3. Formats

This standard defines four floating-point formats in two groups, basic and extended, each having two widths, single and double. The standard levels of implementation are distinguished by the combinations of formats supported.

3.1 Sets of Values

This section concerns only the numerical values representable within a format, not the encodings. The only values representable in a chosen format are those specified by way of the following three integer parameters:

p = the number of significant bits (precision)
E_{max} = the maximum exponent
E_{min} = the minimum exponent

Each format's parameters are given in Table 1. Within each format only the following entities shall be provided:

Numbers of the form $(-1)^s 2^E (b_0 \cdot b_1 b_2 \ldots b_{p-1})$

where

s = 0 or 1
E = any integer between E_{min} and E_{max}, inclusive
b_i = 0 or 1

Two infinities, $+\infty$ and $-\infty$

At least one signaling NaN

At least one quiet NaN

The foregoing description enumerates some values redundantly, for example, $2^0(1 \cdot 0) = 2^1 (0 \cdot 1) = 2^2(0 \cdot 01) = \ldots$. However, the encodings of such nonzero values may be redundant only in extended formats (3.3). The nonzero values of the form $\pm 2^{E_{min}}(0 \cdot b_1 b_2 \ldots b_{p-1})$ are called denormalized. Reserved exponents may be used to encode NaNs, $\pm\infty$, ± 0, and denormalized numbers. For any variable that has the value zero, the sign bit s provides an extra bit of information. Although all formats have distinct representations for +0 and −0, the signs are significant in some circumstances, such as division by zero, and not in others. In this standard, 0 and ∞ are written without a sign when the sign is not important.

Table 1— Summary of Format Parameters

Parameter	Format			
	Single	Single Extended	Double	Double Extended
p	24	≥ 32	53	≥ 64
E_{max}	+127	≥ +1023	+1023	≥ +16383
E_{min}	−126	≤ −1022	−1022	≤ −16382
Exponent *bias*	+127	unspecified	+1023	unspecified
Exponent width in bits	8	≥ 11	11	≥ 15
Format width in bits	32	≥ 43	64	≥ 79

3.2 Basic Formats

Numbers in the single and double formats are composed of the following three fields:

1) **1-bit sign *s***
2) **Biased exponent *e* = *E*+*bias***
3) **Fraction *f* = · $b_1 b_2 \ldots b_{p-1}$**

The range of the unbiased exponent E shall include every integer between two values E_{min} and E_{max}, inclusive, and also two other reserved values $E_{min}-1$ to encode ±0 and denormalized numbers, and $E_{max}+1$ to encode ±∞ and NaNs. The foregoing parameters are given in Table 1. Each nonzero numerical value has just one encoding. The fields are interpreted as follows:

3.2.1 Single

A 32-bit single format number X is divided as shown in Fig 1. The value v of X is inferred from its constituent fields thus

1) If $e = 255$ and $f \neq 0$, then v is NaN regardless of s
2) If $e = 255$ and $f = 0$, then $v = (-1)^s\infty$
3) If $0 < e < 255$, then $v = (-1)^s 2^{e-127}(1 \cdot f)$
4) If $e = 0$ and $f \neq 0$, then $v = (-1)^s 2^{-126}(0 \cdot f)$ (denormalized numbers)
5) If $e = 0$ and f = 0, then $v = (-1)^s 0$ (zero)

3.2.2 Double

A 64-bit double format number X is divided as shown in Fig 2. The value v of X is inferred from its constituent fields thus

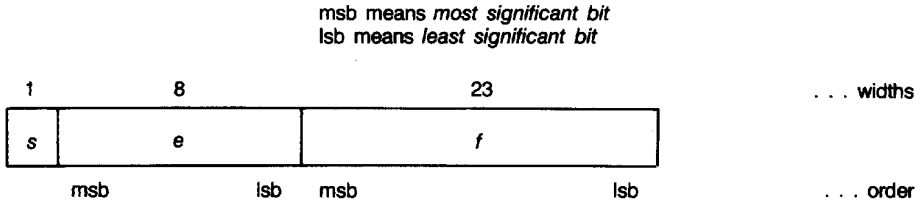

Figure 1— Single Format

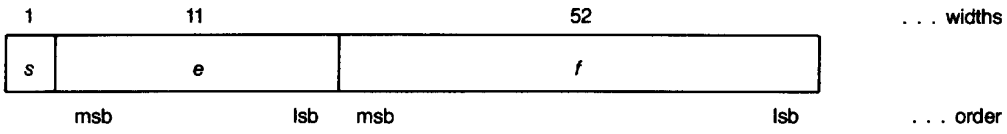

Figure 2— Double Format

1) If $e = 2047$ and $f \neq 0$, then v is NaN regardless of s
2) If $e = 2047$ and $f = 0$, then $v = (-1)^s\infty$
3) If $0 < e < 2047$, then $v = (-1)^s 2^{e-1023}(1 \cdot f)$
4) If $e = 0$ and $f \neq 0$, then $v = (-1)^s 2^{-1022}(0 \cdot f)$ (denormalized numbers)
5) If $e = 0$ and $f = 0$, then $v = (-1)^s 0$ (zero)