Rappresentazione di Numeri Reali

Luca Abeni

March 19, 2014

Rappresentazione dei Numeri Reali

- Rappresentare esattamente un numero reale $x \in \mathcal{R}$ non è possibile...
  - Numero reale: numero infinito di cifre decimali non periodiche...
  - ...Non rappresentabile con un numero finito di bit!
- Rappresentiamo un'approssimazione di numeri reali!!!
  - Alcuni numeri sono rappresentabili correttamente...
  - ...Altri no
- Rappresentazione in virgola mobile
  - Numero fisso di cifre significative, ma numero variabile di cifre dopo la virgola

Virgola Fissa e Mobile

- Numero reale su k cifre: si possono dedicare $k - f$ cifre alla parte intera e f cifre alla parte frazionaria: $\overbrace{\underbrace{c_{k-1}c_{k-2}\cdots c_f}_{k-f} . \underbrace{c_{f-1}\cdots c_0}_{f}}^{k}$
  - $x_{10} = \sum_{i=0}^{k} c_i B^{i-f} = \sum_{i=0}^{k} c_i B^i / B^f = \sum_{i=0}^{k} c_i B^i \cdot B^{-f}$
  - Convertiamo il numero in base 10 come se non avesse virgola e poi moltiplichiamo per B^{-f}: Virgola fissa!
- Virgola mobile: f non è fisso...
  - $x = M \cdot B^E$
  - M: mantissa
  - E: esponente

Virgola Mobile

- Se un numero x è rappresentato in virgola mobile su k bit
  - $x = M \cdot B^E$
  - m bit per la mantissa M...
  - ...e $e = k - m$ bit per l'esponente E
- L'esponente è quello che fa "muovere la virgola"
- Nei computer, $B = 2$ (rappresentazione binaria)
- Mantissa ed esponente possono essere positivi o negativi
  - Segno mantissa: segno di x
  - Esponente negativo: x piccolo; esponente positivo x grande
- Rappresentati in complemento a 2 o con altre tecniche...

Numeri Reali in C

- Standard IEEE (IEEE 754)
- Mantissa M rappresentata in modulo e segno (bit di segno s!)
- Mantissa “normalizzata”: $1.M$
- Esponente $-2^{e-1}+1 < E \leq 2^{e-1}-1$ rappresentato come $E' = E + 2^{e-1} - 1$ $(E' > 0)$
- $E' = 0$ ha un significato speciale: $E = -(2^{e-1} - 2$ e mantissa $0.M$. Anche $E' = 2^k - 1$ ha un significato speciale

$$E' > 0 : x = \begin{cases} 1.M \cdot 2^{E'-(2^{e-1}-1)} & \text{Se } s = 0 \\ -1.M \cdot 2^{E'-(2^{e-1}-1)} & \text{Se } s = 1 \end{cases}$$

$$E' = 0 : x = \begin{cases} 0.M \cdot 2^{-(2^{e-1}-2)} & \text{Se } s = 0 \\ -0.M \cdot 2^{-(2^{e-1}-2)} & \text{Se } s = 1 \end{cases}$$

Tipi Floating Point in C

- `float`: precisione singola
  - $k = 32$
  - $e = 8$, $m = 23$
- `double`: precisione doppia
  - $k = 64$
  - $e = 11$, $m = 52$
- `long double`: precisione estesa o quadrupla
  - $k = 80$ o $k = 128$
  - $e = 15$, $m = 64$ o $m = 112$

Esempio: Precisione Singola

- Approssimazioni di numeri reali rappresentate su 32 bit
- Esponente rappresentato su 8 bit
  - Se $E' \neq 0$, si sottrae $2^7 - 1 = 127$
  - Se $E' = 0$, esponente $-2^7 - 2 = -126$
- Mantissa rappresentata su 23 bit
- Bit s di segno

$$E' \neq 0 \Rightarrow x = (-1)^s \cdot 1.M \cdot 2^{E'-127}$$

$$E' = 0 \Rightarrow x = (-1)^s \cdot 0.M \cdot 2^{-126}$$

Precisione Singola - Valori Massimo e Minimo

- Massimo valore rappresentabile:
  - Massima mantissa: $M = \overbrace{1 \cdots 1}^{23}$
  - Massimo esponente: $E = 254 - 127 = 127$
  - $x = 1.\overbrace{1 \cdots 1}^{23} \cdot 2^{127} = (2 - 2^{-23}) \cdot 2^{127} \simeq 3.4 \cdot 10^{38}$
- Minimo valore rappresentabile: $- \simeq 3.4 \cdot 10^{38}$
- Minimo valore positivo rappresentabile:
  - Minima mantissa: $M = 1$
  - Esponente: $E' = 0 \Rightarrow E = -126$
  - $x = 0.\overbrace{0 \cdots 1}^{23} \cdot 2^{-126} = 2^{-23} 2^{-126} \simeq 1.4 \cdot 10^{-45}$