2004/2005 (Fiumi) (1)

- Note sulla rappresentazione di un singolo reale:
  (da Fiumi, professione Diploma, pg 50-52)

Un numero in singola reale viene rappresentato nel formato

$x = \pm m \cdot 2^{\pm E}$ dove m è la mantissa
E è l'esponente.

- Rappresentazione della mantissa:
  - si rappresenta in modulo e segno
  - si rappresenta normalizzata, ossia, per la base 2, nel formato $1,\ldots$, in modo da evitare di rappresentare l'1.
  - in altre, se ho 1,010, è $1 \cdot 2^{0} + 0 \cdot 2^{-1} + 1 \cdot 2^{-2} + 0 \cdot 2^{-3}$ etc
    (Da Teb. 2,12).

- Rappresentazione dell'esponente.
  - per evitare il segno, non si rappresenta E ma E + Bias, dove Bias è un valore di polarizzazione.
  - Per cui si rappresenta come un Naturale.
  - ~~Esempio~~ Definiamo E + Bias = Eeff con Eeff su n bit
    Il valore di
    - Eeff = 0 indica: → 0 → Numeri denormalizzati.
    - Eeff = $2^n - 1$ indica → Positivo - infinito
      → Negativo - infinito
      → N. a. N.

- Supponiamo di avere 11 bit.
  I valori ~~da~~ rappresentabili sono da [0, 2047]
  - di questi 2 sono eseguiti → 0 → per lo 0 e le denormalizzate
    → 2047 → per infinito e N. a. N.
  - Restano 2046 Valori.

i 2 valori si scelgono di toglierli all'esponente negativo, dunque con 2046 valori scelgo di rappresentare l'intervallo [-1022, 1023] che contiene appunto 2046 valori.

⇒ allora, dato un esponente $E \in [-1022, 1023]$ che numero devo sommare ad E per ottenere un numero nell'intervallo [1, 2046], in quanto ∅ e 2047 sono occupati?

Risposta: 1023. questo è il BIAS per i double.

- RAPPRESENTAZIONE DI ∅ e Numeri Denormalizzati:
  - se $E_{eff} = \phi$, $m = \phi$ è $\pm\phi$ $(-1)^{1}*\phi$
  - se $E_{eff} = \phi$ e $m \neq 0$ è un numero Denormalizzato (vedi testo per il significato)
  - se $E_{eff} = 2^{n}-1$ e $m = \phi$ è Positivo / Negativo > Infinito

- se $E = 2^n - 1$ e $m \neq 0$ allora è N.a.N.

- se $0 < E_{eff} < 2^n - 1$

  allora il num è

  $(-1)^1 * (1,m) * 2^{E_{eff} - Bias}$

  dove $Bias = 2^{n-1} - 1$.

- float STANDARD IEEE 754

  Bit della mantissa: 23
  Bit dell'esponente: 8

  - $Bias = 2^{8-1} - 1 = 127$
  - esponente minimo rappresentabile: -126
  - esponente max rappresentabile: 127

Double standard IEEE 754

Bit della mantissa: 52

Bit dell'esponente: 11

$Bias = 2^{11-1} - 1 = 1023$

- esponente MAX = Bias = 1023
- esponente MIN = -Bias + 1 = -1022

Da qui otteniamo i valori per i double.

RICAPITOLAZIONE DELLO STANDARD IEEE 754 (schematic)

- FLOAT (NUMERI A PRECISIONE SINGOLA)
  - rappresentazione su 32 bit
    - 1 segno
    - 8 esponente
    - 23 mantissa
  - intervallo di rappresentazione
    $[-10^{38}, -10^{-38}] \cup [10^{-38}, 10^{38}]$
  - precisione a 7 cifre decimali

- DOUBLE (numeri a PRECISIONE DOPPIA)
  - rappresentazione su 64 bit
    - 1 segno
    - 11 esponente
    - 52 mantissa
  - INTERVALLO DI RAPPRESENTAZIONE
    $[-10^{308}, -10^{-308}] \cup [10^{-308}, 10^{308}]$
  - precisione a 17 cifre decimali significative.

- significo di NaN, positive-infinite, negative infinite.

- Se Eff = $2^m - 1$ ⇒ • m ≠ 0 invalid NaN
  • m = 0; 1 = ∅ invalid positive-infinite
  • m = ∅; 1 = 1 invalid negative-infinite

- Se Eff = ∅ e m != ∅ ⇒ numero denormalizzato
- Se Eff = ∅ e m = ∅ ⇒ S = ∅ ⇒ invalid ∅ ?
  ⇒ S = 1 ⇒ invalid -∅

(Da INTEL ADT)

- Per il calcolo dell'esponente effettivo, esclusi i valori ∅ e $2^m - 1$, vale la relazione

$$Eff = e + \left(2^{m-1} - 1\right)$$

(Eff ← esponente di ±1.m2^e)

Calcolo della mantissa:

- Caso $m = 1, a_1 a_2 a_3 a_4 a_5 \ldots$

  espressione: $m = 1 \cdot 2^0 + a_1 \cdot 2^{-1} + a_2 \cdot 2^{-2} + \ldots$

  $= 1 \cdot 2^0 + \frac{a_1}{2} + \frac{a_2}{4} \ldots$

  da cui $m - 1 = \frac{a_1}{2} + \frac{a_2}{4} \ldots$

  $(m-1) \cdot 2 = a_1 + \frac{a_2}{2} + \frac{a_3}{4} + \ldots$

  dunque se $(m-1) \cdot 2 = 0 \rightarrow a_1 = 0$

  se $(m-1) \cdot 2 = 1 \rightarrow a_1 = 1$.

  quindi $a_1 \in \{0, 1\}$

e così via

(vedi libro sui DSP).

Esempi:

- Siano 8 i bit per l'esponente
- Siano 3 i bit per la mantissa.

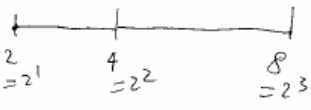
la costante che serve è $2^{n-1}-1 = 2^7-1 = 127$

per cui la corrispondenza fra Eff ed E è la seguente

Eff	E
~~0~~	~~NESSUNA (Eff=0 indica 0)~~
0	−126
1	−125
2	
3	
⋮	
253	126
254	127
255	NESSUNA (Eff = 2^n-1 indica NaN, positivo-infinito, negativo-infinito)

- l'intervallo per gli esponenti vale allora $[-126, +127]$

- è poco simmetrico per la densità:
  i numeri più grandi sono "meno densi"
  - Se m = 3 bit, ho 8 numeri nell'intervallo $[2,4]$, ma anche $[4,8]$.
    Questo è più largo, ma contiene gli stessi numeri (è meno denso)



- ovviamente, sia m che Eff vengono rappresentati in binario.

Esempio:

- Calcolare la rappresentazione ~~ft~~ di 123, con $m=3$ $e=3$

- PASSO 1: occorre normalizzare il numero, cioè rappresentarlo in forma $\pm 1.m \cdot 2^e$

  Per normalizzare si moltiplica e divide per 2.

  $123 = 61{,}5 \cdot 2 = 30{,}75 \cdot 2^2 = 15{,}375 \cdot 2^3 = 7{,}6875 \cdot 2^4$

  $= 3{,}84375 \cdot 2^5 = 1{,}921875 \cdot 2^6$

  allora $1{,}921875 \cdot 2^6$ è la rappresentazione normalizzata di 123

- PASSO 2: Calcolo di Eff

  - l'esponente di $1{,}921875 \cdot 2^{(6)}$ è 6
  - Eff = 6 + 127 = 133
  - $Eff|_{2\text{ cifre}}$ = 1000 0101 $(133_{|62})$

- PASSO 3: calcolo di m

è $1,921875$ $\Rightarrow a_0 = 1$ (si sapeva per la normalizzazione, non si ripete)

$1,921875 - a_0 = 0,921875$

$0,921875 * 2 = 1,84375 \Rightarrow a_{-1} = 1$

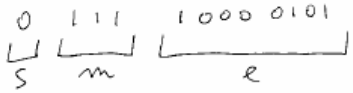
$1,84375 - a_{-1} = 0,84375$

$0,84375 * 2 = 1,6875 \Rightarrow a_{-2} = 1$

$0,6875 * 2 = 1,-- \Rightarrow a_{-3} = 1$

$\Rightarrow$ ci si ferma, perché la mantissa è a 3 cifre.

Doppia in base rappresentazione fp, con $m=3$, $e=8$, di 123 è



- Osservazione: qual è l'errore che si commette nel rappresentare 123 in fp con questi dati sopra?

Il numero che si rappresenta è

$$1.111|_2 \cdot 2^6 = \left(1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}\right)\cdot 64$$

$$= 64+32+16+8 = 120$$

- Sempre si commette un errore di 3

$\Rightarrow$ per aumentare la precisione, occorre aumentare i bit della mantissa (per questo, in IEEE 754, sono molti i bit riservati alla mantissa).

- Esercizi:

1) calcolare l'errore commesso nel rappresentare 123 in fp, con $e=8$ ed $m=3$, $m=4$, $m=7$

2) calcolare la rappresentazione fp, con $e=8$ ed $m=3$, di:

1.0
0,027
123,5
$3,1\cdot 10^{-4}$

calcolando anche l'errore commesso.