ARITMETICA di MACCHINA

$x = 123456.789$ notazione posizionale

10^{-2}, 10^{0}, 10^{-1}, 10^{1}

forma esponenziale

esponente della base

$$x = 0.\underbrace{123456789}_{\text{mantissa} \in \mathbb{Z}} \cdot \underbrace{10}_{\text{base di rappr.}}^{6}$$

$x = 1.23456789 \cdot 10^{5}$ 2 forma esponenziale

$$\| \; x = (-1)^s \, 0.a_1 a_2 \cdots\cdots a_t \cdot \beta^e \; \|$$ forma esponenziale

con $s \in \{0,1\}$ $a_i \in \{0, 1, \dots \beta-1\}$

β = base di rappresentazione

t = n° cifre mantissa

$e \in \mathbb{Z}$ esponente della base

Sistema FLOATING POINT

$$\mathbb{F}(\beta, t, L, U) = \{ x = (-1)^s \, 0.a_1 a_2 \cdots a_t \cdot \beta^e \text{ con}$$
$$s \in \{0,1\},\ a_i \in \{0, 1, \dots, \beta-1\} \text{ per } i = 1, \dots, t,\ \underline{a_1 \neq 0}$$
$$L < 0,\ U > 0,\ L \le e \le U \} \cup \{0\}$$

• sistema $\mathbb{F}$ semplice precisione

$\beta = 2$ registro di 4 Byte

1	8	23
s	e	mantissa

$= 32$ bit

float in C, C++

se $a_1 \neq 0 \Rightarrow a_1 = 1$

23 bit per m, ma $t = 24$

L, U

sistema doppia precisione

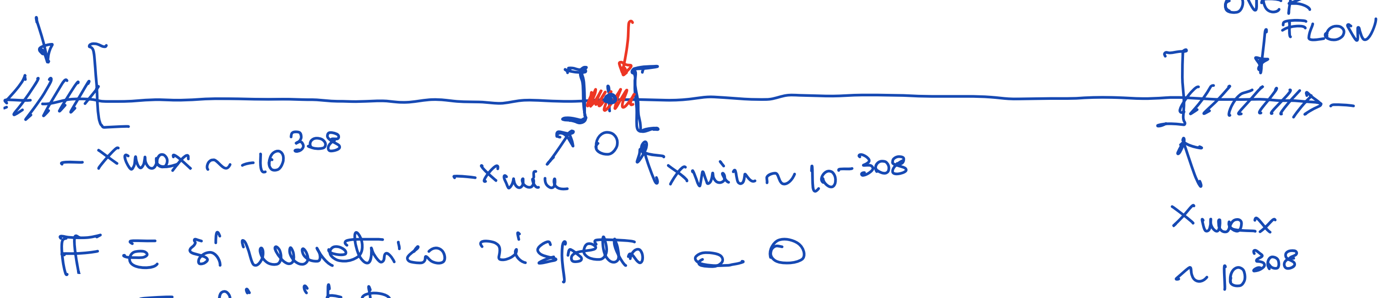
$\beta = 2$ 8 Byte = 64 bit double in C, C++

1	11	52
s	e	m

52 bit per $m \Rightarrow t = 53$

$L = -1021$ $U = 1024$ $1021 + 1 + 1024 = 2046$

$2^{11} = 2048$



$\mathbb{F}$ è simmetrico rispetto a 0
è limitato
è finito (cardinalità finita)

$$card(\mathbb{F}) = 2\beta^{t-1}(\beta-1)(1-L+U)$$

i floating point non sono equamente distribuiti

Esempio

$\mathbb{F}(\beta=2,\ t=3,\ L=-2,\ U=3)$ $\qquad x=(-1)^s\, 0.a_1a_2a_3 \cdot \beta^e$

$-2 \le e \le 3$

$x_{min} = \underbrace{0.100}_{\text{base } 2} \cdot 2^{-2} = \frac{1}{2}\cdot\frac{1}{4} = \frac{1}{8} = \frac{4}{32}$

leggo in base 10

$0.101 \cdot 2^{-2} = \left(\frac{1}{2}+\frac{1}{8}\right)\cdot\frac{1}{4} = \frac{5}{8}\cdot\frac{1}{4} = \frac{5}{32}$

$0.110 \cdot 2^{-2} = \left(\frac{1}{2}+\frac{1}{4}\right)\cdot\frac{1}{4} = \frac{3}{4}\cdot\frac{1}{4} = \frac{3}{16} = \frac{6}{32}$

$0.111 \cdot 2^{-2} = \left(\frac{1}{2}+\frac{1}{4}+\frac{1}{8}\right)\cdot\frac{1}{4} = \frac{7}{8}\cdot\frac{1}{4} = \frac{7}{32}$

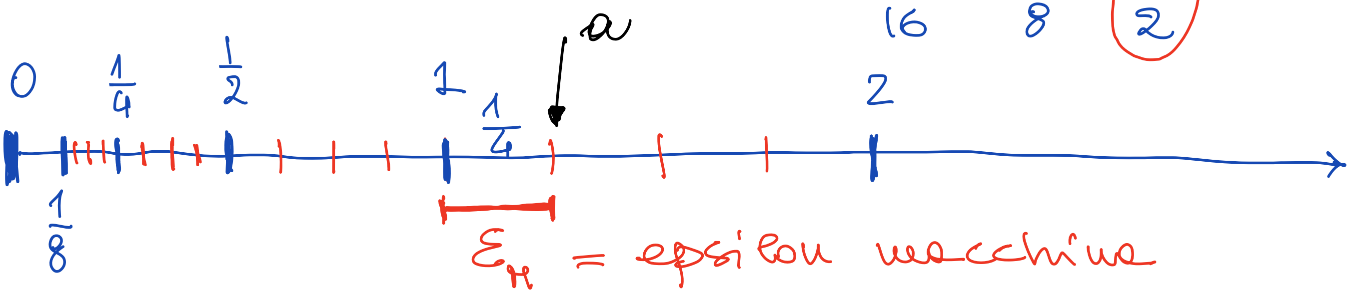
$0.100 \cdot 2^{-1} = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4} = \frac{8}{32} = \frac{4}{16}$

$0.101 \cdot 2^{-1} = \frac{5}{8}\cdot\frac{1}{2} = \frac{5}{16}$

$\frac{6}{16}$

$\frac{7}{16}$

$0.100 \cdot 2^{0}$ $\qquad \frac{8}{16} = \frac{4}{8} = \frac{1}{2}$



ε_M = epsilon macchina o precisione di macchina

è la distanza tra 1 e il + numero di $\mathbb{F}$ maggiore stretto di 1.

26/09/2023

$$1 = 0.100 \cdot \beta^1$$

$$a = 0.101 \cdot \beta^1$$

$$\varepsilon_M = a - 1 = 0.001 \cdot \beta^1 = \beta^{-t} \cdot \beta^1 = \beta^{1-t}$$

posiz. t

$$\boxed{\varepsilon_M = \beta^{1-t}}$$

nell'esempio avevo $\beta = 2$ e $t = 3$ $\Rightarrow$ $\varepsilon_M = 2^{-2} = \frac{1}{4}$

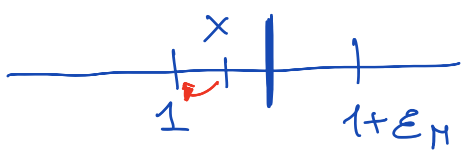
in semplice precisione $\varepsilon_M = 2^{1-24} \sim 10^{-7}$

in doppia precisione $\varepsilon_M = 2^{1-53} = 2^{-52} \sim 2.22 \cdot 10^{-16}$

eps

$x \in \mathbb{R}$, $\boxed{1 < x < 1 + \varepsilon_M}$ $\Rightarrow$ $x \notin \mathbb{F}$

$fl_t(x) = 1$



$fl_t(x)$ è l'arrotondam. di x in $\mathbb{F}$

$$|x - fl_t(x)| \leq \frac{1}{2} \varepsilon_M \qquad u = \frac{1}{2} \varepsilon_M$$

unità di arrotondamento (roundoff unit)

Più in generale :

$$|x - fl_t(x)| \le \frac{1}{2}\, \varepsilon_M \cdot |x|$$

se $x \neq 0$

$$\boxed{\frac{|x - fl_t(x)|}{|x|} \le \frac{1}{2}\, \varepsilon_M}$$

errore di arrotondamento relativo

Proprietà dell'aritmetica classica che cadono in $\mathbb{F}$

1) $\nexists\, !$ lo zero come elemento neutro della somma

cioè $\exists\, x \in \mathbb{F}$ con $x \neq 0$: $1 + x = 1$

es: $x = \beta^{-2t} \underset{(t=53)}{=} 2^{-106} \in \mathbb{F}$

$$1 = 0.1 \,\text{- - - - - - - - - -}\, \overset{\boxed{t}}{0} \cdot \beta^{1}$$

$$x = 0.1 \,\text{- - - - - - - - -}\, 0 \cdot \beta^{-2t+1}$$

allineando gli esponenti ho:

$$1 = 0.1 \,\text{- - - - - - - - - -}\, 0 \;|\; \cdot \beta^{1}$$

$$x = 0.0 \,\text{- - - - - - - - - -}\, \underset{\uparrow}{0} \;|\; \boxed{\text{- - - - - - - -}\; \underset{\uparrow}{1}} \cdot \beta^{1}$$

t-sima cifra

2t-sima cifra

si perde nel fare la somma

$$1 + x = 0.1 \,\text{- - - - - - - - - - -}\, 0 \cdot \beta^{1}$$

2) Le F cade la prop associativa di + e −

prop ass : $(a_1 + a_2) + a_3 = a_1 + (a_2 + a_3)$

$(a_1 \cdot a_2) \cdot a_3 = a_1 \cdot (a_2 \cdot a_3)$

es

$a_1 = 10^{308}$

$a_2 = 10^{308}$

$a_3 = -4 \cdot 10^{307}$

F doppia precisione

$x_{max} \sim 1.7 \cdot 10^{308}$

$$(a_1 + a_2) + a_3 = \underbrace{2 \cdot 10^{308}}_{\substack{\downarrow \\ > x_{max} \\ NaN}} - 4 \cdot 10^{307} = NaN$$

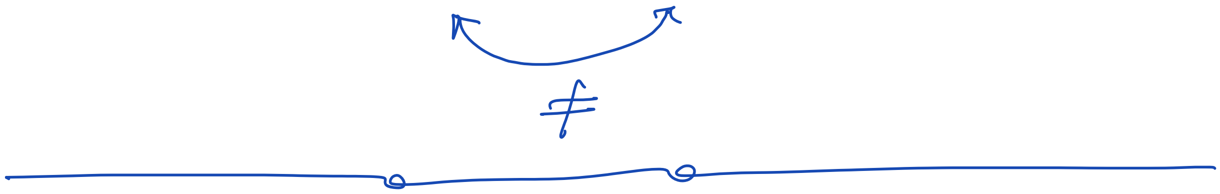
$$a_1 + (a_2 + a_3) = \underbrace{10^{308} + \underbrace{(10^{308} - 4 \cdot 10^{307})}}_{< x_{max}}$$

ES 2

$x \neq 0$

$$\frac{1 + (-1 + x)}{x} = \frac{(1-1) + x}{x} = 1$$

ne canta



Propagazione degli errori di arrotondamento

(+)

$x, y \in \mathbb{R}$ $\qquad \bar{x} = fl_t(x)$ $\qquad \bar{y} = fl_t(y)$

in conto ho $x+y$

in $\mathbb{F}$ ho $fl_t(\bar{x}+\bar{y})$

si riesce a dire

$$\frac{|fl_t(\bar{x}+\bar{y}) - (x+y)|}{|x+y|} \le \underbrace{\left(\frac{|x|+|y|}{|x+y|}+1\right)}_{\sim \frac{2}{10^{-15}}+1 \;\sim 2\cdot 10^{15}} u$$

se $x = 1+10^{-15}$ $\qquad y = -1$

$$\le 2\cdot 10^{15}\cdot 10^{-16} \sim 2\cdot 10^{-1}$$

La somma è un'operazione potenzialmente instabile

- $$\frac{\left| fl_t(\bar{x} \cdot \bar{y}) - (x \cdot y) \right|}{|x \cdot y|} \leq 3u$$

Il prodotto è un'op. stabile,

a piccoli errori sui dati corrispondono piccoli errori nella soluzione