應用於神經網路之電阻式記憶體內運算

ReRAM-based Computing-in-Memory Architecture for CNN

組別：A57　黃詩瑜、劉家蓁
指導老師：張孟凡 教授

Abstract

隨著 AI 人工智慧時代來臨，大量數據與資料密集應用，電腦對於運算速度要求提高，為了解決馮紐曼瓶頸（Von Neumann Bottleneck），記憶體內運算應運而生，將邏輯運算於記憶體內完成，大幅減少資料搬移的時間與功耗，提高運算效能。

此研究利用電阻式記憶體（ReRAM）為架構，實現能加速卷積神經網路運算的硬體電路，電阻式記憶體具有非揮發性、低功耗、高速之優點，適合應用於邊緣運算。

本篇採用的架構利用電阻式記憶體內運算(ReRAM Computing-in-Memory)來執行 2-bit input, 4-bit weight 的乘加運算(MAC, Multiply Accumulate)。ReRAM memory cell 的 BL 與 SL 間的跨壓作為 2-bit input，4-bit weight 的部分則以縱向的方式儲存於 ReRAM memory cell 中。將電流總和透過 down-scaling current mirror 做 1/8、3/8 或 1 倍的倍數處理並累積於電容，後 3 bits weight 累積於一個電容，Weight[3]（MSB）累積於另一電容，再將兩個電容上的電壓轉換成 digital signal 後，後 3 bits weight 的 digital signal 減去 MSB 的 digital signal 即為最後的結果，也就是 MAC 值。

Introduction

為了解決 Von Neumann bottleneck，記憶體內運算(Computing-in-Memory)開始受到關注，希望能提升電腦運算的整體效率。本篇利用電阻式記憶體內運算(ReRAM-based Computing-in-Memory)來執行 2-bit input, 4-bit weight 的乘加運算(MAC, Multiply Accumulate)。

ReRAM memory cell 一般分為高阻值狀態(HRS, High Resistance State)和低阻值狀態(LRS, Low Resistance State)，但因為我們沒有辦法取得此一元件來進行模擬，因此本篇以 1000k ohm 和 3k ohm 的電阻串連一個 NMOS 來模擬 HRS cell 及 LRS cell，以 WL 控制 NMOS 的開關來選擇 cell、NMOS 端接 SL、電阻端接 BL。

本篇的 4-bit weight 以 2’s complement 的形式來表示，也就是 weight 的範圍介於 -8 至 7，這 4 bits 儲存在 ReRAM cell 中，以 LRS cell 表示「1」，HRS cell 表示「0」。假設流經 LRS cell 的電流為 I_L、HRS cell 的電流為 I_H，所有 cell 的電流總和決定 mac 值。理論上 I_H 應為 0，為了使 I_H 等於 0，以免影響到 mac 運

算結果，我們選擇 1000k ohm 的電阻作為 HRS cell。2-bit input 是從 4 種電壓中：0.9V、0.75V、0.6V 以及 0.45V 選擇其一給在 BL 上，8 條 BL 各自選擇不同的電壓。SL 則是維持在 0.9V，如此每個 ReRAM cell 能擁有 0V、0.15V、0.3V 或 0.45V 不同的跨壓，以此得到 0、1、2、3 四種 input。

我們使用的 Memory Array 大小為 8*8，共 64 個 cell，而每個 weight 需要 4 個 cell 來儲存，因此整個 array 全部可存 16 個 weight。常見的 weight 儲存方式是橫向的為一組，這裡則是以縱向的 4 個 cell 為一組，如下方 Fig.1 所示。

為了給予 weight 每個 bit 應有的權重，此架構利用 WL 開啟的時機配合不同倍數的 down-scaling current mirror 來達成目的，如下方 Fig.2 所示。另外，因為 weight 是以 2's complement 的形式儲存，所以會有負的 MAC 值，為了區分正負，我們將電流分成兩部分處理，C1 累積後 3 個 bits 的電流，C2 則儲存了 MSB 的電流，先將 C1 上累積的跨壓送至 voltage type sense amplifier 做 sensing，並將結果以 digital 的形式儲存，再來對 C2 做 sensing，一樣轉成 digital 的形式，將這兩部份結果相減即能得到 MAC 值。

下方 Fig.3 為我們測試 8 個 input 與 8 組 weight 的乘積運算結果。Fig.4 為任選一組 input 和 weight 做完整的乘加運算，其中 3-bit weight 電壓為 1.236V，MSB 電壓為 1.463V，各自經過 SA sensing 後可得到的 digital value 為 46 與 56，相減得 MAC = -10。

此架構能在記憶體內做 8 個 input 與八組 4-bit weight 的乘積運算，節省了資料在記憶體與 CPU 間傳輸的時間。

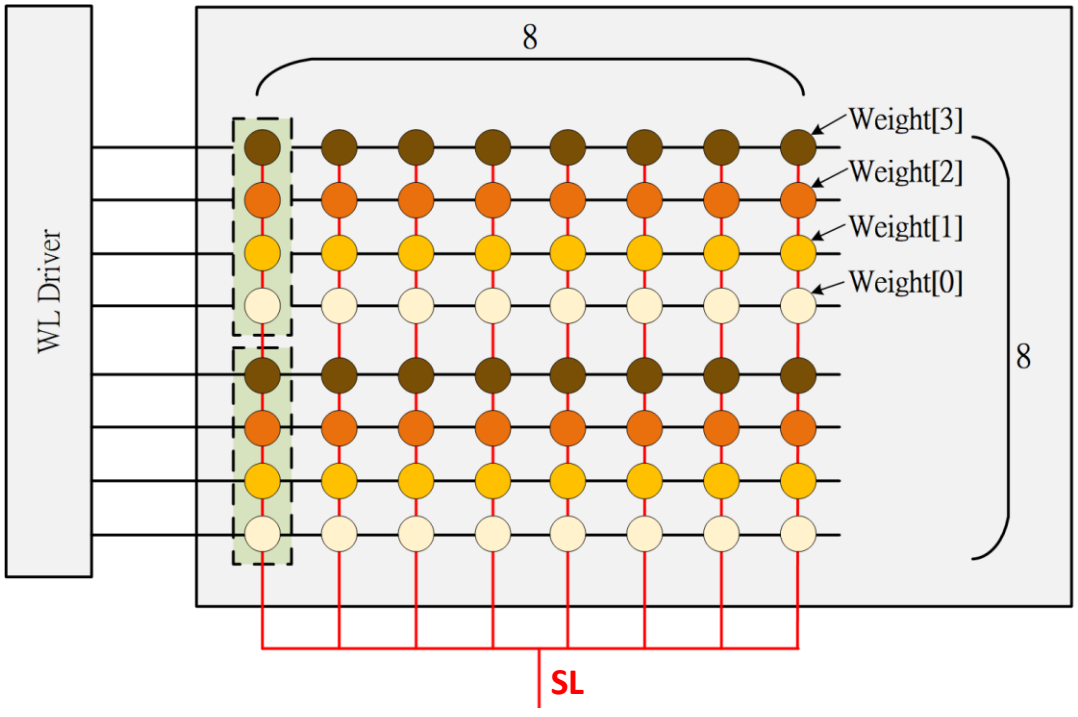

Fig.1

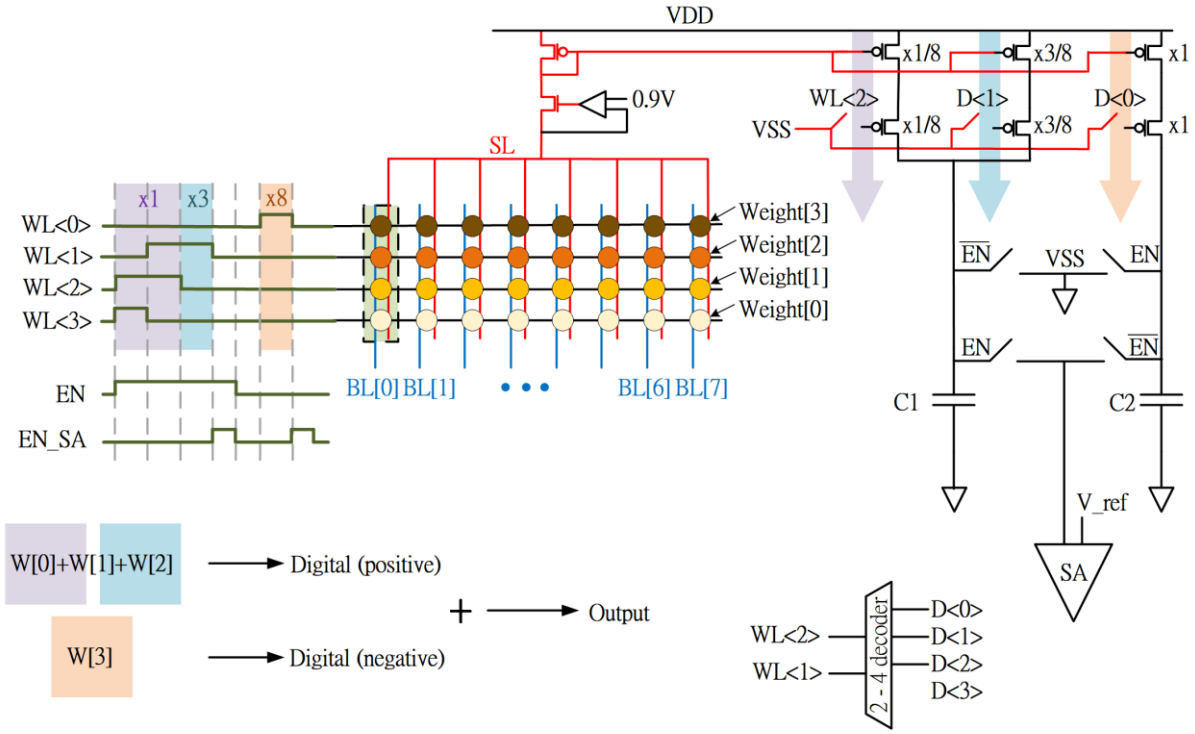
VDD
0.9V
SL
x1/8
x3/8
x1
WL<2>
D<1>
D<0>
VSS
x1/8
x3/8
x1
x1
x3
x8
WL<0>
WL<1>
WL<2>
WL<3>
Weight[3]
Weight[2]
Weight[1]
Weight[0]
EN
VSS
EN
EN
EN
EN
BL[0] BL[1]
BL[6] BL[7]
C1
C2
EN
EN_SA
V_ref
SA
W[0]+W[1]+W[2]
Digital (positive)
Output
W[3]
Digital (negative)
WL<2>
WL<1>
2 - 4 decoder
D<0>
D<1>
D<2>
D<3>

Fig.2

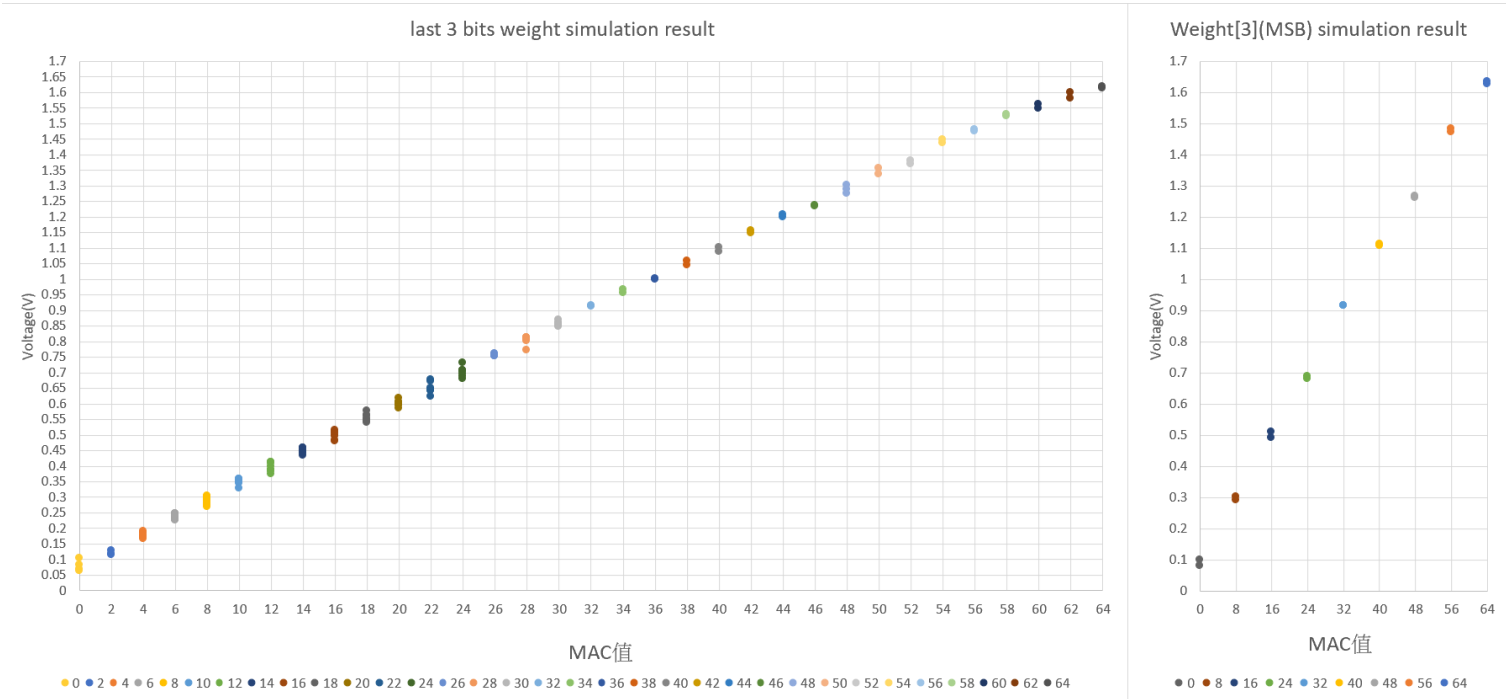
last 3 bits weight simulation result
Voltage(V)
MAC值
Weight[3](MSB) simulation result
Voltage(V)
MAC值

Fig.3

Input	2	0	0	3	2	2	3	1
Weight	-7	-5	-5	3	5	-2	-4	1

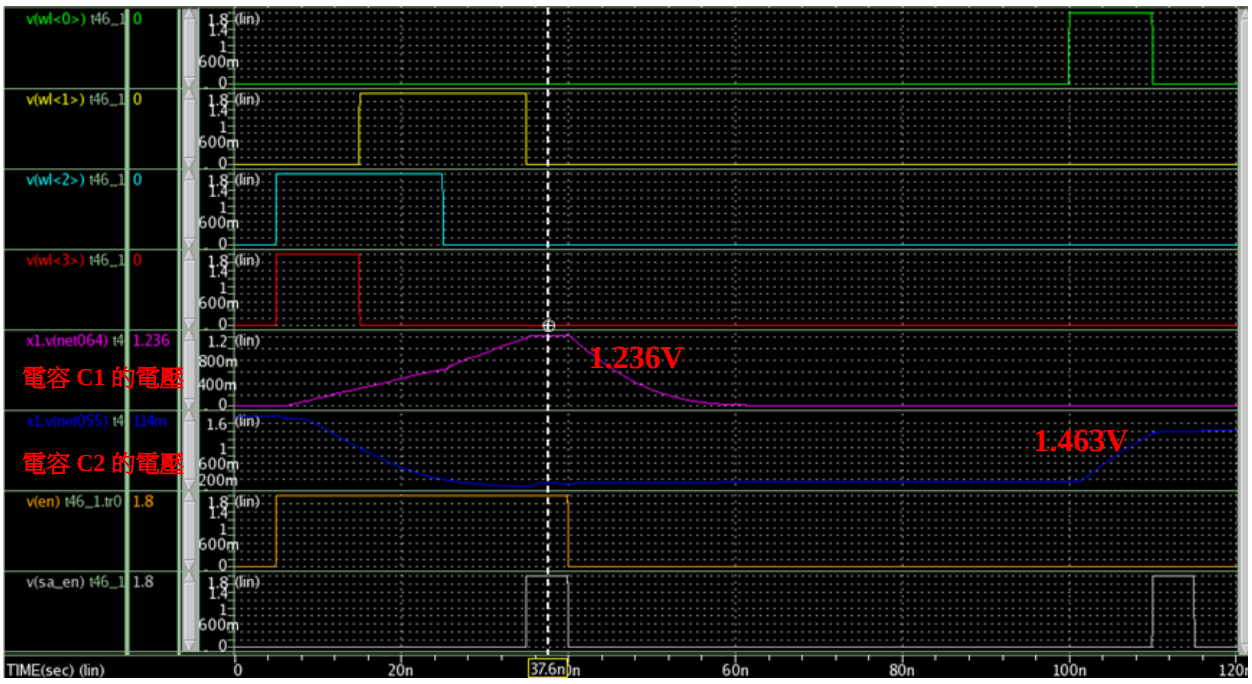

Fig.4 (MAC = -10)

心得感想

劉家蓁

最初對記憶體領域的認識基本為零到現在能完成這個專題，真的很感謝這一年來教授以及學長姐的指導與幫助。從 Tool 的使用、文獻閱讀與報告，到最後完成專題，我們在這過程中學到了許多，不僅是專業知識，還有合作的經驗。特別的是這是在大學部課程中難以接觸到的領域，能透過這機會來了解是相當珍貴的經驗。雖然模擬電路的過程中時常會遇到問題，但 mentor 都能適時的指引我們方向，我們才得以順利完成專題實作。

黃詩瑜

經過 2 個學期的實作專題，閱讀許多 SRAM 和 ReRAM 等記憶體方面的相關論文，透過每周的 meeting 與同學、學長姐們和教授分享自己吸收到的知識，教授會帶著我們思考論文當中哪些地方可以再改進或創新，要求我們擁有獨立思考的習慣，從剛開始設計 decoder，畫 layout 跑模擬，到了這次專題競賽的大架構，了解到許多問題要實際模擬才會發現，思考可能的原因並驗證，再透過不斷的改進將結果趨近理想狀態。非常感謝教授和學長姐，在遇到問題時能不厭其煩地給予幫助，讓我們完成這次的專題競賽。