**Precession of the number:**

Both exponent, fractional or significand part play an important role in precession but mainly fractional part or significand part. As we know that the Logical representation of normal number is (-1)s \*(1.f)\*(2)e-12 and represents significand or fractional bits. As, number of significand or fractional bits increases the precision also increases that is why 64 bit floating point representation is also known as double precession [1].

Ex: Representation of same floating point hex number 0x15afe01e

Single precession: 7.103553E-26

Double precession: 1.79764653829008578644058390943E-315

**Normal and Subnormal Values:**

According to IEEE 754 standard to increase precision of the significand or fractional part they use a normalized significand which implies that it’s most significand bit is always 1. This means that in 32-bit floating point the significand is 24 bits long, out of which 23 bits are stored in memory with an implicit 1 as most significand 24th bit [1].

Logical representation of normal number is (-1)s \*(1.f)\*(2)e-127

Ex:

|  |  |  |
| --- | --- | --- |
| 0 | 00101011 | 01011111110000000011110 |
| Sign 1bit | exponent 8 bits | significand 23 bits |

Decimal representation is 7.103553E-26

When all exponent bits are 0 and leading hidden bit is 0 then floating point number is called a subnormal number.

Logical representation of subnormal number is (-1)s \*(0.f)\*(2)-127

Ex:

|  |  |  |
| --- | --- | --- |
| 0 | 00000000 | 01011111110000000011110 |
| Sign 1bit | exponent 8 bits | significand 23 bits |

Decimal representation is 4.396666E-39

**IEEE standard for rounding floating point numbers:**

**References:**

[1] <http://www.ias.ac.in/article/fulltext/reso/021/01/0011-0030>