Sistemi numerici: numeri in virgola mobile
Esercizi risolti

1 Esercizio

Un numero relativo è rappresentato in virgola mobile secondo lo standard IEEE 754 su 32 bit nel seguente modo:

$$\begin{aligned} s &= 1 \\ e &= 10000111 \\ m &= 11011000000000000000000 \end{aligned}$$

Ricavare il corrispondente valore decimale.

Soluzione

Dato che:

$$e = 10000111_2 = 135_{10}$$

Si ha:

$$\begin{aligned} N &= (-1)^s \cdot 2^{(e-127)} \cdot 1.m \\ &= -1 \cdot 2^{135-127} \cdot 1.11011 \\ &= -1 \cdot 2^8 \cdot 1.11011 \\ &= -111011000_2 \\ &= -(2^8 + 2^7 + 2^6 + 2^4 + 2^3)_{10} \\ &= -472_{10} \end{aligned}$$

2 Esercizio

Convertire i seguenti numeri decimali in virgola mobile in singola precisione secondo lo standard IEEE 754:

1. -23.375_{10}
2. -127.25_{10}
3. $+131.5_{10}$
4. -300.25_{10}
5. -3.6_{10}

Soluzione

Procedendo in base alla definizione dello standard e indicando con

- N_{10} in numero originale in base dieci
- s il segno del numero
- e l'esponente del numero
- e_r l'esponente rappresentato in floating point
- m la mantissa della rappresentazione
- N_{FP} il numero nella rappresentazione in floating point

si ha:

1. $$\begin{aligned} N_{10} &= -23.375_{10} = -10111.011_2 = -1.0111011_2 \cdot 2^4 \\ s &= - = 1 \\ e &= 4 \\ e_r &= 4 + 127 = 131_{10} = 10000011_2 \\ m &= 1.0111011 = 0111011 \text{ (\textit{con hidden bit})} \\ N_{FP} &= 1 \mid 10000011 \mid 0111011000\ldots = C1BB0000_{16} \end{aligned}$$

2. $$\begin{aligned} N_{10} &= -127.25_{10} = -1111111.01_2 = -1.11111101_2 \cdot 2^6 \\ s &= - = 1 \\ e &= 6 \\ e_r &= 127 + 6 = 133_{10} = 10000101_2 \\ m &= 1.11111101 = 11111101 \text{ (\textit{con hidden bit})} \\ N_{FP} &= 1 \mid 10000101 \mid 111111010000\ldots = C2FE8000_{16} \end{aligned}$$

3. $$\begin{aligned} N_{10} &= 131.5_{10} = 10000011.1_2 = 1.00000111_2 \cdot 2^7 \\ s &= + = 0 \\ e &= 7 \\ e_r &= 127 + 7 = 134 = 10000110_2 \\ m &= 1.00000111 = 00000111 \text{ (\textit{con hidden bit})} \\ N_{FP} &= 0 \mid 10000110 \mid 00000111000\ldots = 43038000_{16} \end{aligned}$$

4. $$\begin{aligned} N_{10} &= -300.25_{10} = -100101100.01_2 = -1.0010110001_2 \cdot 2^8 \\ s &= - = 1 \\ e &= 8 \\ e_r &= 127 + 8 = 135 = 10000111_2 \\ m &= 1.0010110001 = 0010110001 \text{ (\textit{con hidden bit})} \\ N_{FP} &= 1 \mid 10000111 \mid 00101100010\ldots = C3962000_{16} \end{aligned}$$

5. In questo caso il numero decimale possiede infinite cifre nella rappresentazione binaria. La conversione in binario può essere interrotta una volta ottenute 24 cifre totali (comprensive di parte intera e parte frazionaria, considerando il bit nascosto) oppure una volta ottenuta la precisione desiderata.

$$\begin{array}{lcl} N_{10} & = & 3.6_{10} = 11.1001\ldots_2 = 1.11001\ldots_2 \cdot 2^1 \\ s & = & - = 1 \\ e & = & 1 \\ e_r & = & 127 + 1 = 128 = 10000000_2 \\ m & = & 1.11001\ldots = 11001\ldots \ (\textit{con hidden bit}) \\ N_{FP} & = & 1 \mid 10000000 \mid 11001100110011001100110 = C0666666_{16} \end{array}$$

3 Esercizio

Siano dati i seguenti numeri binari in virgola mobile in singola precisione secondo lo standard IEEE 754, espressi in base sedici:

1. $BE900000_{16}$
2. $438AA000_{16}$
3. $C301A000_{16}$
4. $C1FA8000_{16}$

Calcolare il corrispondente valore decimale.

Soluzione

Procedendo in base alla definizione dello standard e utilizzando la stessa notazione adottata nell'esercizio 2, si ha:

1.

$$\begin{array}{lcl} N_{FP} & = & BE900000_{16} = 1 \mid 01111101 \mid 00100000000000000000000_2 \\ m & = & 0010\ldots \ (\textit{con hidden bit}) \ = 1.001 \\ e_r & = & 01111101_2 = 125_{10} \\ e & = & 125 - 127 = -2 \\ s & = & 1 = - \\ N_{10} & = & -1.001 \cdot 2^{-2} = -0.01001 \\ & = & -(1 \cdot 2^{-2} + 1 \cdot 2^{-5}) = -0.28125_{10} \end{array}$$

2.

$$\begin{aligned}
N_{FP} &= 438AA000_{16} = 0 \mid 10000111 \mid 00010101010000000000000_2 \\
m &= 00010101010\ldots \ (\textit{con hidden bit}) \ = 1.0001010101 \\
e_r &= 10000111 = 135_{10} \\
e &= 135 - 127 = 8 \\
s &= 0 = + \\
N_{10} &= +1.0001010101 \cdot 2^8 = +100010101.01 \\
&= +(256 + 16 + 4 + 1 + 0.25) = +277.25_{10}
\end{aligned}$$

3.

$$\begin{aligned}
N_{FP} &= C301A000_{16} = 1 \mid 10000110 \mid 00000011010000000000000_2 \\
m &= 0000001101\ldots \ (\textit{con hidden bit}) \ = 1.0000001101 \\
e_r &= 10000110 = 134_{10} \\
e &= 134 - 127 = 7 \\
s &= 1 = - \\
N_{10} &= -1.0000001101 \cdot 2^7 = -10000001.101 \\
&= -(128 + 1 + 0.5 + 0.125) = -129.625_{10}
\end{aligned}$$

4.

$$\begin{aligned}
N_{FP} &= C1FA8000_{16} = 1 \mid 10000011 \mid 11110101000000000000000_2 \\
m &= 111101010\ldots \ (\textit{con hidden bit}) \ = 1.11110101 \\
e_r &= 10000011 = 131_{10} \\
e &= 131 - 127 = 4 \\
s &= 1 = - \\
N_{10} &= -1.11110101 \cdot 2^4 = -11111.0101 \\
&= -(16 + 8 + 4 + 2 + 1 + 0.25 + 0.0625) \\
&= -31.3125_{10}
\end{aligned}$$

4 Esercizio

Sia dato un formato in virgola mobile definito come segue:

- 1 bit di segno
- E bit di esponente in modulo e segno
- M bit di mantissa in forma normalizzata con bit nascosto.

Si voglia rappresentare in tale formato il numero $N = 266.1_{10}$ con precisione pari a 1/5. Si determini il numero minimo di bit da assegnare all'esponente e alla mantissa. Si rappresenti quindi il numero in tale formato.

Soluzione

Supponendo di rappresentare il numero 266.1 con n cifre per la parte intera e m per la parte frazionaria, si ha che:

$$n \geq \lceil log_2(266) \rceil = 9 \qquad \begin{aligned} 2^{-m} &\leq \tfrac{1}{5} \\ 2^m &\geq 5 \\ m &\geq 3 \end{aligned}$$

Occorrono quindi 9 cifre per la parte intera e 3 per la parte frazionaria di cui uno nascosto, quindi occorrono 11 bit di mantissa. Dato che:

:2	
266	
133	0
66	1
33	0
16	1
8	0
4	0
2	0
1	0
0	1

$$\begin{aligned} 0.1 \cdot 2 &= 0.2 \Rightarrow 0 \\ 0.2 \cdot 2 &= 0.4 \Rightarrow 0 \\ 0.4 \cdot 2 &= 0.8 \Rightarrow 0 \end{aligned}$$

segue che:

$$N = 266.1_{10} = 100001010.000 = 1.0000101000 \cdot 2^8$$

L'esponente pari a 8 richiede 5 bit per essere rappresentato in modulo e segno. La rappresentazione è la seguente:

$$\begin{aligned} s &= + = 0 \\ e_r &= 01000 \\ m &= 0000101000 \\ N &= 266.1_{10} = 0 \;\; 01000 \;\; 0000101000 \end{aligned}$$

5 Esercizio

Si debbano rappresentare dei numeri reali senza segno il cui valore assoluto sia sempre minore di $2 \cdot 10^6$ e la precisione sia la maggiore possibile. Quanti bit si possono dedicare alla parte intera e quanti alla parte frazionaria se si adotta una rappresentazione in virgola fissa su 32 bit? Giustificare la risposta.

Soluzione

Dato che:

$$2 \cdot 10^6 \simeq 2 \cdot 2^{20} = 2^{21}$$

e con n bit si rappresentano i numeri N il cui valore V_N è tale che

$$0 \leq V_N \leq 2^n - 1$$

in virgola fissa la parte intera deve essere rappresentata su 21 bit e, di conseguenza, alla parte frazionaria possono essere dedicati 11 bit.

6 Esercizio

Siano $x_1 = 100000000_{10}$ e $x_2 = 100000001_{10}$ numeri rappresentati in virgola mobile nel formato IEEE 754 a 32 bit. Quale sarà il valore di $y = x_1 - x_2$ supponendo di effettuare tale operazione nella stessa notazione? Giustificare la risposta.

Soluzione

Poiché:

$$x_1 = 100000000_{10} = 100 \cdot 10^6 \simeq 100 \cdot 2^{20}$$

x_1 richiede $7+20 = 27$ bit per essere rappresentato in virgola fissa. Poichè in virgola mobile su 32 bit si possono rappresentare solo 24 bit di mantissa, i valori x_1 e x_2 coincidono in tale rappresentazione. La loro differenza y risulterà quindi uguale a 0 essendo y ottenuto per differenza tra due valori identici.

7 Esercizio

Se si esegue l'operazione $32.416_{10} \cdot 128.776_{10}$ "sulla carta", il risultato avrà 6 cifre decimali. Se si opera mediante la rappresentazione in virgola mobile IEEE 754 su 32 bit, quante cifre decimali avrà il risultato?

Soluzione

Si osservi che non è necessario effettuare le conversioni. Il prodotto della parte intera

$$32 \cdot 128 = 2^5 \cdot 2^7 = 2^{12}$$

può essere rappresentato su 13 bit. Nella rappresentazione in virgola mobile si dispone di 24 bit di mantissa. Se 13 vengono utilizzati per la parte intera, alla parte frazionaria possono essere dedicti i restanti 11 bit. Questi corrispondono a circa $11/3.3 = 3$ cifre decimali.

8 Esercizio

Quanti bit occorrono per rapprentare due milioni di miliardi in binario puro? Se la stessa grandezza si rappresentasse nella notazione in virgola mobile standard IEEE 754 su 32 bit, quale sarebbe l'ordine di grandezza dell'errore commesso?

Soluzione

Dato che

$$2\ milioni\ di\ miliardi \;=\; 2 \cdot 10^6\ \ miliardi \;=\; 2 \cdot 10^6 \cdot 10^9 \;\simeq\; 2 \cdot 2^{20} \cdot 2^{30}$$

sono necessari 51 bit per rappresentare tale grandezza in binario puro.

Per esprimere un numero di 51 bit in virgola mobile, si deve avere esponente uguale a 2^{50} e la mantissa pari a $1.xxxxx$ (con 1 = hidden bit, x = 23 bit di mantissa = 0/1). Dato che l'ultimo bit della mantissa ha peso (relativo) uguale a 2^{-23}, il suo peso assoluto sarà pari a

$$2^{50} \cdot 2^{-23} \;=\; 2^{27} \;=\; 2^{-3} \cdot 2^{30} \;\simeq\; 0.125 \cdot 10^9$$

che è anche l'errore commesso.