

MLC ReRAM-Based Computing-in-Memory

Macro for MAC Operations

利用多層單元之電阻式隨機存取記憶體建立

適用乘積累加運算之記憶體內運算架構

組別：A111

指導教授：張孟凡

組員姓名：游得奇

Abstract

近年來人工智慧(AI)處理器崛起，其多採用基於乘積累加(MAC)運算的卷積神經網路(CNN)，因而發展出高效能、低功耗及小面積的記憶體內運算(CIM)架構。非揮發性 CIM (nvCIM)為 CIM 架構的分支，大多採用新興記憶體技術。

電阻式隨機存取記憶體(ReRAM)為新興記憶體成員之一，其基本結構為一顆電晶體加上一顆憶阻器(1T1R)組成的記憶體單元(MC)，並藉由偏壓改變其阻態(RS)。單層單元(SLC) ReRAM 支援兩種阻態，分別為高阻態(HRS、'0')與低阻態(LRS、'1')，HRS 阻值與 LRS 阻值的比值稱為阻比(R-ratio)。

採用 ReRAM 作為 nvCIM 架構基礎的好處主要有兩點，其一是具 CMOS 製程相容性進而節省下線成本，其二則是元件大 R-ratio 允許高精度運算的可能。然而先前提出的 1T1R SLC 架構，存在以下挑戰：

- (1) 再大的 R-ratio 亦非無窮大，因此 1T1R 結構仍存在非理想因素；
- (2) SLC 僅能儲存一位元(1b)資料，資料密度低落。

本作針對(1)提出 2T2R 結構優化，並針對(2)提出 MLC 策略改良，接著額外設計適合 2T2R MLC 架構的讀取電路。相較於 1T1R SLC 架構的 1b-IN-1b-W-1b-OUT，2T2R MLC 架構能夠支援 1b-IN-3b-W-5b-OUT，並使用 180nm 製程檔於 TT 角 25°C 環境下進行模擬。

Implementation

一、憶阻器模組

憶阻器(圖1左)元件具有正端(PLS)與負端(MNS)，其單層式行為如下：

- (1) 若正端電壓(V_{PLS})與負端電壓(V_{MNS})之差大於設定電壓(V_{SET})，則阻態變為 LRS (1)；
- (2) 若負端電壓(V_{MNS})與正端電壓(V_{PLS})之差大於重設電壓(V_{RST})，則阻態變為 HRS (0)。

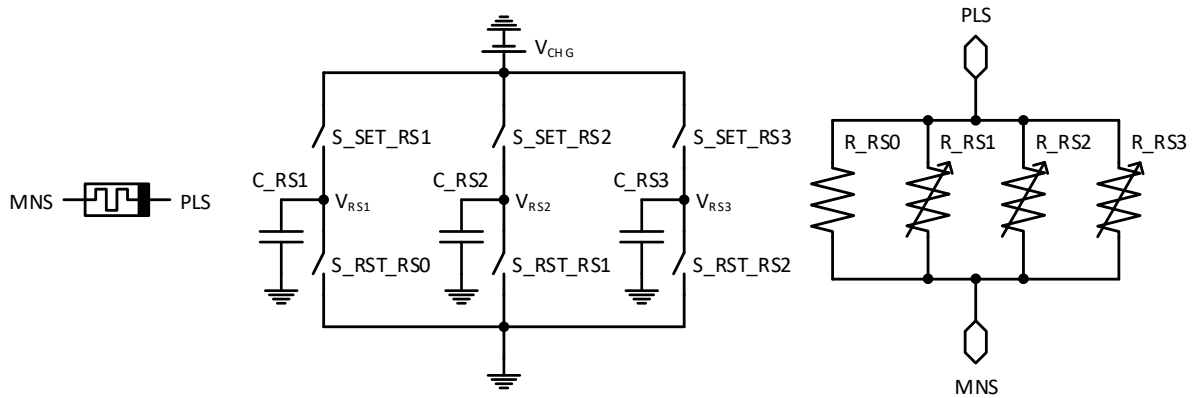


圖1 憶阻器電路符號與模組設計圖

相較於單層式憶阻器，本作設計之多層式憶阻器支援四種阻態，阻值由高至低分別為 RS0至 RS3。模組設計(圖1右)分為左記憶區塊與右電阻區塊，共使用了十四個元件，包含三組開關(S)、三個電容(C)、一電壓源(V)與四顆電阻(R)，並由若干參數控制。

二、2T2R 結構與 M 子運算

2T2R 結構(圖2)是由兩個 1T1R 結構組成，並由兩條字元線(WLP 與 WLN)、兩條位元線(BLP 與 BLN)以及來源線(SL)五條控制線隨機存取。

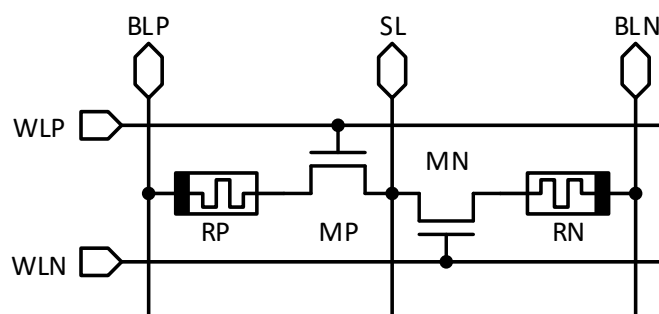


圖2 2T2R 結構

結構的精隨，在於 M 子運算電流互相抵銷的機制。若將 SL 電壓箝(固定)於 V_{CLP} 、BLP 電壓箝於 $(V_{CLP} - V_{READ})$ 、BLN 電壓箝於 $(V_{CLP} + V_{READ})$ ，則 RP (W_P)與 RN (W_N)跨壓皆為 V_{READ} ，利用克希荷夫電流定律(KCL)，便能在 SL 上將 I_{HRS} 消去。而 $W (=W_P - W_N)$ 須保持一般性(表1)，因此規定 W_P 與 W_N 至少一者為0，且 WLP 與 WLN 輸入(IN)同步。

表1 M 子運算真值表

IN	W_P	W_N	W	IWP	I_{MC}
0	3	0	+3	0	0
	2	0	+2	0	0
	1	0	+1	0	0
	0	0	0	0	0
	0	1	-1	0	0
	0	2	-2	0	0
	0	3	-3	0	0
1	3	0	+3	3	3I
	2	0	+2	2	2I
	1	0	+1	1	I
	0	0	0	0	0
	0	1	-1	-1	-I
	0	2	-2	-2	-2I
	0	3	-3	-3	-3I

三、8X8陣列與 AC 子運算

以64個2T2R 結構組織而成的8X8陣列(圖3)，支援兩種操作模式，分別為讀寫操作之記憶體模式與 MAC 運算之 CIM 模式。在記憶體模式中，陣列本質上視為1T1R 結構，WLP 相當於偶數列(0, 2, 4, 6)，WLN 相當於奇數列(1, 3, 5, 7)，BLP 則與 BLN 同步為 BL。在 CIM 模式中，SL 負責 AC 子運算(I_{sum})，因此共有八埠 MAC 輸出。

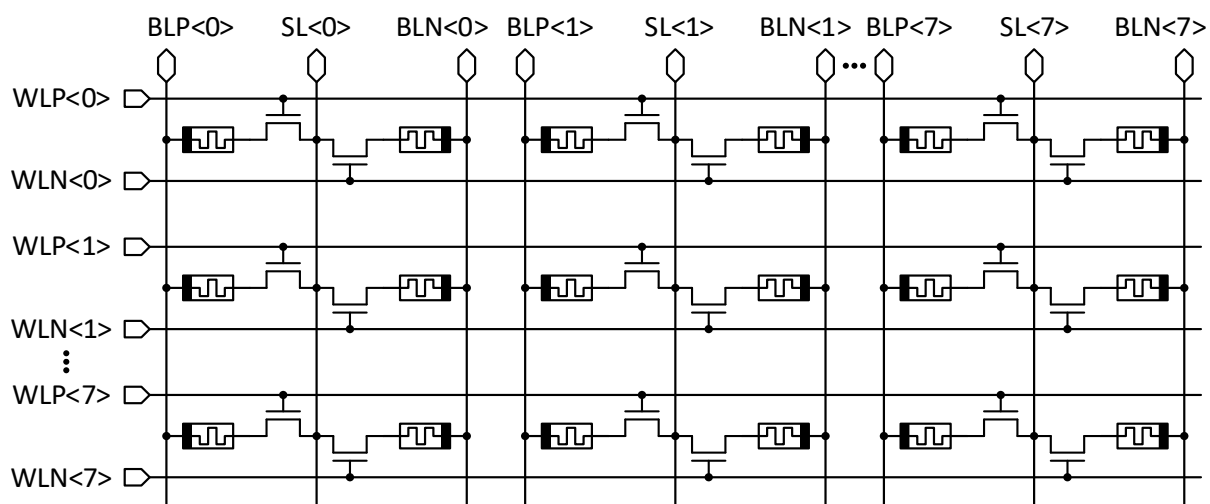


圖3 8X8陣列

四、ADC 電路與讀取功能

本作設計之 ADC 電路由積分器(INTEG)、數位類比轉換器(DAC)以及比較器(COMP)三個子電路組成，操作具有三相，第一相(PH1)為電路重設相，第二相(PH2)為輸入積分相，第三相(PH3)為比較輸出相。輸出精度設計本為八位元，實測卻為五位元，導因於 COMP 之補償電壓(V_{Offset})。

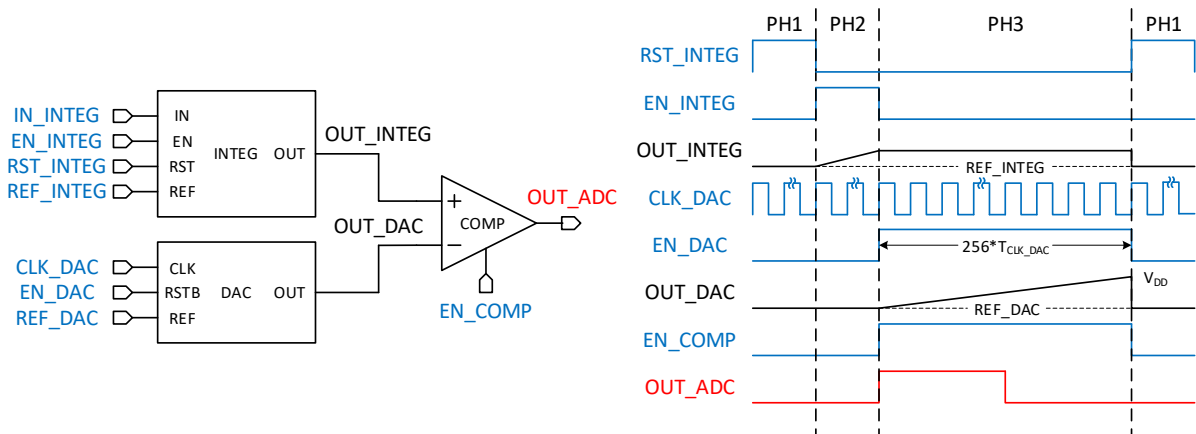


圖4 ADC 方塊圖與波形圖

IN_INTEG 為輸入端，輸入電流(I_{IN})經由 INTEG 積分成電壓後，DAC 開始輸出斜坡電壓。兩者的電壓藉由 COMP 比較，當 INTEG 輸出電壓大於 DAC 輸出電壓時，COMP 便輸出高電平，因此 I_{IN} 的大小決定 COMP 輸出時間長短，此大致為 ADC 的運作原理。在 MAC 運算中，ADC 輸出時間長短便反映 MAC 值大小。

五、最終架構與 MAC 運算結果

將8X8陣列各條 SL 接上一組 ADC，再加上一顆接地的大尺寸 NMOS (M0至 M7)，並以 EN_WRITE 控制線輔助寫操作，便完成最終架構(圖5上)。本作將 SL<0>至 SL<7>上之總權重分別配置為+3、+6、...、+24，藉此觀察倍數 MAC 值間的輸出關係。

模擬結果(圖5下)大致上可區分各種 MAC 值，然而觀察當 MAC 值愈大，輸出比例愈失真；此現象乃因 INTEG 子電路隨著 I_{SUM} 增大，參考端電壓箝位效果降低， I_{SUM} 積分的结果進而脫離線性。

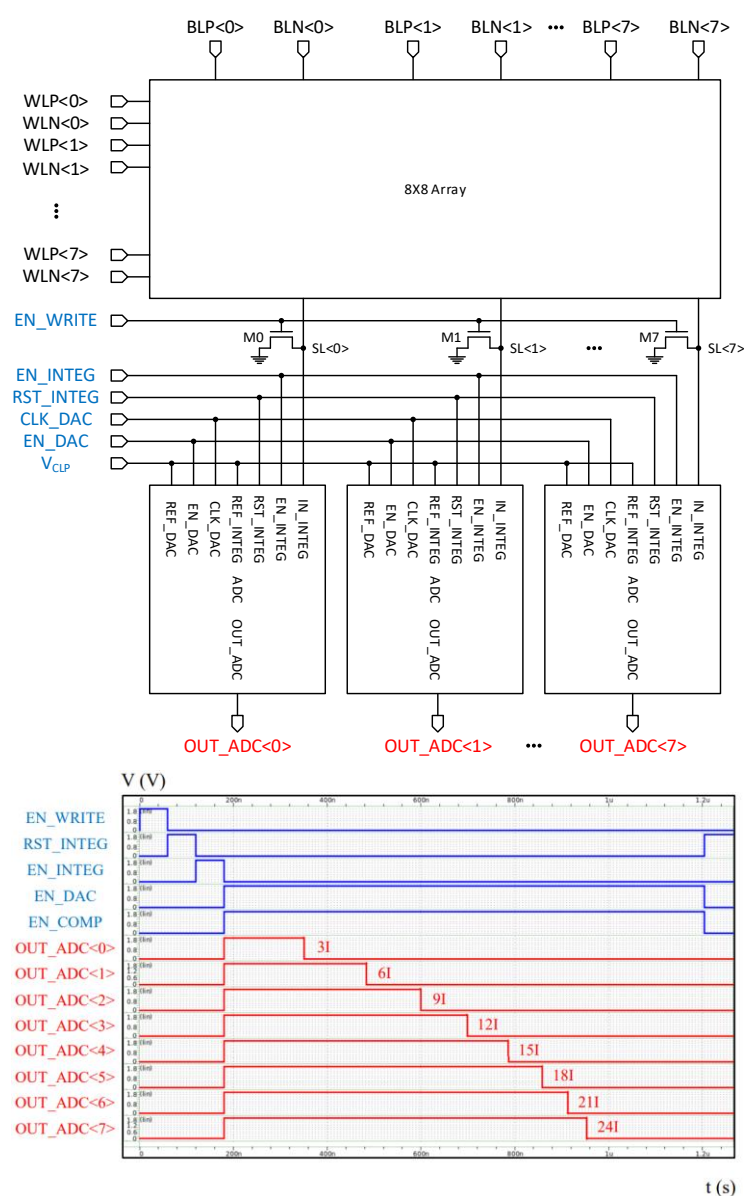


圖5 最終架構方塊圖與 MAC 運算模擬波形圖

Conclusion

本作實作出具有記憶體與 CIM 雙模式的 2T2R MLC ReRAM 架構，能夠順利讀寫並支援 1b-IN-3b-W-5b-OUT 之 MAC 運算。然而讀取電路(ADC)面臨兩大挑戰：

- (1) 積分器(INTEG)子電路：SL 電壓本應箝於 V_{CLP} ，卻隨電流增大而降低；
- (2) 比較器(COMP)子電路：OUT 輸出結果本應為 8b，卻因 V_{Offset} 而僅達 5b。

以上兩點大幅影響架構 MAC 運算之精準度，因此未來將持續針對此兩點改良。

Reference

- [1] C.-X. Xue et al., “Embedded 1-Mb ReRAM-Based Computing-in-Memory Macro With Multibit Input and Weight for CNN-Based AI Edge Processors,” IEEE J. Solid-State Circuits (JSSC), vol. 55, no. 1, Jan. 2020.
- [2] Q. Liu et al., “A Fully Integrated Analog ReRAM Based 78.4TOPS/W Compute-In-Memory Chip with Fully Parallel MAC Computing,” IEEE Int. Solid-State Circuits Conf. (ISSCC), session 33, non-volatile devices for future architecture, Feb. 2020.

心得感想

在我的概念中，「電」是稍縱即逝的物理現象，也因此我一直覺得記憶體很酷，對於專題題目也樂在其中。在實作過程，我將系統設計與驗證分為五個階段，然而每個階段都碰到許多大小不等的問題，不過只要耐心地檢查每處細節，各種錯誤都能迎刃而解，也因此累積不少成就感。在這將近一年的專題研究期間，從學習電路設計工具、閱讀 Paper 到架構實作方面的訓練都相當紮實，因此特別感謝實驗室學長姐與張孟凡老師的指導。專題現階段的架構並非完美，尚有改進之處，因此將來須持續構思改良方案。