

假設一浮點數表示法如下：

Sign bit: 1 bit

Exponent bits: 4 bits

Mantissa bits: 3 bits

Total: 8 bits

Bias: $2^{4-1} - 1_{(10)} = 7_{(10)} = 111_{(2)}$

Sign	Exponent				Mantissa		

使用浮點數表示法計算 $\frac{1}{3}_{(10)} \times 3_{(10)}$ ：

$$\frac{1}{3}_{(10)} = 0.\bar{3}_{(10)} = 0.\overline{01}_{(2)} = 1.\overline{01}_{(2)} \times 2^{-2}_{(10)}$$

使用浮點數表示法：

Sign	Exponent				Mantissa		
0	0	1	0	1	0	1	1

Sign: 正數為 0，負數為 1。

Exponent: $-2_{(10)} + 7_{(10)} (bias) = 5_{(10)} = 101_{(2)}$

Mantissa:

(1) $1.\overline{01}_{(2)} - 1_{(2)} (leading\ 1\ is\ implicit) = 0.\overline{01}_{(2)}$

(2) Rounding:

有效位數	無效位數	Rounding 結果
XX0	0XXXX...	XX0
XX0	1XXXX...1...XXX	XX1
X00	1000.....(All 0 except first 1)	X00
X01	1000.....(All 0 except first 1)	X10

(3) 結果：011。

$$3_{(10)} = 11_{(2)} = 1.1_{(2)} \times 2^1_{(10)}$$

使用浮點數表示法：

Sign	Exponent				Mantissa		
0	1	0	0	0	1	0	0

Sign: 正數為 0，負數為 1。

Exponent: $1_{(10)} + 7_{(10)} (bias) = 8_{(10)} = 1000_{(2)}$

Mantissa:

(1) $1.1_{(2)} - 1_{(2)} (\text{leading 1 is implicit}) = 0.1_{(2)}$

(2) 結果：100。

浮點數相乘：

(1) Sign 做 XOR 運算： $0 \wedge 0 = 0$ 。

(2) Exponent 相加後減掉重複的一個 bias：

$$0101_{(2)}(5) + 1000_{(2)}(8) - 0111_{(2)}(7) = 0110_{(2)}(6)$$

(3) Mantissa 加上 leading 1 後相乘：

(3-1) $1.011_{(2)} \times 1.1_{(2)} = 10.0001_{(2)}$

	1.011
X	1.1
	1011
	10110
	10.0001

(3-2) Rounding:

有效位數	無效位數	Rounding 結果
XX0	0XXXX...	XX0
XX0	1XXXX...1...XXX	XX1
X00	1000.....(All 0 except first 1)	X00
X01	1000.....(All 0 except first 1)	X10

(3-3) 結果：10.0₍₂₎

(4) 使用浮點數表示法：

因為 $10 > 1$ （2 進制一位數最大可表達的數字），所以要將小數點向左移一位（Exponent + 1），新的 Exponent = 0111，新的 Mantissa = 000。

Sign	Exponent			Mantissa		
0	0	1	1	1	0	0

(5) 十進制表示法：

$$(1 + \text{Mantissa}) \times 2^{\text{Exponent} - \text{bias}} = (1 + 0.0) \times 2^{111 - 111} = 1 \quad (\#)$$