

Soluzione esercizio 3

Il numero n è:

F F 5 A B 9
~~1111~~ ~~1111~~ 0101 1010 1011 1001

Trascurando i bit piu' significativi posti a 1 (tranne l'ultimo) e lasciando invariati i bit fino al primo 1 e complementando poi i successivi si ottiene il modulo del numero in questione:

01010010101000111

Spostando la virgola a sinistra di 15 posti (dividendo e moltiplicando per 2^{15}) risulta:

$1,010010101000111 * 2^{15}$

Quindi l'esponente da rappresentare è $(15)_{10}$.

15 ricade nell'intervallo $[-16, +15]$ per cui si ricava che il numero e di bit minimo per la rappresentazione dell'esponente in eccesso 2^{e-1} è dato da: $e = 5$.

La mantissa sarà rappresentata da $15 - e = 10$ bit.

Il numero rappresentato è quindi:

1 11111 0100101010

La parte non rappresentata è $0,000000000000111 * 2^{15}$ e coincide con l'errore assoluto, mentre l'errore relativo è $\epsilon_r = 2^{-13}$.

Soluzione esercizio 4

I numeri da rappresentare sono:

F A 5 3
 1111 1010 0101 0011
 segno esponente mantissa

L'esponente in eccesso 64 è 1111010 e in complemento a due equivale a complementare il bit più significativo, quindi 0111010 che in decimale vale 58.

In definitiva $m = 1,01010011 * 2^{58}$

F 9 F 2
 1111 1001 1111 0010
 segno esponente mantissa

L'esponente in eccesso 64 è 1111001 e in complemento a due equivale a complementare il bit più significativo, quindi 0111001 che in decimale vale 57.

In definitiva $n = 1,11110010 * 2^{57}$

Per eseguire la somma è necessario trasformare uno dei due numeri al fine di ottenere lo stesso esponente.

Trasformando il più piccolo (n) si ha:

$n = 0,111110010 * 2^{58}$

A questo punto è possibile sommare le mantisse:

1,01010011+
 0,11111001
 10,01001100

Si ha dunque: $h = m + n = 10,01001100 * 2^{58} = 1,001001100 * 2^{59}$

L'esponente del numero da rappresentare in eccesso 64 è

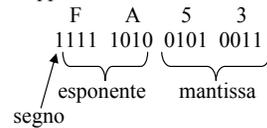
$(59+64)_{10} = (123)_{10} = (1111011)_2$

Il numero nella rappresentazione considerata è quindi

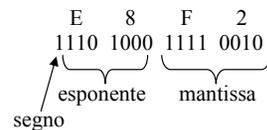
1111 1011 0010 0110
 segno esponente mantissa

che in esadecimale diventa FB26.

Consideriamo ora un'altra coppia di numeri:



L'esponente in eccesso 64 è 1111010 che in complemento a due equivale a complementare il bit più significativo, quindi 0111010 che in decimale vale a 58.
 In definitiva $m = 1,01010011 * 2^{58}$



L'esponente in eccesso 64 è 0101000 e in complemento a due equivale a complementare il bit più significativo, quindi 0111001 che in decimale vale a 40.
 In definitiva $n = 1,11110010 * 2^{40}$

Come prima, per poter eseguire la somma è necessario trasformare uno dei due numeri al fine di ottenere lo stesso esponente.

Trasformando il più piccolo (n) si ha (spostando la virgola a sinistra di 18 posti):
 $n = 0,000000000000000011110010 * 2^{58}$

A questo punto è possibile sommare le mantisse:

$$\begin{array}{r} 1,01010011+ \\ 0,00000000 \\ \hline 1,01010011 \end{array}$$

Si ha dunque: $h = m + n = 1,01010011 * 2^{58}$ che è proprio uguale al numero più grande.

Quello che si vuole mettere in evidenza in questo esercizio è che se gli esponenti dei due numeri rappresentati m ed n differiscono per un valore maggiore del numero di bit utilizzati per rappresentare l'esponente, il risultato della somma è pari al numero più grande rappresentato (in questo caso m).