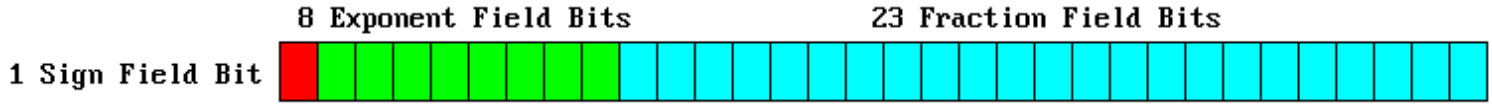
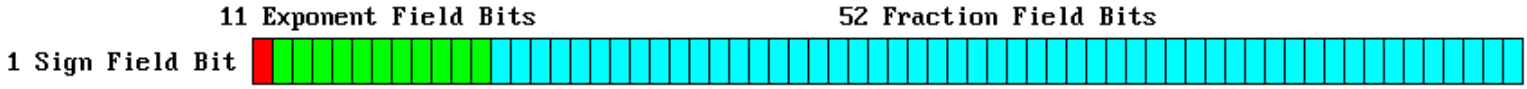
Floating Point Reference Sheet

32-Bit Float



64-Bit Double



Floating Point Representation

$V=(-1)^s \times M \times 2^E$ S = Sign M = Significand (Mantissa) E = Exponent	$Bias=2^{k-1}-1$ k = number of bits in exponent field 32-Bit Float: Bias = 127 64-Bit Double: Bias = 1023
Denormalized All exponent field bits are equal to "0" Significand has implied leading ZERO $E=1-Bias$ $M=0+\frac{Unsigned\ Value\ of\ Fraction\ Field\ Bits}{2^{Number\ of\ Fraction\ Field\ Bits}}$	**Normalized** Exponent field contains at least one "1" Significand has implied leading ONE $E=Unsigned\ Value\ of\ Exponent\ Field\ Bits-Bias$ $M=1+\frac{Unsigned\ Value\ of\ Fraction\ Field\ Bits}{2^{Number\ of\ Fraction\ Field\ Bits}}$

Special Cases

Description	*Bit Representation*
Zero	0 0000...0 0000...0
Smallest Denormalized	0 0000...0 0000...1
Largest Denormalized	0 0000...0 1111...1
Smallest Normalized	0 0000...1 0000...0
One	0 0111...1 0000...0
Largest Normalized	0 1111...0 1111...1
Infinity	0 1111...1 0000...0
Not-a-Number (NaN)	0 1111...1 <Anything non-zero>