Quan Nguyen

**The IEEE Standard for Floating-Point Arithmetic** (IEEE 754) is a technical standard for floating-point arithmetic established in 1985 by the Institute of Electrical and Electronics Engineers (IEEE). The standard addressed many problems found in the diverse floating-point implementations that made them difficult to use reliably and portably. Many hardware floating-point units use the IEEE 754 standard. [1]

**IEEE 754 has 3 basic components**:

The Sign of Mantissa – This is as simple as the name. 0 represents a positive number while 1 represents a negative number. [2]

The Biased exponent – The exponent field needs to represent both positive and negative exponents. A bias is added to the actual exponent in order to get the stored exponent. [2]

The Normalised Mantissa – The mantissa is part of a number in scientific notation or a floating-point number, consisting of its significant digits. Here we have only 2 digits, i.e. O and 1. So a normalised mantissa is one with only one 1 to the left of the decimal. [2]

**Two binary floating-point basic formats**: single precision and double precision [2]:

Single precision has:

Sign: 1 bit

Bias exponent: 8 bits

Normalised mantisa: 23 bits

bias: 127

Double precision has:

Sign: 1 bit

Bias exponent: 11 bits

Normalised mantisa: 52 bits

bias: 1023

**The standard definitions**:

- Arithmetic formats: sets of binary and decimal floating-point data, which consist of finite numbers (including signed zeros and subnormal numbers), infinities, and special "not a number" values (NaNs) [1]

Zero is a special value denoted with an exponent and mantissa of 0. -0 and +0 are distinct values, though they both are equal. [2]

If the exponent is all zeros, but the mantissa is not then the value is a denormalized number. [2]

The values +infinity and -infinity are denoted with an exponent of all ones and a mantissa of all zeros. The sign bit distinguishes between negative infinity and positive infinity. [2]

The value NAN is used to represent a value that is an error. This is represented when exponent field is all ones with a zero sign bit or a mantissa that it not 1 followed by zeros. [2]

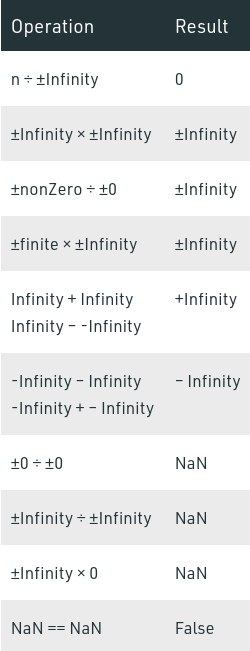
- Interchange formats: encodings (bit strings) that may be used to exchange floating-point data in an efficient and compact form [1]

- Rounding rules: properties to be satisfied when rounding numbers during arithmetic and conversions [1]

- Operations: arithmetic and other operations (such as trigonometric functions) on arithmetic formats [1]

- Exception handling: indications of exceptional conditions (such as division by zero, overflow, etc.) [1]

**Special operations**: [2]



[1]: https://en.wikipedia.org/wiki/IEEE\_754

[2]: https://www.geeksforgeeks.org/ieee-standard-754-floating-point-numbers/