1.2 IEEE standard 754 : FLOATING POINT

↳ Institute of Electrical & Electronics Engineers

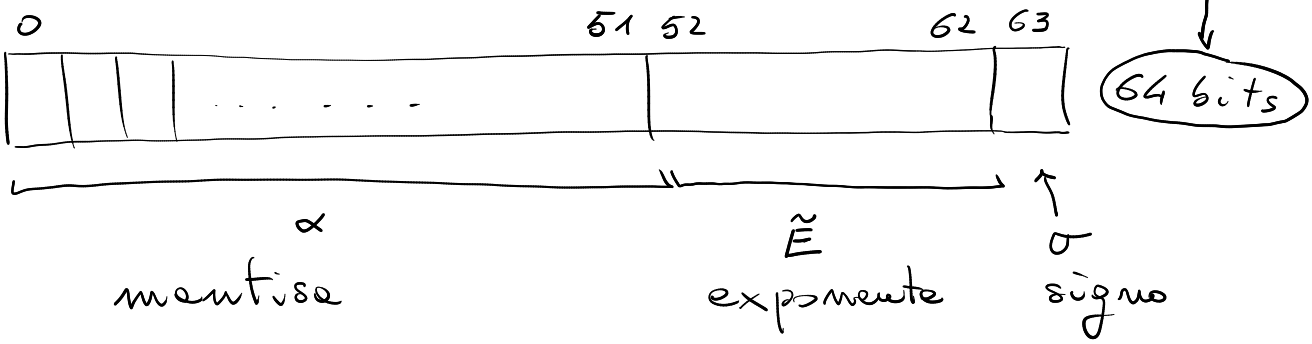
los numeros reales se representan en la máquina como

$$X = (-1)^{\sigma} \; (1.\alpha_1\alpha_2 \ldots \alpha_m)_2 \; 2^E$$

[obs: $(1.011)_2 \cdot 2 = 10.11$
$(11.001)_2 \cdot 2^{-1} = 1.1001$]

$\sigma \in \{0,1\}$, $\alpha_j \in \{0,1\}$, $j = 1 \ldots m$, $m = 52$ doble precisión

$E \in \{-1022, \ldots, 1023\}$



$\tilde{E}$: 11 bits => se pueden representar todos los enteros entre 0 y $2^{11}-1 = 2047$

$\tilde{E} = 0$ reservado para representar 0
(y para tratar los problemas de underflow)

$\tilde{E} = 2047$ reservado para representar ∞

↳ quedan 2046 valores disponibles para definir E

$E = \tilde{E} - 1023$

. realmin = número más pequeño que se puede representar (en valor absoluto)

$(1.000\ldots0)_2 \cdot 2^{-1022} = 2^{-1022} \simeq 2.225 \cdot 10^{-308}$

. realmax = número más grande que se puede representar (en valor absoluto)

$(1.11\ldots1)_2 \cdot 2^{1023} = (2 - 2^{-52}) \cdot 2^{1023} \simeq 1.8 \cdot 10^{308}$

. flintmax = número entero más grande hasta el que se pueden representar todos los enteros

es 2^{53}. entre 2^{53} y 2^{54} tenéis 1 entero cada 2

entre 2^{54} y 2^{55} " " 4

. eps : ε - máquina / machine precision

precisión con la que se pueden representar los números cerca de 1

. el número IEEE 754 64 bits $x > 1$ más pequeño es

$x = (1.00\ldots01)_2 \cdot 2^0 = 1 + 2^{-52}$

(2^{-52})

observar que $1 + 2^{-53} = (1.000\ldots0\,1) \cdot 2^0 = 1$

(52 ceros; se necesitaría un bit 53)

. el número IEEE 754 64 bits $x < 1$ más grande es

$x = (1.11\ldots1)_2 \cdot 2^{-1} = 1 - 2^{-53}$

observación: la aritmética en floating point no es asociativa

$$1 - (1 + 2^{-53}) = 0$$

$$(1 - 1) - 2^{-53} = -2^{-53}$$

$$1 + 2^{-53} - 1 = 0$$

$$-1 + 1 + 2^{-53} = 2^{-53}$$

} ORDEN: de izquierda a derecha