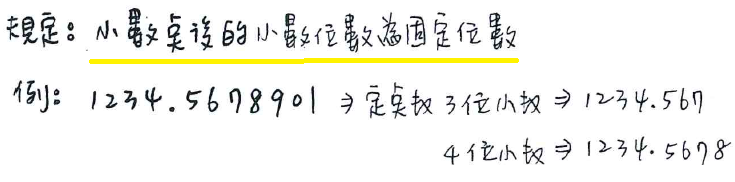
**CH 4 浮點數 Floating Point Number**

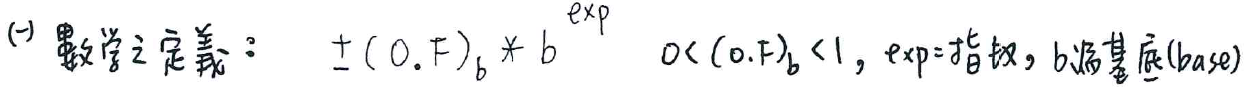
1. 定點數(Fixed Point Number)
2. 浮點數(Floating Point Number) – 正規化後，第一個位數為0.xxxx表示
3. 一般浮點數(Floating Point Number)格式 – sign bit、exponential、matnissa
4. IEEE754(國際格式)(有保留特殊值) –1.xxx表示；超碼 : – 1；單 : 1S 8E 23M；雙 : 1S 11C 52M；全1全0不能用
5. 誤差 – 固有(Inhernet)、截尾(truncation)、捨棄(rounding)
6. **定點數(Fixed Point Number)**

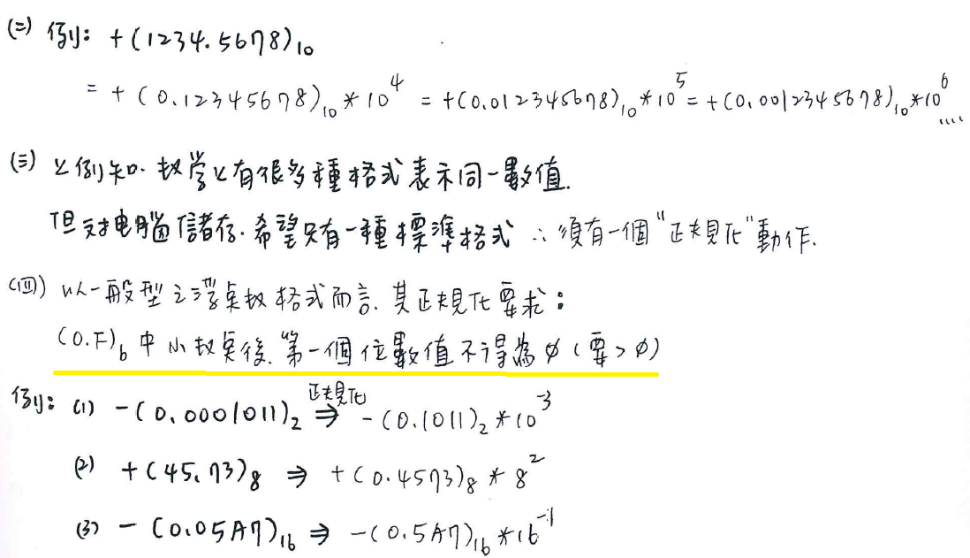


可參考台大OCW CH02 計概

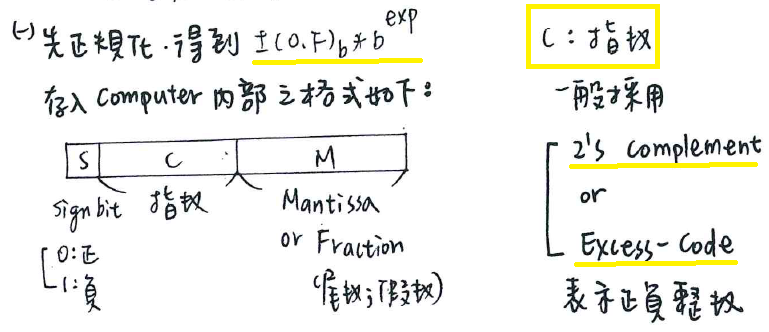
<http://ocw.aca.ntu.edu.tw/ntu-ocw/ocw/cou/101S210/2>

1. **浮點數(Floating Point Number)**

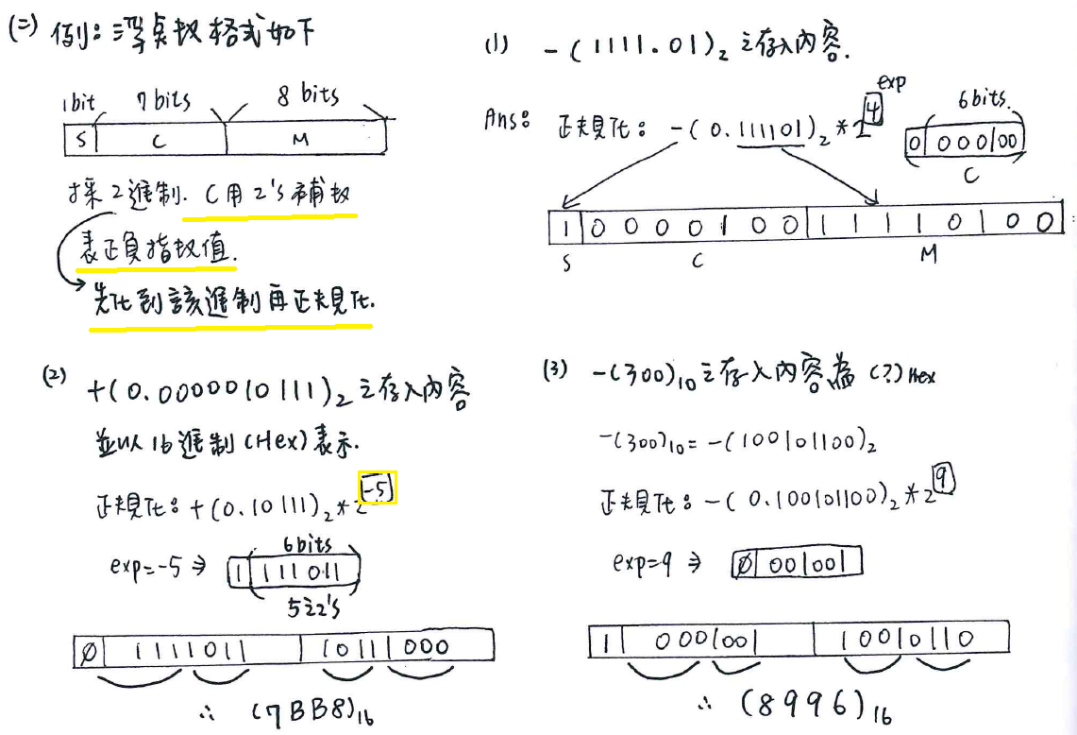


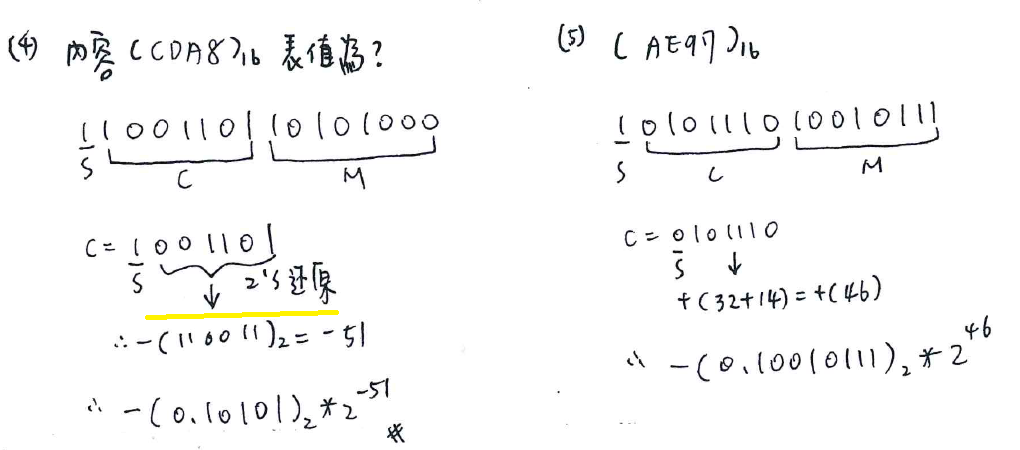
(b幾位，例如:10進位，b = 10)

1. **一般浮點數(Floating Point Number)格式**

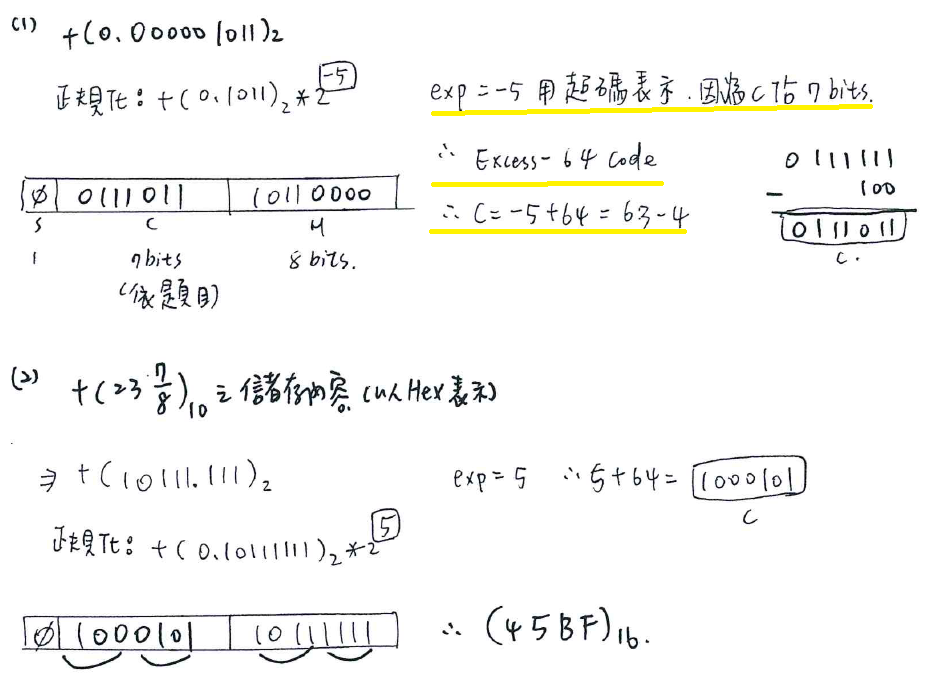


* 換成2進位，再換成2補數(下列例子)，再換成對應之進位系統，Ex:16進位等等

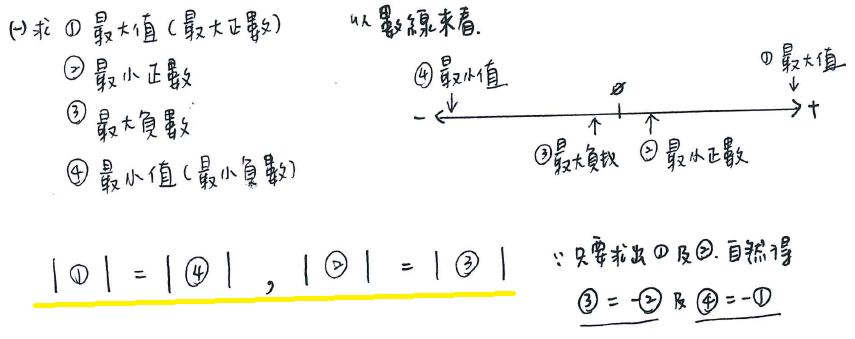


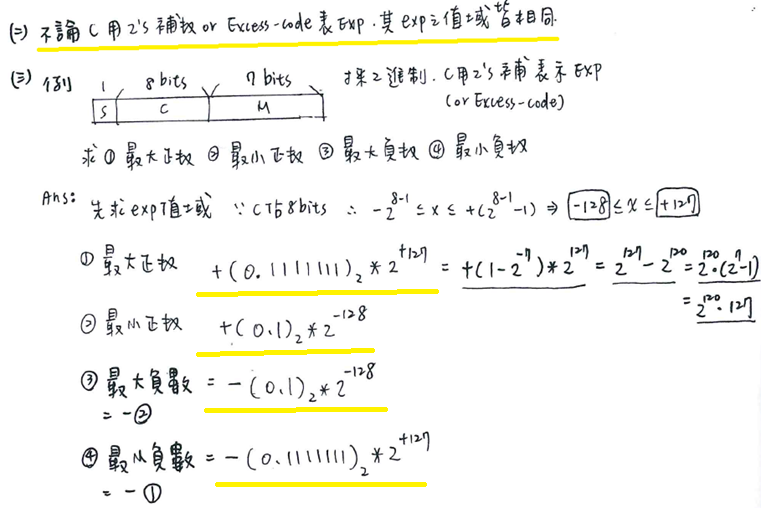


* 換成2進位，再轉成對應之Excess(+，再換成對應之進位系統，Ex:16進位等等



* 1. 不看正負(最左位元)正數最大與負數最小相同，反之

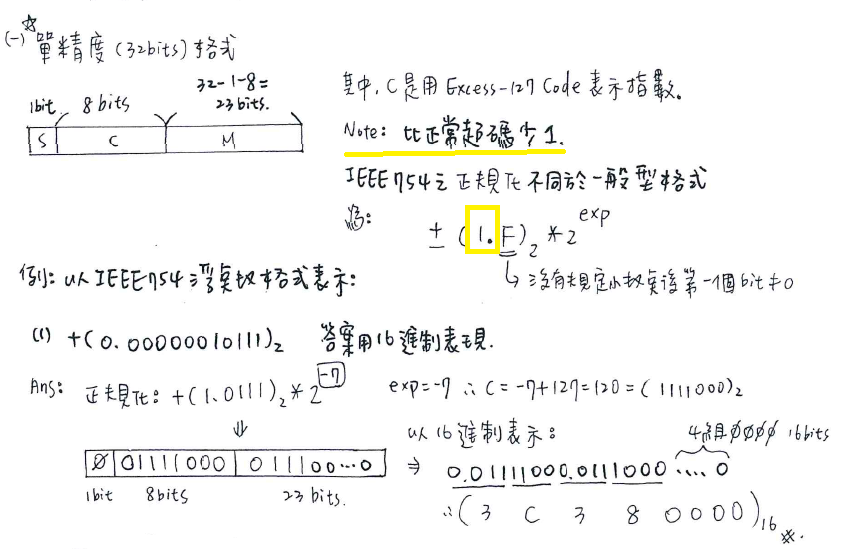


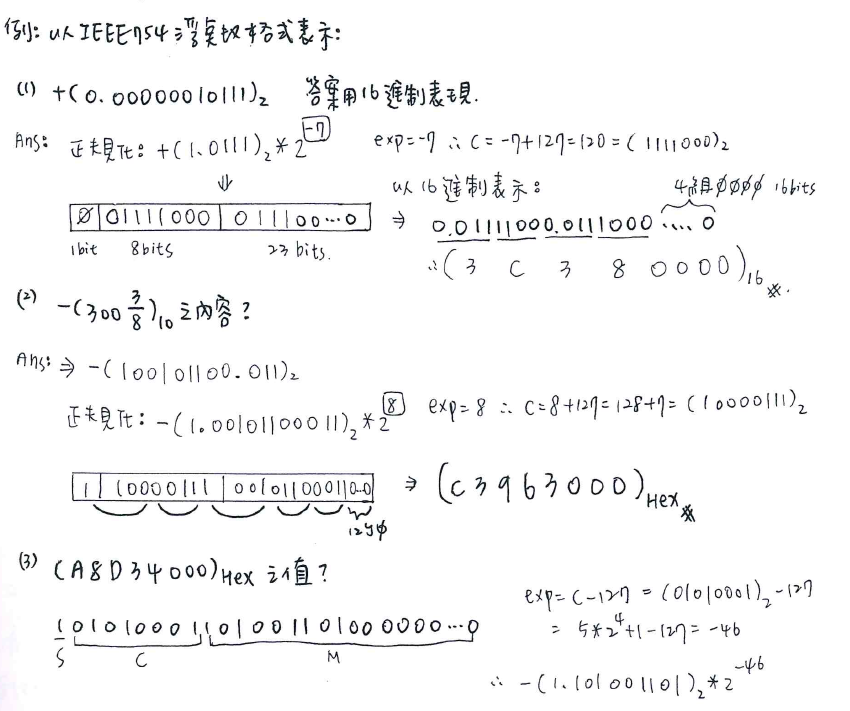
* 1. 且2補數與excess值域相同 🡪 - ( ) ~ - 1
  2. 以最大數為例0.11111111 \* = (1 – 0.00000001) \* = ( 1 – ) \* 

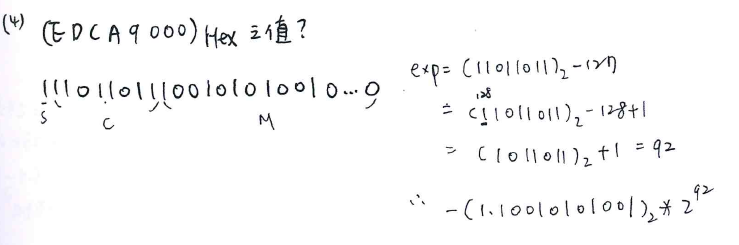
1. **IEEE – 754(國際格式)(有保留特殊值)**

* 該標準之固定值為 – 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 – 1 = 127
  2. Exp : 全1或全0為特殊值，表示範

00000001(1) – 127 exp 11111110(254) – 127 🡪 -126 exp 127

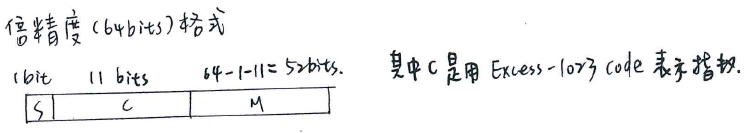






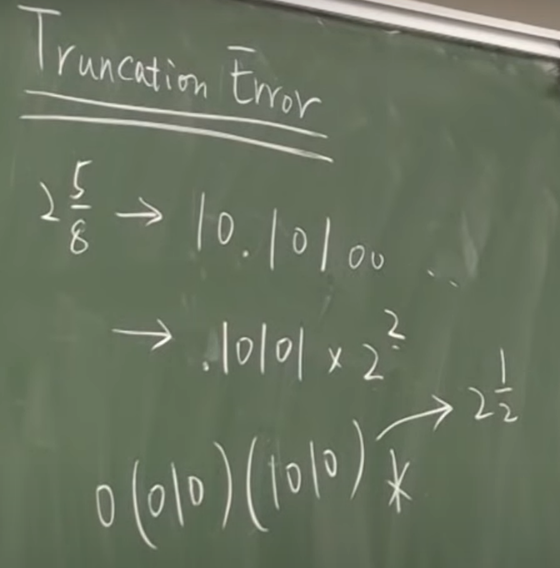
* **倍精度 double precision 64 bits**

1. Exp 部分加 – 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