Floating point Half-precision

Lo standard half-precision segue le stesse regole dello standard IEEE 754, utilizzando però la metà dei bit, cioè 16. Se ne utilizza sempre uno per il segno, 5 per l'esponente e i 10 restanti per la mantissa. All'esponente verrà sottratto 14.

$$\pm 1,M \times 2^{E-14}$$

M = 10 bit
E = 5 bit
Segno = 1 bit

Il numero più piccolo denormalizzato rappresentabile è 2^{-24} cioè $5{,}96 \times 10^{-8}$
Che sarebbe in binario 0000000000000001
Il numero più grande è 65504 cioè 0000010000000000

Floating point Double-precision

Lo standard double-precision segue le stesse regole dello standard IEEE 754, utilizzando però il doppio dei bit, cioè 64. Se ne utilizza sempre uno per il segno, 11 per l'esponente e i 52 restanti per la mantissa. All'esponente verrà sottratto 1023.

Addizione tra 2 floating point

Per eseguire l'addizione fra 2 numeri floating point vanno seguiti vari step, prendiamo per esempio i 2 numeri :

0x3FC00000
0x40500000

1) Prima di tutto estraiamo la mantissa e l'esponente

   3FC00000 = 00111111110000000000000000000000 cioè
   + 10000000000000000000000* 2^01111111= 1,5 * 2^127-127 = 1,5 * 2^0 = 1,5

   40500000 = 01000000010100000000000000000000 cioè

   + 10100000000000000000000* 2^10000000 = 1,625*2^128-127 =1,625*2^1 = 3,25

2) Ora che abbiamo le cifre decimali procediamo alla somma

   1,5+3,25 = 4,75

3) Possiamo anche fare la somma direttamente in binario

1,100 * 2^0 +
1,101 * 2^1 =

Per sommare 2 numeri devo prima portare gli esponenti allo stesso valore

SHIFTO la virgola del numero con l'esponente minore verso sinistra, ed aggiungo un unità all'esponente

1,100 * 2^0 = 0,110 * 2^1

Adesso eseguo la somma

```
  0,110 * 2^1 +
  1,101 * 2^1 =
--------------------
10,011 * 2^1
```

10,011 * 2^1 equivale a 2,375 * 2 cioè 4,75

Normalizzo la mantissa spostando la virgola verso sinistra ed aggiungendo un unità all'esponente
10,011 * 2^1 = 1,0011 * 2^2

1,0011 * 2^2in floating point è :

si ricordi di sommare 127 all'esponente

01000000100110000000000000000000
Cioè
0x40980000

Moltiplicazione tra 2 floating point

Prendiamo come esempio sempre i due numeri 1,5 * 2^0 e 1,375 * 2^1

Trasformiamo i 2 valori in binario, moltiplichiamoli trascurando gli esponenti, il risultato avrà come esponente quello maggiore.

1,5 = 1,100
1,375 = 1,011

Moltiplicazione in colonna :

1) Come primo step eseguiamo la moltiplicazione fra i 2 numeri trascurando la virgola :

```
 001100 *
 001011 =
--------
 001100 +
 011000 +
 000000 +
1100000 =
--------
10000100
```

2)Adesso del numero 10000100 bisogna capire dove mettere la virgola, per capirlo bisogna sommare la cardinalità dei numeri dopo la virgola dei valori iniziali.
1,100 = 3 numeri dopo la virgola
1,011 = 3 numeri dopo la virgola

3 + 3 = 6 La virgola verrà spostata di 6 posizioni a partire dal primo numero a destra

1 0, 0 0 0 1 0 0

Il risultato della moltiplicazione sarà quindi 10,011100 * 2^1. Normalizziamo il valore rendendo la parte intera della mantissa uguale ad 1 :

10,000100 * 2^1 = 1,0000100 * 2^2

Il nostro risultato è 1,0000100 * 2^2, trasformiamolo in IEEE 754 adesso :

binario : 01000000100001000000000000000000
esadecimale : 0x40940000