Floating-point & BF16

DD Final Project (#10)

助教:胡祐嘉、劉宸彥

Department of Electrical Engineering and SoC Research Center National Chung Cheng University

Outline

- 課程目的
- IEEE754浮點數表示法
- BF16浮點數表示法
- Lab作業
- 課程評分
- 附錄

課程目的

經過了先前的實驗課，我們已經了解如何設計整數硬體，在本次課程中，同學們將:

- 學習IEEE754浮點數表示法
- 學習BF16浮點數表示法

IEEE754浮點數表示法 (1 / 4)

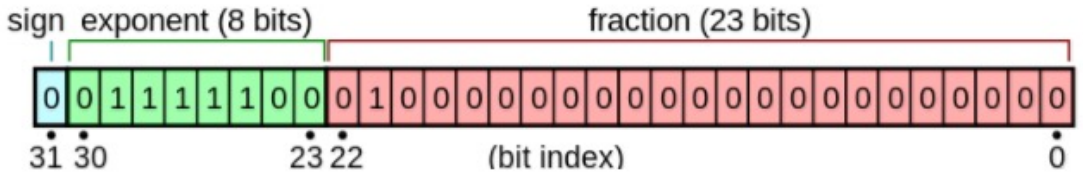
- IEEE 二進位浮點數算術標準（IEEE 754）是當前最廣泛使用的浮點數運算標準，在 IEEE 754 中表示浮點數值的方式，包含半精確度（16 位元）、單精確度（32 位元）、雙精確度（64 位元）等等。
- 其浮點數表示為：Value = Sign × Exponent × Fraction

IEEE754浮點數表示法 (2 / 4)

■ Sign為符號位，以0表示正值，1表示負值

■ Exponent為二進位科學計數法表示下的指數值加上指數偏移值
因為IEEE 754中以無號整數（Unsigned Integer）表示指數，其中一半值域在表示負數，因此將$2^{e-1}-1$定為指數偏移值（Exponent Bias），其中e為儲存指數的位元長度。

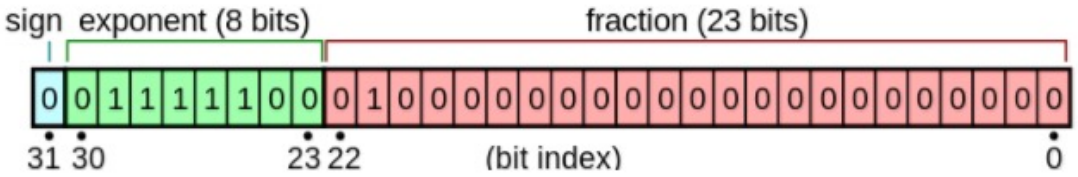
■ Fraction
當浮點數的指數部分編碼值在 0 < Exponent ≤ 2^e-2之間，則Fraction值為二進位科學計數法的尾數（Mantissa），亦即1.Fraction
如果指數部分編碼值是0，二進位科學計數法的尾數部分非零，則該實際值比前述涵蓋情況更接近0，因此其Fraction代表的值實際為0.Fraction



單精確度浮點數表示法

IEEE754浮點數表示法 (3 / 4)

以單精確度浮點數為例，在 32 bits 中，我們使用 1 bit 表示正值或負值，8 bits 表示指數，23 bits 表示尾數精度



Sign = +1

Exponent = $(01111100)_2 - 127 = -3$

Fraction

$= 1 + (0.01000000000000000000000)_2$

$= 1 + 2^{-2}$

$= 1.25$

Value = $(+1) \times 1.25 \times 2^{-3} = + 0.15625$

IEEE754浮點數表示法 (4 / 4)

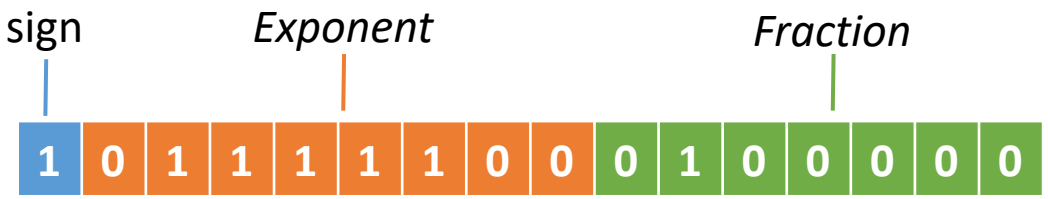
特殊值

如果指數是0且Fraction亦為0，該值為正負0（視Sign Bit而定）
如果指數 = 2^e-1且Fraction為0，該值為正負無限大（視Sign Bit而定）
如果指數 = 2^e-1且Fraction不為0，表示該不為一個數（NaN）
總結規則如下：

形式	指數	小數部分
零	0	0
非正規形式	0	大於0小於1（0.*Fraction*）
正規形式	1到$2^e - 1$	大於等於1小於2（1.*Fraction*）
無窮	$2^e - 1$	0
NaN	$2^e - 1$	非0

BF16浮點數表示法 (BFloat16)

BF16主要概念在於透過降低數字的精度，從而減少運算資源和功耗。
在 BF16中，我們使用 1 bit 表示正值或負值，8 bits 表示指數，7 bits 表示尾數精度



Sign = +1

Exponent = $(01111100)_2 - 127 = -3$

Fraction

$= 1 + (0.0100000)_2$

$= 1 + 2^{-2}$

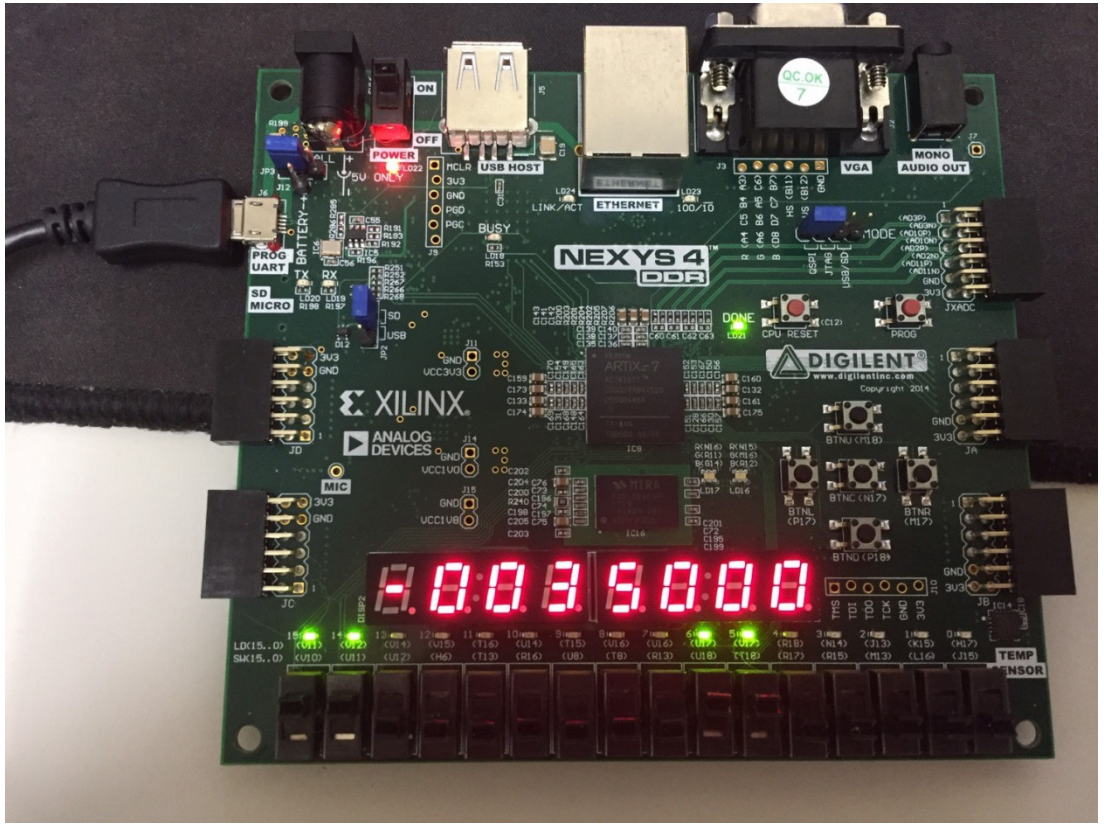
$= 1.25$

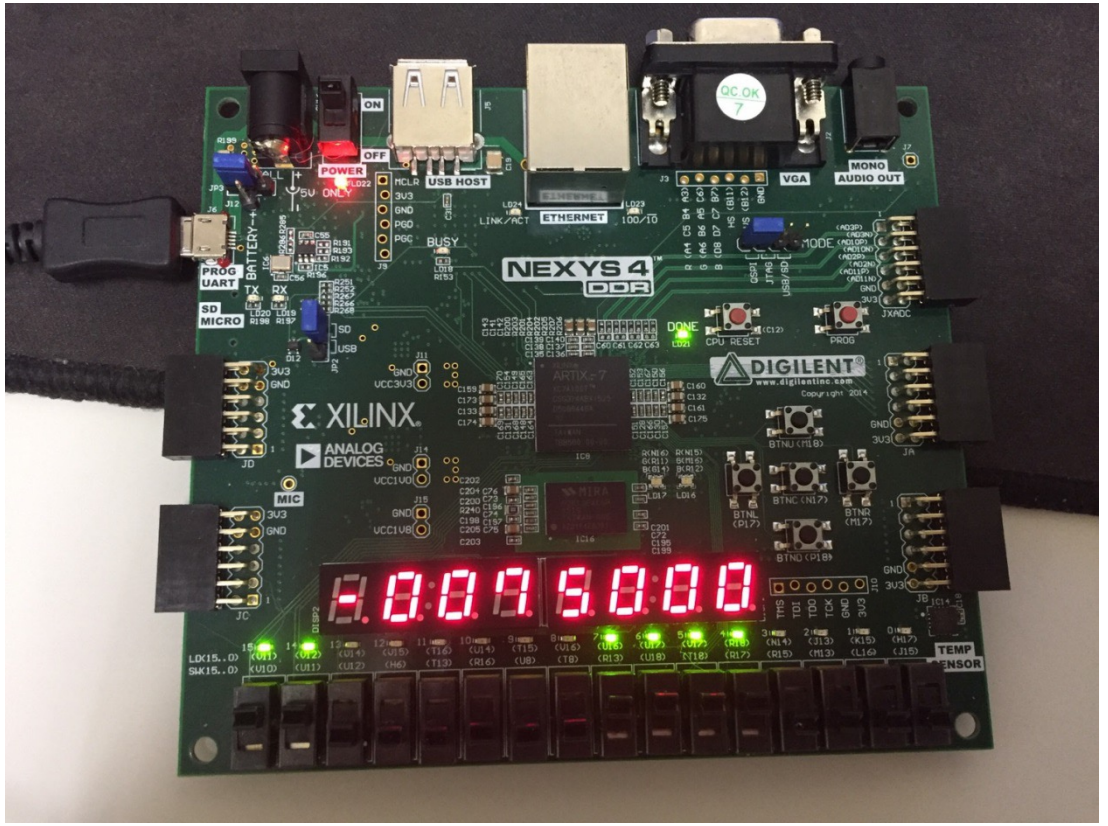
Value = $(+1) \times 1.25 \times 2^{-3} = +0.15625$

LAB作業

請同學透過FPGA版上的switch輸入一BF16浮點數，並將其顯示在七段顯示器上

1. 若值為負，第一個七段顯示器顯示負號，若值為正則不顯示
2. 第二到四個七段顯示器顯示整數位，後四個七段顯示器顯示小數位
3. 超出最大值(999.9999)或最小值(-999.9999)時顯示FFFF FFFF





課程評分

- Demo時間:6/15，6/17
- Demo地點:資工館501A
- 評分方式

1. 輸入一BF16浮點數，按下N17將其顯示在七段顯示器上 (60%)
2. 按下P17按鈕將輸入數值+7，並將其顯示在七段顯示器上 (20%)
   按下M17按鈕將輸入數值*3，並將其顯示在七段顯示器上 (20%)

記得填寫意見回饋表，否則不予以計分