

## 1 IEEE 754 Number Representation

As you can see in your textbook, the IEEE754 Floating Point representation is composed of three parts, the Mantissa Sign,  $S$ , the Signed Exponent,  $E$ , and the Mantissa Magnitude,  $M$ . In single precision floating point representation, the Signed Exponent,  $E$ , is 8 bits, whereas the Mantissa Magnitude,  $M$ , is composed of the remaining 23 bits. In double precision floating point representation, the Signed Exponent,  $E$  is 11 bits, whereas the Mantissa Magnitude,  $M$ , is composed of the remaining 52 bits. In both cases, the hidden-1 representation for the Mantissa Magnitude holds, effectively extending its representational power by one bit.

The value of a single precision IEEE754 Floating Point number is typically given by the following formula:

$$N = (-1)^S 2^{E-127} (1.M) \quad (1)$$

Yet, one of the things to keep in mind is that this interpretation only holds for  $0 < E < 255$ . For  $E = 0$  (i.e.,  $E$  being the bit string “00000000”) and for  $E = 255$  (i.e.,  $E$  being the bit string “11111111”) alternate value interpretations hold as given below.

Condition	N value
$E = 255$ and $M \neq 0$	NaN
$E = 255$ and $M = 0$	$(-1)^S \infty$
$E = 0$ and $M \neq 0$	$(-1)^S 2^{-126} (0.M)$
$E = 0$ and $M = 0$	$(-1)^S 0$

Similarly, the following interpretations hold for the case of *double precision* IEEE754 Floating Point numbers:

Condition	N value
$E = 2047$ and $M \neq 0$	NaN
$E = 2047$ and $M = 0$	$(-1)^S \infty$
$0 < E < 2047$	$(-1)^S 2^{E-1023} (1.M)$
$E = 0$ and $M \neq 0$	$(-1)^S 2^{-1022} (0.M)$
$E = 0$ and $M = 0$	$(-1)^S 0$