Exponent bit pattern	Unsigned value	Signed value	
		Bias = 127	Bias = 128
11111111	255	+128	+127
11111110	254	+127	+126
⋮	⋮	⋮	⋮
10000001	129	+2	+1
10000000	128	+1	0
01111111	127	0	-1
01111110	126	-1	-2
⋮	⋮	⋮	⋮
00000001	1	-126	-127
00000000	0	-127	-128

* Observation: for SM mantissas

$\rightarrow \underbrace{1.x_M}_{\text{significand}}$

hiding → gain precision

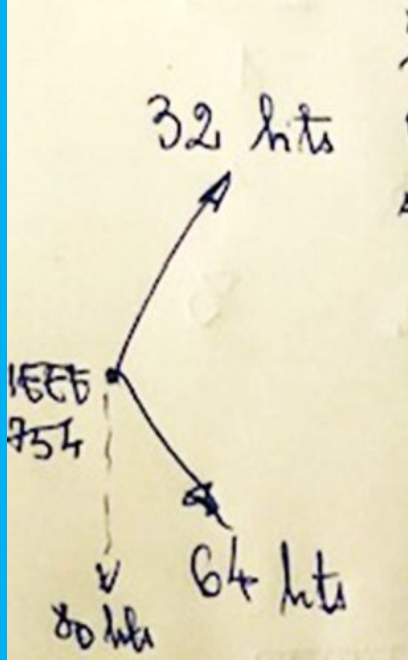
$$1 \leq \underbrace{1.x_M}_{\text{significand}} < 2$$

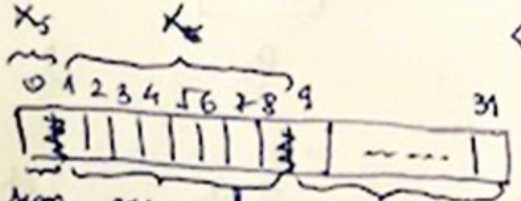
Standards IEEE (Institute of Electrical and Electronics Engineers)

→ IEEE 754 Floating Point Standard (Zuse, 1985 Kahan, 1987 → 854, 2008 Cowlishaw, Revision 2019 - Hough & Cowlishaw)

→ mainframes → IBM Floating Point „Standard"

IEEE 754 Floating Point Standard



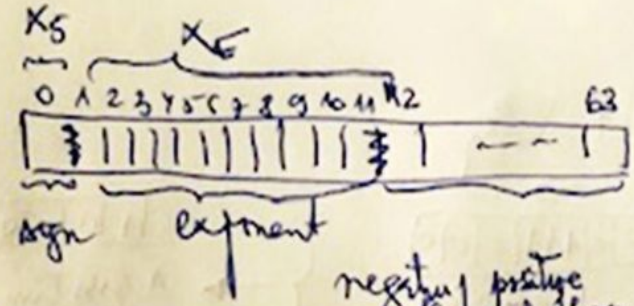


$$X = (-1)^{x_S} \cdot 2^{x_E - 127} \cdot (1.x_M)$$

hidden bit

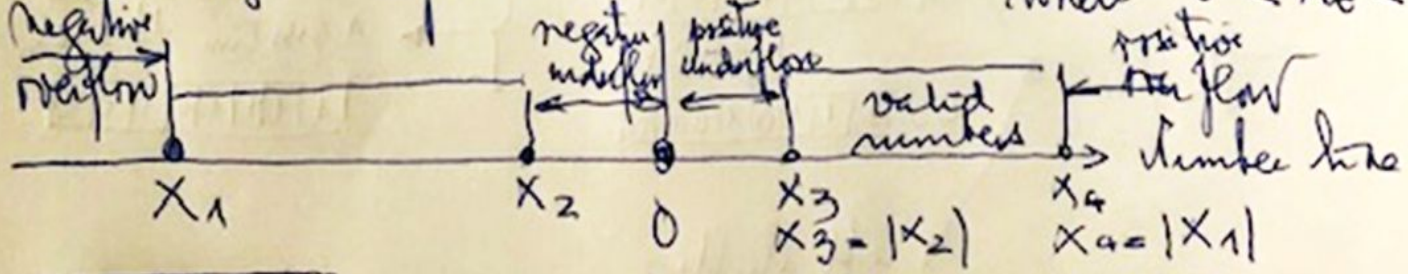
where $0 < x_E < 255$

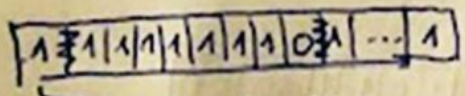
significand



$$X = (-1)^{x_S} \cdot 2^{x_E - 1023} \cdot (1.x_M)$$

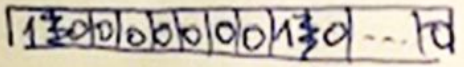
where $0 < x_E < 2047$





$$x_1 = (-1)^1 \cdot 2^{254-127} \cdot (1 + \tfrac{1}{2} + \cdots \tfrac{1}{2^{23}}) = -2^{127}(2 - 2^{-23})$$

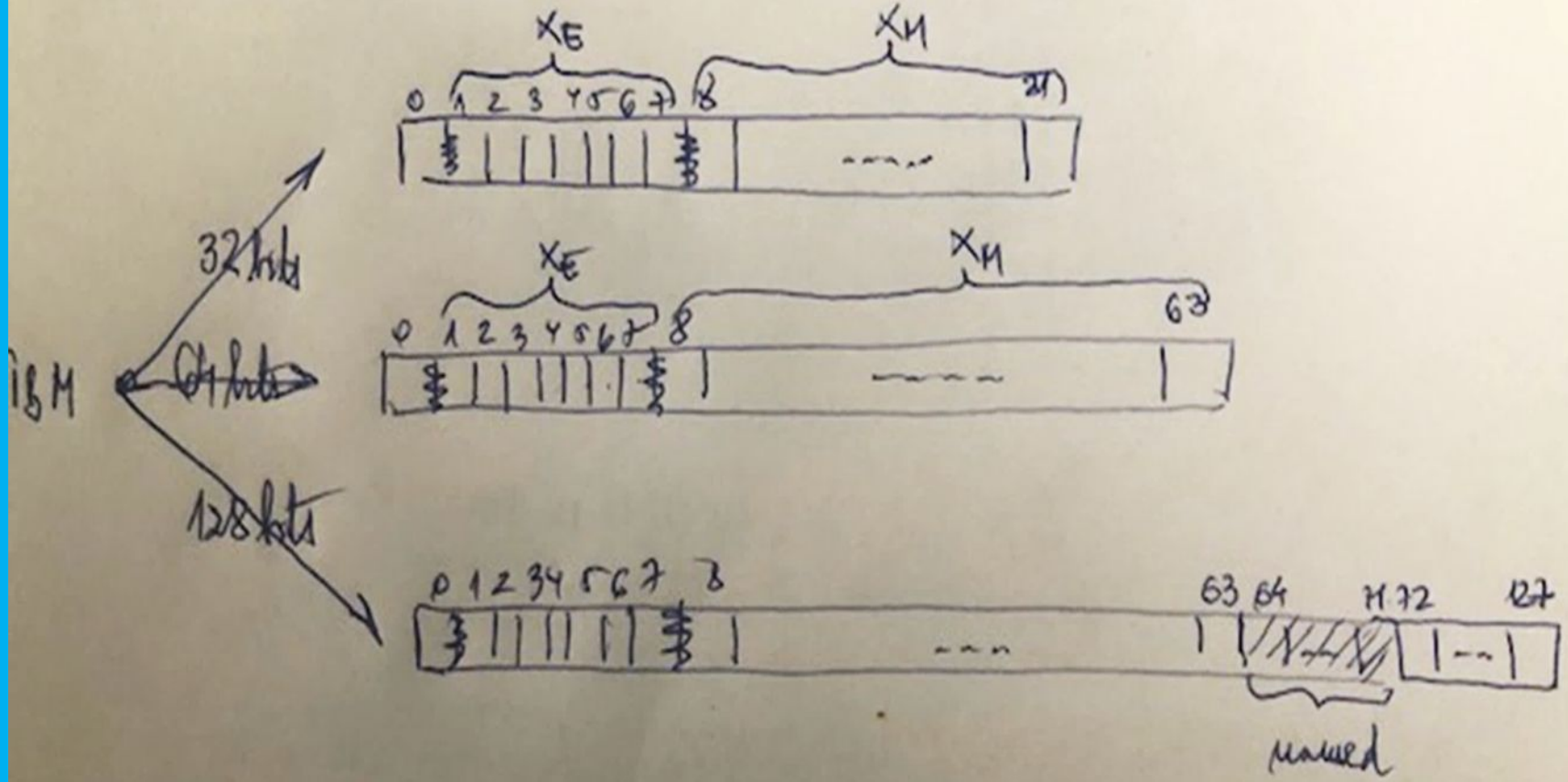
hidden bit



$$x_2 = (-1)\, 2^{1-127} \cdot (1 + 0 \cdots + 0) = -2^{-126}$$

IBM Floating Point „Standard"

$$X = (-1)^{X_S} \cdot 16^{X_E - 64} \cdot (0.X_M)$$



$$X = -794,0898437\,5_{10} = -1100011010,00010111_2$$

$$X = (-1)^{x_s} \cdot 2^{x_E - 127} \cdot (1.x_M)$$

794 | 2
0 | 397 | 2
1 | 198 | 2
0 | 99 | 2
1 | 49 | 2
1 | 24 | 2
0 | 12 | 2
0 | 6 | 2
0 | 3 | 2
1 | 1

0,08984375 × 2
0,17968750 × 2
0,35937500 × 2
0,71875000 × 2
1,43750000 × 2
0,87500000 × 2
1,75000000 × 2
1,50000000 × 2
1,00000000

$$X = -1,1000\,1101\,0000\,1011\,1 \times 2^9$$

$$X_E - 127 = 9 \Rightarrow X_E = 136$$

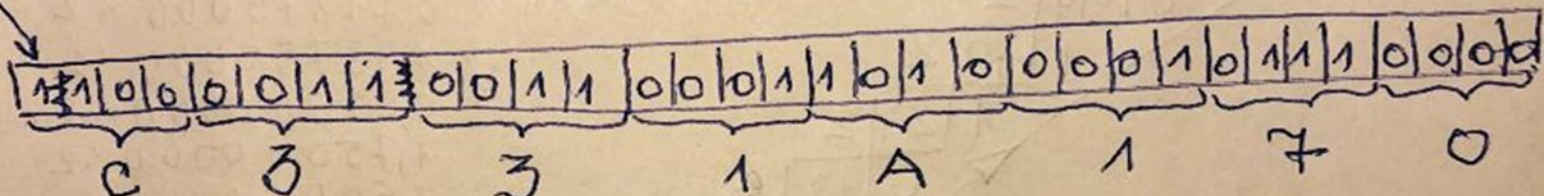
1	1	0	1	0	0	0	1	1	0	0	0	1	1	0	1	0	0	0	0	1	0	1	1	1	0	0	0	0
C			4			4			6			8			5			C			0								

$$X_{16} = C4468\,5C0_{16}$$

$X = -794{,}08984375_{10} = -1100011010{,}00010111_2 =$

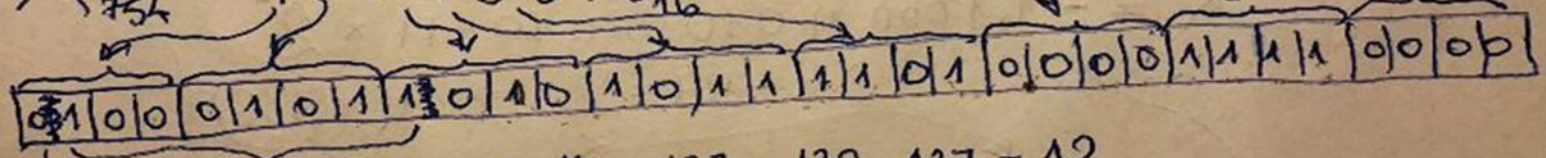
$= -0{,}0011000110100001011 \times 16^3$

$X = (-1)^{x_S} \cdot 16^{x_E - 64} \cdot (0 x_M)$ $\quad x_E - 64 = 3 \Rightarrow x_E = 67$



$X_{IBM} = C331A170_{16}$

$X_{754} = 45ABD070_{16}$



$139 = x_E \Rightarrow x_E - 127 = 139 - 127 = 12$

$X_{754} = (-1) \cdot 2^{12} \cdot 1{,}0101011110100001111 =$

hidden bit

$= X_{754} = 1010101111010{,}0001111_2 =$

$= + (4096 + 1024 + 256 + 64 + 32 + 16 + 8 + 2 + \frac{15}{128}) =$

$= + 5498{,}117187_{10}$