaller than the given number to the highest possible precision for that format. (-inf)
v) To Zero: It's similar to rounding a number down, except that now it is rounded off to the smallest number magnitude-wise to the highest possible precision for that format.

Let's see an example, number to be rounded off = 1.6,1.5,-1.5
According to the various methods;
i) Nearest Even: 1, 1, -2
ii) Nearest Away: 1, 2, -2

iii) Up: 2, 2,-1
iv) Down: 1, 1,-2
v) To Zero: 1, 1,-1

References:

[1] https://docs.microsoft.com/en-us/sql/t-sql/data-types/precision-scale-and-length-transact-sql?view=sql-server-2017
[2] https://steve.hollasch.net/cgindex/coding/ieeefloat.html
[3] http://mathworld.wolfram.com/Floating-PointNormalNumber.html
[4] https://www.ias.ac.in/article/fulltext/reso/021/01/0011-0030
[5] https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4610935&tag=1
[6] https://cs.nyu.edu/courses/spring16/CSCI-UA.0201-001/resources/lecture10.pdf