CH 4 浮點數 Floating Point Number

A. 定點數(Fixed Point Number)

B. 浮點數(Floating Point Number) – 正規化後，第一個位數為 0.xxxx 表示

C. 一般浮點數(Floating Point Number)格式 – sign bit、exponential、matnissa

D. IEEE754(國際格式)(有保留特殊值) – 1.xxx 表示；超碼 ： $2^{n-1}-1$；單 ：1S 8E 23M；雙 ：1S 11C 52M；全 1 全 0 不能用

E. 誤差 – 固有(Inhernet)、截尾(truncation)、捨棄(rounding)

---

1. 定點數(Fixed Point Number)

規定：小數點後的小數位數為固定位數

例：1234.5678901 ⇒ 定點取 3 位小取 ⇒ 1234.567

4 位小取 ⇒ 1234.5678

可參考台大 OCW CH02 計概

http://ocw.aca.ntu.edu.tw/ntu-ocw/ocw/cou/101S210/2

2. 浮點數(Floating Point Number)

(一) 數學之定義： $\pm(0.F)_b * b^{exp}$ $0<(0.F)_b<1$，exp=指數，b為基底(base)

(b 幾位，例如:10 進位，b = 10)

(二) 例: $+(1234.5678)_{10}$

$= +(0.12345678)_{10} * 10^{4} = +(0.012345678)_{10} * 10^{5} = +(0.00123456 78)_{10} * 10^{6}$ .....

(三) 以例知: 指數之有很多種格式表示同一數值.

但對電腦儲存, 希望有一種標準格式 ∴ 須有一個"正規化"動作.

(四) 以一般型之浮點數格式而言, 其正規化要求:

$(0.F)_b$ 中小數點後第一個位數值不得為 0 (零>0)

例: (1) $-(0.00001011)_2 \overset{正規化}{\Rightarrow} -(0.1011)_2 * 10^{-3}$

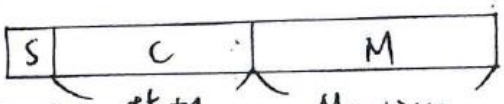
(2) $+(45.73)_8 \Rightarrow +(0.4573)_8 * 8^{2}$

(3) $-(0.05A7)_{16} \Rightarrow -(0.5A7)_{16} * 16^{-1}$

3. 一般浮點數(Floating Point Number)格式

(一) 先正規化, 得到 $\pm(0.F)_b * b^{exp}$

存入 Computer 內部之格式如下:



Sign bit 指數 Mantissa or Fraction (定點;浮點)

0: 正
1: 負

C: 指數

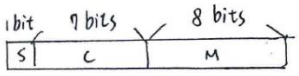
一般採用

2's complement or Excess-Code

表示正負指數

- 換成 2 進位，再換成 2 補數(下列例子)，再換成對應之進位系統，Ex:16 進位等等

(二) 例: 浮點數格式如下



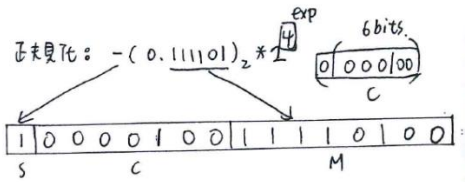
指數之進制. C用2's補數

表正負指數值.

先化到該進制再正規化.

(1) $-(1111.01)_2$ 之存入內容.

Ans: 正規化: $-(0.1111 01)_2 * 2^{4}$

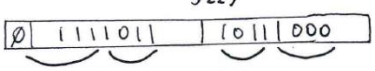


(2) $+(0.0000010111)_2$ 之存入內容

並以 16 進制 (Hex) 表示.

正規化: $+(0.10111)_2 * 2^{-5}$

exp=-5 ⇒ 6 bits 1 111011 (5 2's)



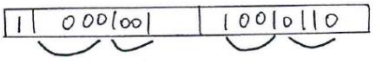
∴ $(7BB8)_{16}$

(3) $-(300)_{10}$ 之存入內容為 (?) Hex

$-(300)_{10} = -(100101100)_2$

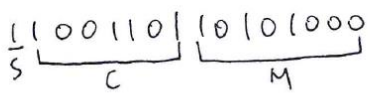
正規化: $-(0.100101100)_2 * 2^{9}$

exp=9 ⇒ 0 001001



∴ $(8996)_{16}$

(4) 內容 $(CDA8)_{16}$ 表值為?

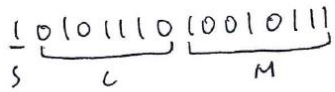


$C = 1\,001110$ → 2's 補原

$\therefore -(11001)_2 = -51$

$\therefore -(0.1010)_2 * 2^{-51}$ ※

(5) $(AE97)_{16}$



$C = 0\,101110$

$+(32+14) = +(46)$

$\therefore -(0.10101011)_2 * 2^{46}$

- 換成 2 進位，再轉成對應之 Excess($+2^{n-1}$)，再換成對應之進位系統，
  Ex:16 進位等等

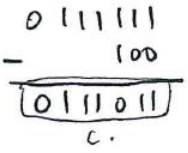
(1) $+(0.00000\,1011)_2$

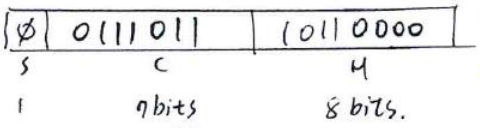
正規化: $+(0.1011)_2 * 2^{-5}$

exp = -5 用超碼表示. 因為 C 佔 7 bits.

∴ Excess-64 Code

∴ C = -5+64 = 63-4





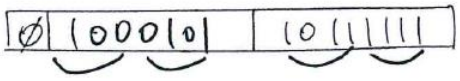
(依題目)

(2) $+(23\frac{7}{8})_{10}$ = 儲存內容 (以 Hex 表示)

$\Rightarrow +(10111.111)_2$

exp = 5 ∴ 5+64 = 1000101 (C)

正規化: $+(0.10111111)_2 * 2^5$



∴ $(45BF)_{16}$.

(1) 不看正負(最左位元)正數最大與負數最小相同，反之

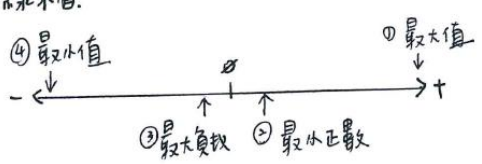
(一)求 ① 最大值(最大正數)

② 最小正數

③ 最大負數

④ 最小值(最小負數)

以數線來看.



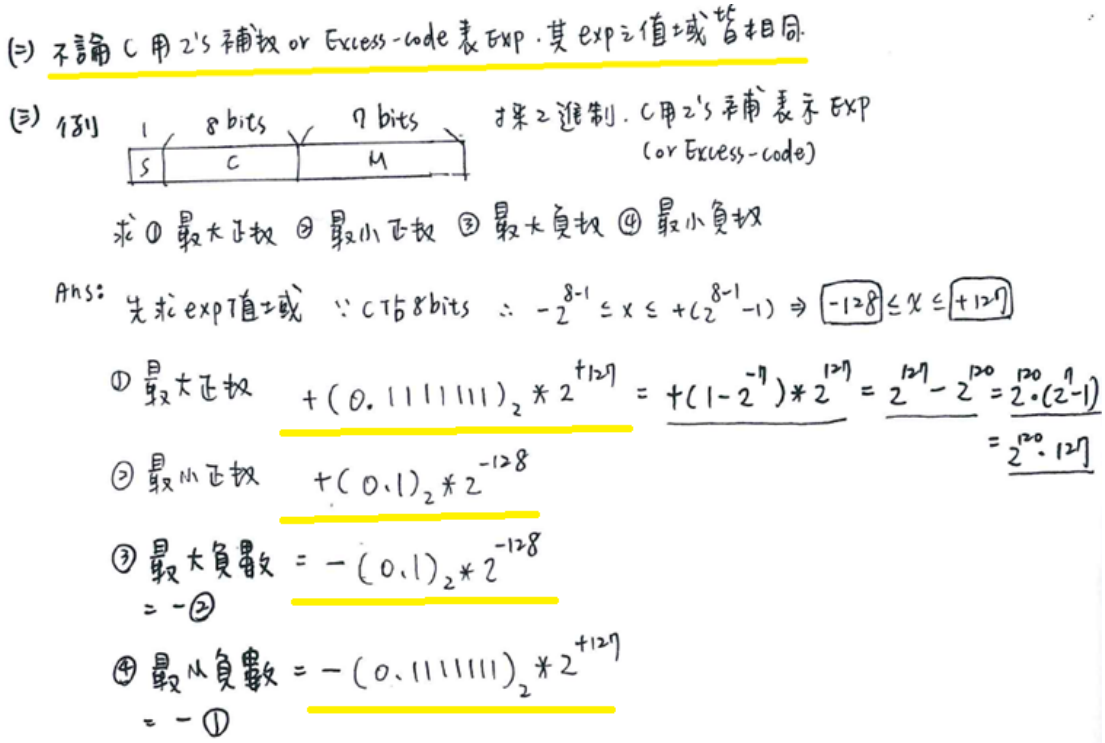
$|①| = |④|$, $|②| = |③|$

∵只要求出①及②. 自然得

③ = -② 及 ④ = -①

(2) 且 2 補數與 excess 值域相同 → - (2^{n-1}) ~ 2^{n-1} - 1

(3) 以最大數為例 0.11111111 * 2^{127} = (1 – 0.00000001) * 2^{127}= (1 – 2^{-7}) * 2^{127}



4. IEEE – 754(國際格式)(有保留特殊值)

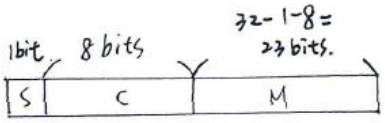
- 該標準之固定值為$2^{Exp-1} - 1$ → 跟一般 excess 相比還要多減 1
- Mantissa 越多越精確
- **單精度 single precision 32 bits(紅點為小數點)**

(1) Exp 部分加$2^{8-1} - 1 = 127$

(2) Exp： 全 1 或全 0 為特殊值，表示範圍

00000001(1) – 127 ≤ exp ≤ 11111110(254) – 127 → -126 ≤ exp ≤ 127

(一) ☆單精度 (32bits) 格式



其中, C是用 Excess-127 Code 表示指數.

Note: 比正常起碼少1.

IEEE754之正規化不同於一般型格式

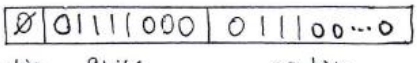
為: $\pm (1.F)_2 * 2^{exp}$

↳ 沒有規定小數點後第一個bit=0

例: 以IEEE754浮點格式格式表示:

(1) $+(0.0000010111)_2$ 答案用16進制表現.

Ans: 正規化: $+(1.0111)_2 * 2^{-7}$ exp=-7 ∴ C=-7+127=120=$(1111000)_2$



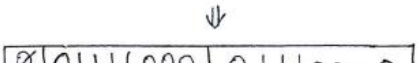
以16進制表示: 4組0000 16bits

$0.01111000.0111000\overbrace{\cdots 0}^{\text{4組0000 16bits}} \Rightarrow (3\ C\ 3\ 8\ 0000)_{16}$ ※.

例: 以IEEE754浮點格式格式表示:

(1) $+(0.0000010111)_2$ 答案用16進制表現.

Ans: 正規化: $+(1.0111)_2 * 2^{-7}$ exp=-7 ∴ C=-7+127=120=$(1111000)_2$



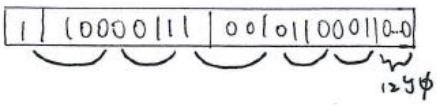
以16進制表示: 4組0000 16bits

$0.01111000.0111000\overbrace{\cdots 0}^{\text{4組0000 16bits}} \Rightarrow (3\ C\ 3\ 8\ 0000)_{16}$ ※.

(2) $-(300\frac{3}{8})_{10}$ 之內容?

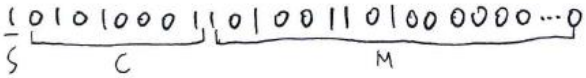
Ans: ⇒ $-(100101100.011)_2$

正規化: $-(1.00101100011)_2 * 2^{8}$ exp=8 ∴ C=8+127=128+7=$(10000111)_2$



⇒ $(C3963000)_{Hex}$ ※

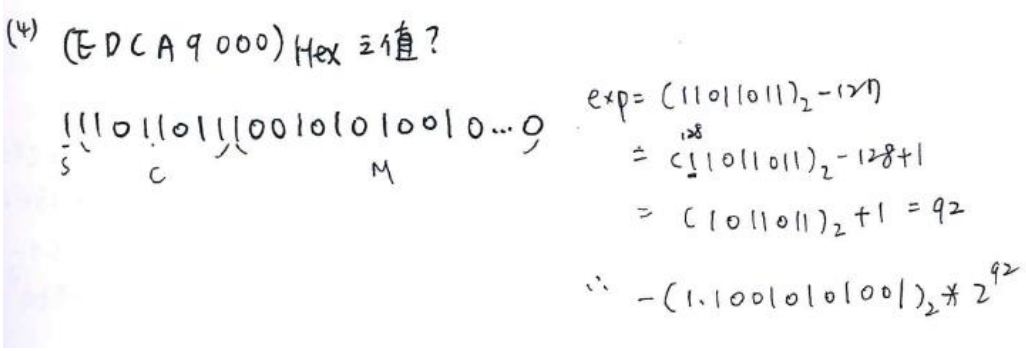
(3) $(A8D34000)_{Hex}$ 之值?



exp=C-127 = $(01010001)_2$ - 127

= $5*2^4+1-127=-46$

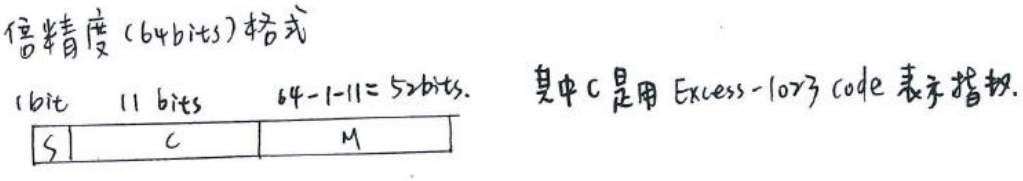
∴ $-(1.10100110)_2 * 2^{-46}$



- **倍精度 double precision 64 bits**

(1) Exp 部分加2^{11-1} – 1 = 1023

(2) Exp： 全 1 或全 0 為特殊值，表示範圍:00000000001(1) - 1023<= exp <= 11111111110(2046) – 1023 → -1022 <= exp <= +1023



- **特殊值規定**

exp	mantissa	result
全 1	為 0	正無窮大
全 1	非 0	NaN(未定義或不可表示)
全 0	為 0	0
全 0	非 0	無法正規化

- **誤差種類**

(1) 固有(Inhernet)誤差(Error)，ex. 圓周率

(2) 截尾誤差(Trucation Error)

(3) 捨棄誤差(Rounding Error)，ex.四捨五入行程之誤差，5/2=2