Rappresentazione di un reale in virgola fissa

Sia dato il numero reale a = 9.7. Rappresentarlo su 16 bit in virgola fissa, con 8 bit di parte intera e 8 di parte frazionaria.

Soluzione

Prima di tutto occorre sapere se la parte intera, cioè 9, è rappresentabile su 8 bit. L'intervallo di rappresentabilità su 8 bit è [0, 255], che comprende 9. Quindi la parte intera è rappresentabile. Con il metodo del div&mod troviamo la sua rappresentazione:

q0 = 9

q1 = q0 div 2 = 4 **a0** = q0 mod 2 = **1**

q2 = q1 div 2 = 2 **a1** = q1 mod 2 = **0**

q3 = q2 div 2 = 1 **a2** = q2 mod 2 = **0**

q4 = q3 div 2 = **0** **a3** = q3 mod 2 = **1**

Quindi 9 = (**00001001**)$_2$ su 8 bit.

Andiamo a rappresentare la parte frazionaria, cioè 0.7, con il metodo "parte frazionaria – parte intera":

$f_0 = 0.7$

$f_{-1} = F(f_0 \cdot 2 = 0.7 \cdot 2 = \mathbf{1.4}) = 0.4$ $a_{-1} = I(1.4) = \mathbf{1}$

$f_{-2} = F(f_{-1} \cdot 2 = 0.4 \cdot 2 = \mathbf{0.8}) = 0.8$ $a_{-2} = I(0.8) = \mathbf{0}$

$f_{-3} = F(f_{-2} \cdot 2 = 0.8 \cdot 2 = \mathbf{1.6}) = 0.6$ $a_{-3} = I(1.6) = \mathbf{1}$

$f_{-4} = F(f_{-3} \cdot 2 = 0.6 \cdot 2 = \mathbf{1.2}) = 0.2$ $a_{-4} = I(1.2) = \mathbf{1}$

$f_{-5} = F(f_{-4} \cdot 2 = 0.2 \cdot 2 = \mathbf{0.4}) = 0.4$ $a_{-5} = I(0.4) = \mathbf{0}$

$f_{-6} = F(f_{-5} \cdot 2 = 0.4 \cdot 2 = \mathbf{0.8}) = 0.8$ $a_{-6} = I(0.8) = \mathbf{0}$

$f_{-7} = F(f_{-6} \cdot 2 = 0.8 \cdot 2 = \mathbf{1.6}) = 0.6$ $a_{-7} = I(1.6) = \mathbf{1}$

$f_{-8} = F(f_{-7} \cdot 2 = 0.6 \cdot 2 = \mathbf{1.2}) = 0.2$ $a_{-8} = I(1.2) = \mathbf{1}$

Non importa procedere oltre con il metodo, dato che abbiamo solo 8 bit per rappresentare la parte frazionaria. Quindi 0.7 = (**0.10110011…**)$_2$. I puntini di sospensione stanno a indicare che ci sarebbero altre cifre binarie ma che ignoriamo. La rappresentazione in virgola fissa di 0.7 non avrà quindi precisione infinita, ma avrà la parte frazionaria troncata.

Quindi, la rappresentazione finale sarà A = **00001001|10110011**.

Semplificazione

Notare che, durante il metodo "parte frazionaria – parte intera", i passi finali si ripetono uguali da un certo punto in poi, perché si ripete il valore di f. Questo ci consente di abbreviare il metodo in questo modo:

$f_0 = 0.7$

$f_{-1} = F(f_0 \cdot 2 = 0.7 \cdot 2 = \mathbf{1.4}) = 0.4$ $a_{-1} = I(1.4) = \mathbf{1}$

$f_{-2} = F(f_{-1} \cdot 2 = 0.4 \cdot 2 = \mathbf{0.8}) = 0.8$ $a_{-2} = I(0.8) = \mathbf{0}$

$f_{-3} = F(f_{-2} \cdot 2 = 0.8 \cdot 2 = \mathbf{1.6}) = 0.6$ $a_{-3} = I(1.6) = \mathbf{1}$

$f_{-4} = F(f_{-3} \cdot 2 = 0.6 \cdot 2 = \mathbf{1.2}) = 0.2$ $a_{-4} = I(1.2) = \mathbf{1}$

$f_{-5} = F(f_{-4} \cdot 2 = 0.2 \cdot 2 = \mathbf{0.4}) = 0.4$ $a_{-5} = I(0.4) = \mathbf{0}$ f_{-5} è uguale a f_{-1}, quindi il metodo si ripete uguale a partire dal passo successivo

$a_{-6} = a_{-2} = \mathbf{0}$

$a_{-7} = a_{-3} = \mathbf{1}$

$a_{-8} = a_{-4} = \mathbf{1}$

Notare anche che, siccome i passi del metodo si ripetono all'infinito, ci sono infinite cifre binarie dopo la virgola, che si ripetono periodicamente: 0.7 = (**0.10110011001100110011…**)$_2$. Un numero infatti può essere non periodico in base 10 ma diventarlo in base 2.

Somma di due reali tramite le loro rappresentazioni in virgola fissa

Siano date le rappresentazioni su sedici bit (8 parte intera, 8 parte frazionaria) dei reali a e b, rispettivamente A = 00011010|11010000 e B = 01000111|01010000. Si calcoli la rappresentazione C della somma c = a + b.

Soluzione

La somma delle rappresentazioni si fa in colonna, stando attenti ad "allineare" verticalmente le virgole di A e di B.

```
A = 00011010.11010000 +
B = 01000111.01010000 =

    ________________
      111111 1 1     (riporti)
    01100010.00100000
```

Non c'è riporto finale, quindi possiamo concludere che non c'è stato traboccamento. La parte intera e la parte frazionaria della rappresentazione C della somma c = a + b saranno quelle rispettivamente a sinistra e a destra della virgola. Quindi C = **01100010|00100000**.