M.L. OVERTON - "Numerical Computing with IEEE Floating Point Arithmetic" - SIAM (2001

1 := 3ff00 — 0
2 := 40000 — 0
4 := 40100 — 0
⋮

-1 := BFF00 — 0
-2 := C0000 — 0

Doppia precisione: 64 bit (1+11+52)

- Table 4.2: IEEE Double Format

±	$a_1a_2a_3\ldots a_{11}$	$b_1b_2b_3\ldots b_{52}$

0 / 1

If exponent bitstring is $a_1\ldots a_{11}$	Then numerical value represented is
$(00000000000)_2 = (0)_{10}$	$\pm(0.b_1b_2b_3\ldots b_{52})_2 \times 2^{-1022}$
$(00000000001)_2 = (1)_{10}$	$\pm(1.b_1b_2b_3\ldots b_{52})_2 \times 2^{-1022}$
$(00000000010)_2 = (2)_{10}$	$\pm(1.b_1b_2b_3\ldots b_{52})_2 \times 2^{-1021}$
$(00000000011)_2 = (3)_{10}$	$\pm(1.b_1b_2b_3\ldots b_{52})_2 \times 2^{-1020}$
$\downarrow$	$\downarrow$
$(01111111111)_2 = (1023)_{10}$	$\pm(1.b_1b_2b_3\ldots b_{52})_2 \times 2^{0}$
$(10000000000)_2 = (1024)_{10}$	$\pm(1.b_1b_2b_3\ldots b_{52})_2 \times 2^{1}$
$\downarrow$	$\downarrow$
$(11111111100)_2 = (2044)_{10}$	$\pm(1.b_1b_2b_3\ldots b_{52})_2 \times 2^{1021}$
$(11111111101)_2 = (2045)_{10}$	$\pm(1.b_1b_2b_3\ldots b_{52})_2 \times 2^{1022}$
$(11111111110)_2 = (2046)_{10}$	$\pm(1.b_1b_2b_3\ldots b_{52})_2 \times 2^{1023}$
$(11111111111)_2 = (2047)_{10}$	$\pm\infty$ if $b_1 = \cdots = b_{52} = 0$, NaN otherwise

All ≡ ∅ Sub normal →

NORMAL (−1023)

At least one 1

All ≡ 1 Exception →

:= realmin se $b_i = 0$

:= 1 se $b_i = 0$

:= 2 se $b_i = 0$

realmin := 2^{-1022} (NORMALE)

2^{-1074} := più piccolo subnormale