FLOATING POINT ARITHMETIC

A (nonzero, normalized) base β, t-digit floating point number has the mathematical representation

$$\pm .d_1 d_2 \cdots d_t \times \beta^e,$$

where

$$m \le e \le M, \qquad 1 \le d_1 \le \beta - 1, \qquad 0 \le d_i \le \beta - 1 \text{ for } i = 2, \ldots, t.$$

A real number x is said to be within floating range if $|x| \le \beta^M$. Let $fl(x)$ denote the floating point representation of the real number x within floating point range. Then

$$fl(x) = x(1+\delta), \qquad |\delta| \le u = \begin{cases} \beta^{1-t} & \text{chopped,} \\ \frac{1}{2}\beta^{1-t} & \text{rounded,} \end{cases}$$

depending on whether $fl(x)$ is obtained by chopping or rounding the base β expansion of x. u is called the *unit round-off*.

Assumption. Let $\bowtie$ denote any of the floating point arithmetical operations $+$, $-$, *, or /. Then for any floating point numbers x and y,

$$fl(x \bowtie y) = (x \bowtie y)(1+\delta), \qquad |\delta| \le u.$$

Example.

$$fl\left(\sum_{i=1}^{3} x_i y_i\right) = \left\{\left[x_1 y_1 (1+\delta_1) + x_2 y_2 (1+\delta_2)\right](1+\delta_3) + x_3 y_3 (1+\delta_4)\right\}(1+\delta_5)$$

where all $|\delta_i| \le u$.

Note. An equivalent definition of the unit round-off u is that u is the smallest floating point number such that $fl(1+u) > 1$.

Internal machine representation of floating point numbers—**IEEE 754 Standard.**

32-bit format

31 0

sign bit	biased exponent	significand
1	8	↑1. 23

Exponent bias $= 7F_{16}$.
$\beta = 2$, $t = 24$, $-126 \le e \le 127$.

64-bit format

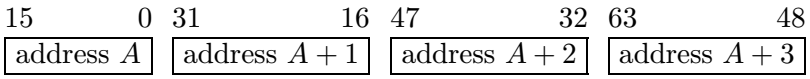
63 0

sign bit	biased exponent	significand
1	11	↑1. 52

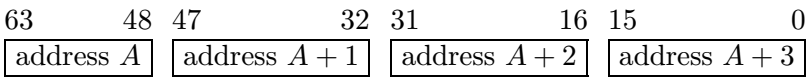
Exponent bias $= 3FF_{16}$.
$\beta = 2$, $t = 53$, $-1022 \le e \le 1023$.

Except for zero and denormals, the significand is assumed to follow 1. (in binary). The stored exponent $E = \text{bias} + e$.

Memory storage for Intel 80*, MIPS R*, DEC Alpha chips:

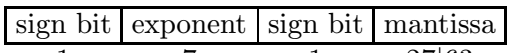


Memory storage for Motorola 68*, IBM RS6000, SUN Sparc, Power 60* chips:



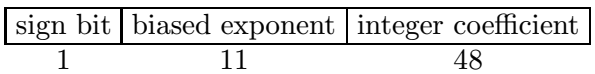
Some other floating point representations.

Honeywell 68/60



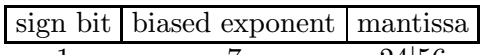
Exponent: fixed point binary integer.
Mantissa: fixed point binary fraction.
The sign bits are part of the exponent and mantissa.
$\beta = 2$, $t = 27 \mid 63$, $-128 \leq e \leq 127$.

CDC 6000 series



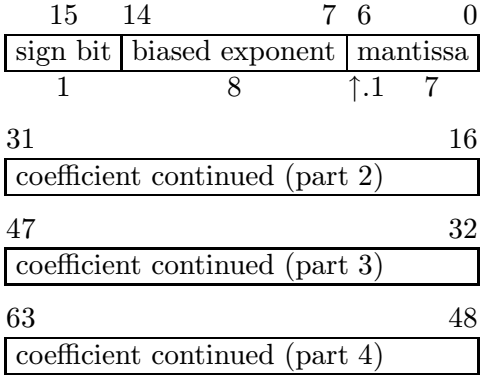
Exponent bias $= 2^{10}$.
Note that the mantissa is an integer, not a fraction.
$\beta = 2$, $t = 48$, $-1024 \leq e \leq 1023$.

IBM SYSTEM/360 series



Exponent bias $= 2^6$.
$\beta = 16$, $t = 6 \mid 14$, $-64 \leq e \leq 63$.

DEC VAX series



Exponent bias $= 2^7$.
The most significant mantissa bit is not stored.
$\beta = 2$, $t = 24 \mid 56$, $-128 \leq e \leq 127$.

IEEE Standard 754 Real and Long Real Encodings

Class			Sign	Biased Exponent	Significand* $\overset{\text{ff} \ldots \text{ff}}{\Delta}$
Positives	NaNs	Quiet	0 ⋮ 0	11...11 ⋮ 11...11	11...11 ⋮ 10...00
		Signaling	0 ⋮ 0	11...11 ⋮ 11...11	01...11 ⋮ 00...01
	∞		0	11...11	00...00
	Reals	Normals	0 ⋮ 0	11...10 ⋮ 00...01	11...11 ⋮ 00...00
		Denormals	0 ⋮ 0	00...00 ⋮ 00...00	11...11 ⋮ 00...01
———		Zero	0	00...00	00...00
		Zero	1	00...00	00...00
		Denormals	1 ⋮ 1	00...00 ⋮ 00...00	00...01 ⋮ 11...11
Negatives		Normals	1 ⋮ 1	00...01 ⋮ 11...10	00...00 ⋮ 11...11
	∞		1	11...11	00...00
	NaNs	Signaling	1 ⋮	11...11 ⋮ 11...11	00...01 ⋮ 01...11
		Quiet	1 ⋮ 1	11...11 ⋮ 11...11	10...00 ⋮ 11...11
			Short: Long:	⟵ 8bits ⟶ ⟵ 11bits ⟶	⟵ 23bits ⟶ ⟵ 52bits ⟶

∗ Integer bit is implied and not stored.