基於 ADFP 改良比較器之無乘法近似 SRAM 記憶體內運算巨集

Based On ADFP Technology Multiply-Less Approximation SRAM In-Memory Computing (IMC) Macro with Improved Comparator

指導教授: 王進賢 特聘教授　專題生: 李紹瑋、王睿超、林品佑　組別:晶片系統組

研究摘要

傳統 Von Neumann architecture 中，處理器與記憶體分離的設計方式，導致資料在兩者之間傳輸時形成瓶頸。尤其在深度學習與神經網路模型中，龐大的權重與中間資料需要頻繁搬移，使得資料傳輸所耗費的時間與能量遠超過實際的計算本身。為了解決此一瓶頸，「**記憶體內運算**（In-Memory Computing, IMC）」的概念被提出。其核心思想是將部分運算功能整合進記憶體單元內部，使資料能夠在記憶體內被處理，從而減少資料搬移、提升效能並降低功耗。本次專題在操作電壓為 0.72 V 的情況下達到 11.9 TOPS/W 和 0.11 TOPS/mm^2 的功耗表現。

本專題參考今年發表於 JSSC 的論文[1]所提出的近似 DIMC 電路設計，使用 8T SRAM 、comparator 和其他周邊電路組成記憶體內運算電路。

AdderNet 介紹

- 使用L1 - distance compute scheme 近似法得到與卷積方法相似的輸出，以節省傳統 IMC 加法樹和乘法器產生的大量功耗和面積。
- $Y = \sum|x_i - w_i| = \sum x_i + \sum w_i - 2\sum min(x_i, w_i)$，讓計算可以使用 bit-serial 的配置。

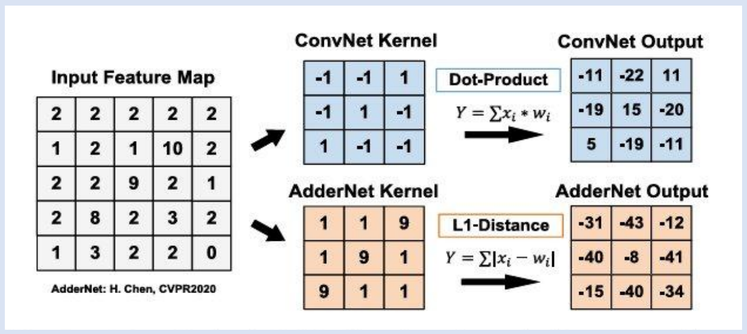

圖1、AdderNet Kernel 達到與乘法相似的輸出

研究局部設計

SRAM Array

- SRAM array 由 8*128個 8T SRAM 組成，分開讀取路徑(RBL、RWL) 、寫入路徑(BL、WL) ，解決 6T SRAM 讀取干擾問題。
- 每個 cycle Control 將 128*1b 的 weight 透過 Write driver 寫入 SRAM ，共8個 cycle。

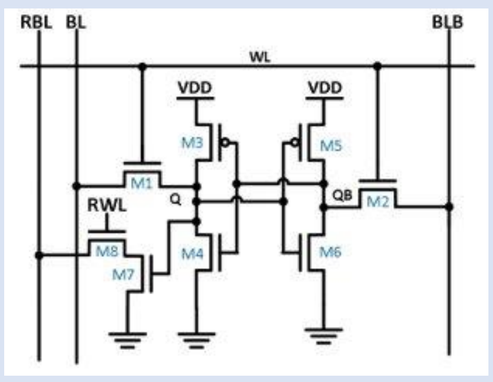

圖2、8T SRAM 之電路

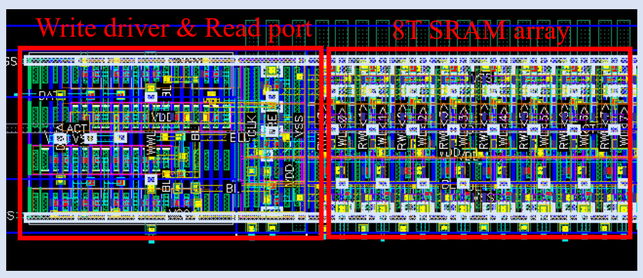

圖3、SRAM array 之 layout

Read port

- Early stopping

透過 OUTB bar 控制減少預充電次數，以節省功耗 (當 W>ACT，把 OUTB bar 充電至 VDD)

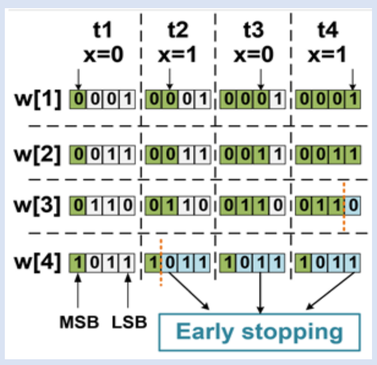

圖4、Early weight-readout stopping mechanism

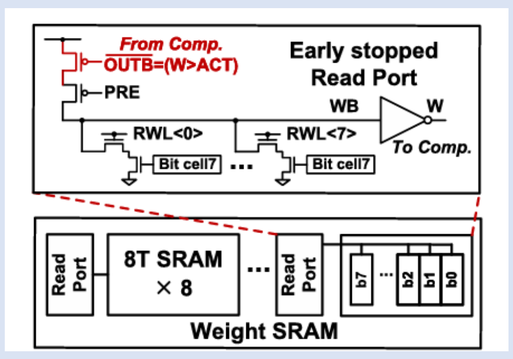

圖5、Schematic of SRAM read port

Adder tree

- 使用樹狀結構並行加法，將 ACT 和比較的結果加總成一個 8-bit 結果值

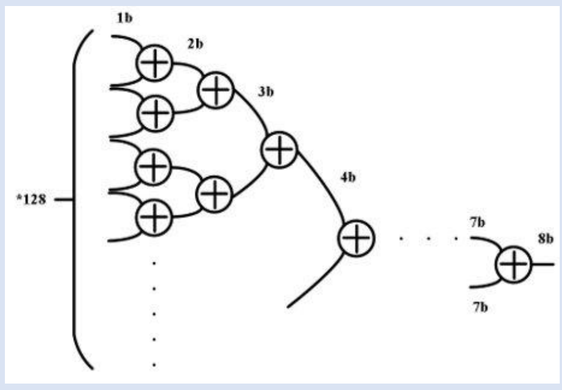

圖6、128 inputs Adder tree 架構圖

Accumulator

- 累加結果：$MSB * 2^M + (MSB - 1) * 2^{M-1} + \ldots + (MSB - 2) * 2^1 + \text{LSB} * 2^0$
- 第8個 cycle 加完後，clko 會變成 1，將結果送到 ABS-ASM。

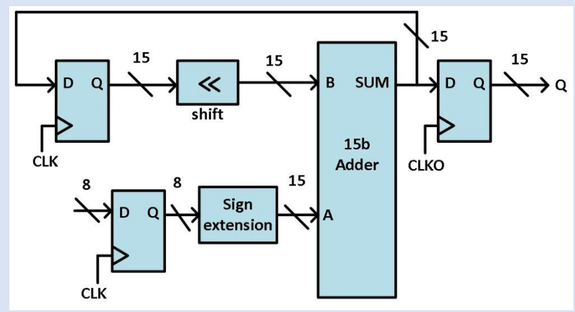

圖7、Accumulator 架構圖

Comparator

原論文架構

C1、C2 透過 WB 放電至 VSS，使 OUT 無法維持高電壓，在ACT和W多"1"值比較容易因 charge-sharing 造成 voltage drop 導致錯誤發生，此架構最多容忍輸入有 5 個 1。這在低位元時影響不大，但對於高位元的準確率大幅降低。

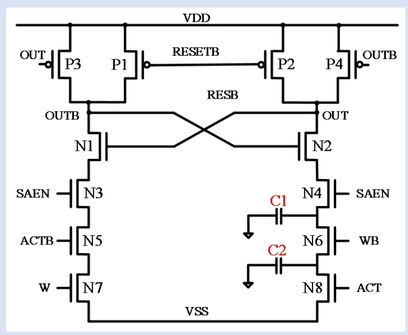

圖8、original comparator

圖9、發生錯誤波形圖

改良比較器架構

- 加入一個 2 to 1的多工器控制 WB 的時序。
- 讓比較時(SAEN=1)才會傳入正確的 WB，其餘時候N6皆不會通，斷開放電的路徑。
- N4、N6間的寄生電容保持充飽電的狀態，避免OUT 發生 voltage drop。

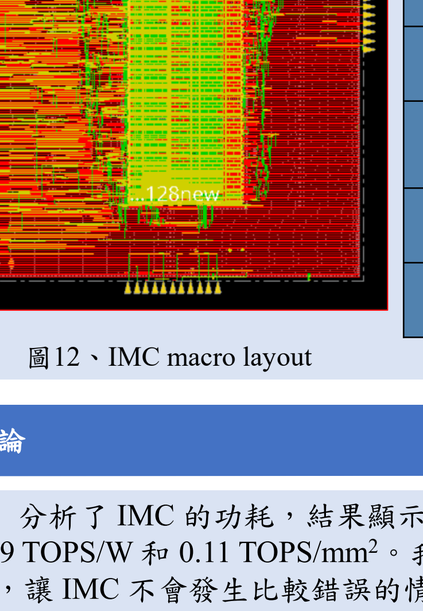

圖10、improved comparator

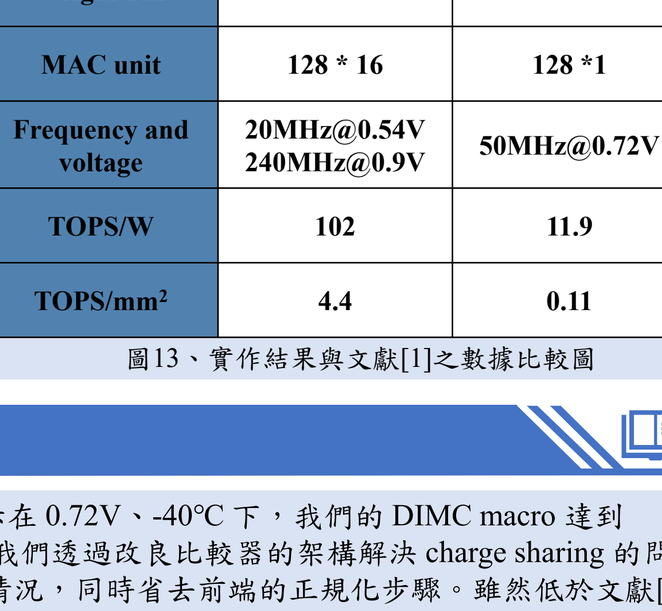

圖11、改善 charge sharing 後波形圖

實作結果

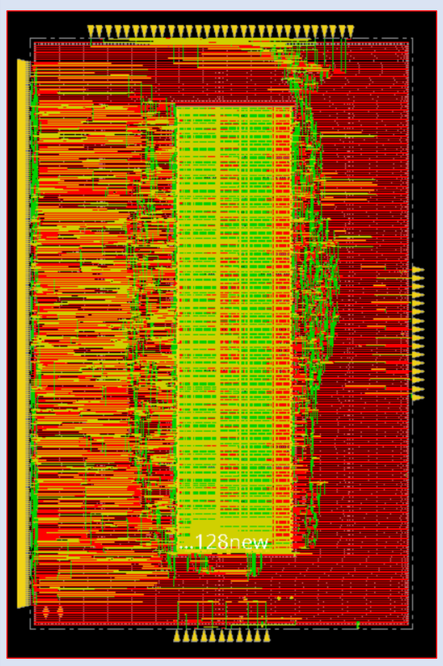

圖12、IMC macro layout

規格與功能	文獻[1](16 macro)	實際實現(1 macro)
Technology	28nm	16nm
NN operand	ABS + ADD	
Input activation	2~8 bits	8 bits
Weight bits	2~8 bits	8 bits
MAC unit	128 * 16	128 *1
Frequency and voltage	20MHz@0.54V 240MHz@0.9V	50MHz@0.72V
TOPS/W	102	11.9
TOPS/mm^2	4.4	0.11

圖13、實作結果與文獻[1]之數據比較圖

結論

分析了 IMC 的功耗，結果顯示在 0.72V、-40℃ 下，我們的 DIMC macro 達到 11.9 TOPS/W 和 0.11 TOPS/mm^2。我們透過改良比較器的架構解決 charge sharing 的問題，讓 IMC 不會發生比較錯誤的情況，同時省去前端的正規化步驟。雖然低於文獻[1]中提及的 102 TOPS/W，但我們找出了佈局設計、 Adder tree 的設計方法和 decoder 的資料傳輸方式等原因。未來規劃採用 programmable 的設計，並且優化 IMC 電路 floorplan 的面積、layout 的手法，以提升 IMC 的多項指標。最終實現完整16個 macro、加入 SRAM IP 完成整個 Data flow 之後，我們會模擬實際神經網路推論流程，盼望功耗表現、準確率改善幅度能夠超越現有的研究成果。

參考文獻

[1] H. Diao et al., "A Multiply-Less Approximate SRAM Compute-In-Memory Macro for Neural-Network Inference," in IEEE Journal of Solid-State Circuits, vol. 60, no. 2, pp. 695-706, Feb. 2025