(1). IEEE 754 表示法

設以科學記號表示 = 1位小數

正規化表示：

$$-\underbrace{1.1011}_{\text{significand}} \times \underbrace{2^{-5}}_{\text{base}}$$

sign: −　fraction: 1011　exponent: -5

A designer need to find a compromise between fraction & exponent

(1). fraction 影響了可表示之數的 precision

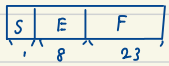
(2). exponent = 可表示之數的 range

∵ 字組大小是固定的，sign + fraction + exponent = fixed length

#. Good design demands good compromise

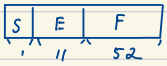
IEEE 754 有兩種精度表示法：① single precision ② double precision

① single precision：32 bit



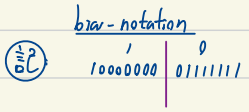
其中：S: sign, E: exponent F: fraction

② double precision：64 bit



其中：E 欄位數值是利用 bias-notation 表示

bias-notation



(記)

Why?: 因為在做浮點數加法時，會先比較 exponent 是否相同，若不同，需要 shift 至相同才能做加法

∴ 需要大量 exponent 的比較，又 bias-notation 的大小排序 和 unsign number 相同，可使用無号數比較器來做比較加速比較速度，

題目：(1). 將 10進位小數轉至 IEEE 754

step: I. 先將 10進位轉至 2進位

II. 轉至正規化表示，並以 IEEE 754 表示

eg $-13.8125_{(10)} \rightarrow -1101.1101_{(2)} \rightarrow -1.101101 \times 2^{3}$

S: 1　E: 3 = 10000010　F: 10110100---0

⇒ 1 10000010 10110100---0

(2). 給定 IEEE 754，求 10進位數

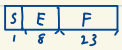
$1.0000\cdots0 \times 2^{1}$

☆☆☆☆☆

IEEE 754 之編碼情形 & 數值範圍

編碼情形共有5種：① ±0：S 000…0 0…0

single precision:



② ±∞：S 11…1 0…0

③ NaN：S 11…1 nonzero

NaN 的情形有：(1). 0/0　(2). 0×∞　(3). NaN 相關運算　(4). ∞ + −∞

④ 正規化數：$\pm 1.xxx \times 2^{E}$

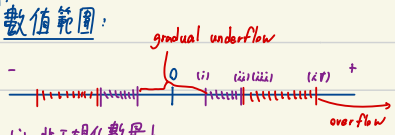
$1 \le E \le 254$，F 為任何數

⑤ 非正規化數：$\pm 0.xxxx \times 2^{1-B}$

S 000…0 nonzero

☆

數值範圍：



(1). 非正規化數最小：

0 000…0 00…01　$0.0\ldots01 \times 2^{-126} = 2^{-149}$　(23)

(2). 非正規化數最大：

0 000…0 1.…1　$0.1\ldots11 \times 2^{-126} = (1-2^{-23}) \times 2^{-126} =$

(3). 正規化數最小

0 000…1 0.…0　$1.0\ldots0 \times 2^{-126} = 2^{-126} = 2 \cdot 2^{-127} = 2 \cdot 10^{-38}$

(4). 正規化數最大

0 1…0 1.…1　$1.1\ldots1 \times 2^{254-127} = 1.1\ldots1 \times 2^{127}$

其中：$1.1\ldots1 \times 2^{127} = (2-2^{-23}) \times 2^{127}$

$\fallingdotseq 2 \cdot 2^{127} = 2 \cdot 10^{127 \cdot \log_{10} 2}$　⇒ 一個 10進位 bit 等於 $\log_2 10$ 個 2進位 bit

$\fallingdotseq 2 \cdot 10^{38}$

∵ 一般數值運算很本用到非正規化數那麼小的數 ∴ MIPS 中不支援，會為 underflow

⇒ 整數不會有 underflow 問題

課本無列到 denormalized number

∴ 小於 2^{-126} 的數皆視為 underflow

例：P217 3題　已知 single precision

(1). $-13.5_{(10)} \rightarrow -1101.1_{(2)} \rightarrow -1.1011 \times 2^{3}$

1 - 10000000

3 - 10000010　⇒ 1 10000010 10110000---0

(2). 最大正規化數：

0 11111110 1111···1　⇒ $1.111\cdots1 \times 2^{127}$

$= (2-2^{-23}) \times 2^{127}$

(3). 最小非正規化數：

0 000…0 00…01　⇒ $0.00\ldots01 \times 2^{-126}$

$= 2^{-23} \times 2^{-126} = 2^{-149}$

(4). E 全 1 ┌ F 為非0 ⇒ NAN

└ F 為全0 ⇒ ±∞

P261 57題：double ⇒ E 欄位 11個 bit　$0 \sim 2^{11}-1$ ⇒ 0~2047

0 ⇒ preserve for ±0, denormalized

2047 ⇒ preserve for ±∞, NAN

1~2046 ⇒ normalized number.

(補)

IBM format：

IBM format 為 excess-64

以 16 為底且 exponent 大小為 7 bit

∴ 1000000 ⟶ 0

∵ 為 base 16，無法保證 significand 小數點左邊必為 1

∴ 設為 0.正規化數為：$\pm 0.xxx\cdots x \times 16^{E}$

ex $-938.8125 \longrightarrow -1110\,1010.1101 \times 2^{0}$

base 16 ⟶ $-3AA.D \times 16^{0}$

Normalize：$-0.3AAD \times 16^{3}$

又 1000000 = 0 ∴ 3 = 1000011

⇒ IBM format：$\underbrace{1}_{1}\ \underbrace{1000011}_{7}\ \underbrace{00111010101011010\cdots0}_{24}$