Rappresentazione binaria floating point

leo

March 5, 2023

- Utilizza il MSB per rappresentare il segno e può adottare sia MS sia CA2.
- un certo numero di bit per rappresentare la **mantissa** (*parte decimale in notazione scientifica*).
- un certo numero di bit per rappresentare **l'esponente** (2^{e_2}).

Prima di effettuare la conversione in binario del numero è necessario **normalizzarlo** ovvero riscriverlo in modo tale che abbia una sola cifra intera maggiore di 0.

Esempio: $0.87_{10} = 1.74{*}2^{-1} = 0\ -\ 0111\ 1110\ -\ 1011\ 1101\ 0111\ 0000\ 1010\ 010_2$

Dal 1985 viene utilizzato lo standard IEEE-754.
Un generico numero $X = 1, mantissa_2 * 2^{esponenete_2}$.

s	e	e	...	e	m	m	...	m
Segno	esponente				mantissa			

La notazione in virgola mobile a 32b utilizza (*precisione singola*):

- 1b per il segno
- 8b per l'esponente in notazione eccesso 8
- 23b per la mantissa

La notazione in virgola mobile a 64b utilizza (*doppia precisione*):

- 1b per il segno
- 11b per l'esponente in notazione eccesso 11
- 52b per la mantissa

Una rappresentazione comune è:

- esponente in **eccesso 127**
- parte intera in $[1, 2)$:
  - se più grande si effettua uno shift right e si incrementa l'esponente
  - se più piccolo si effettua uno shift left e si decrementa l'esponente

Per convertire un numero reale da binario a decimale si applica la formula:

$$(-1)^{segno} * (1 + mantissa) * 2^{esponenete-127}$$

1 Errore assoluto e relativo

Rappresentando un numero reale in virgola mobile si commette un errore di approsimazione.

In realtà viene rappresentato un numero n' con un numero limitato di cifre significative:

- errore assoluto = numero originale - numero rappresentato
- errore relativo = errore assoluto / numero originale