

摘要

傳統的計算架構（如CPU、GPU）皆為馮紐曼架構，數據資料需要由記憶體傳輸到運算單元做運算，再由運算單元傳輸回記憶體做儲存。而因為深度學習所需處理之數據量過於龐大，巨量資料需於記憶體與運算單元間的反覆傳輸，導致資料的傳輸速度遠比不上運算單元的計算速度，而造成大量的時間浪費與功率消耗。為了解決馮紐曼架構造成之瓶頸，提出了一種新型運算架構——記憶體內運算Computing In Memory(CIM)。

記憶體內運算是一種整合記憶體與運算單元的技術，它將計算與數據存儲做緊密的結合，資料無須再於記憶體和運算單元間反覆傳輸資料，克服馮紐曼架構遇到之瓶頸，實現更高效的運算和解決多餘功率消耗的問題。

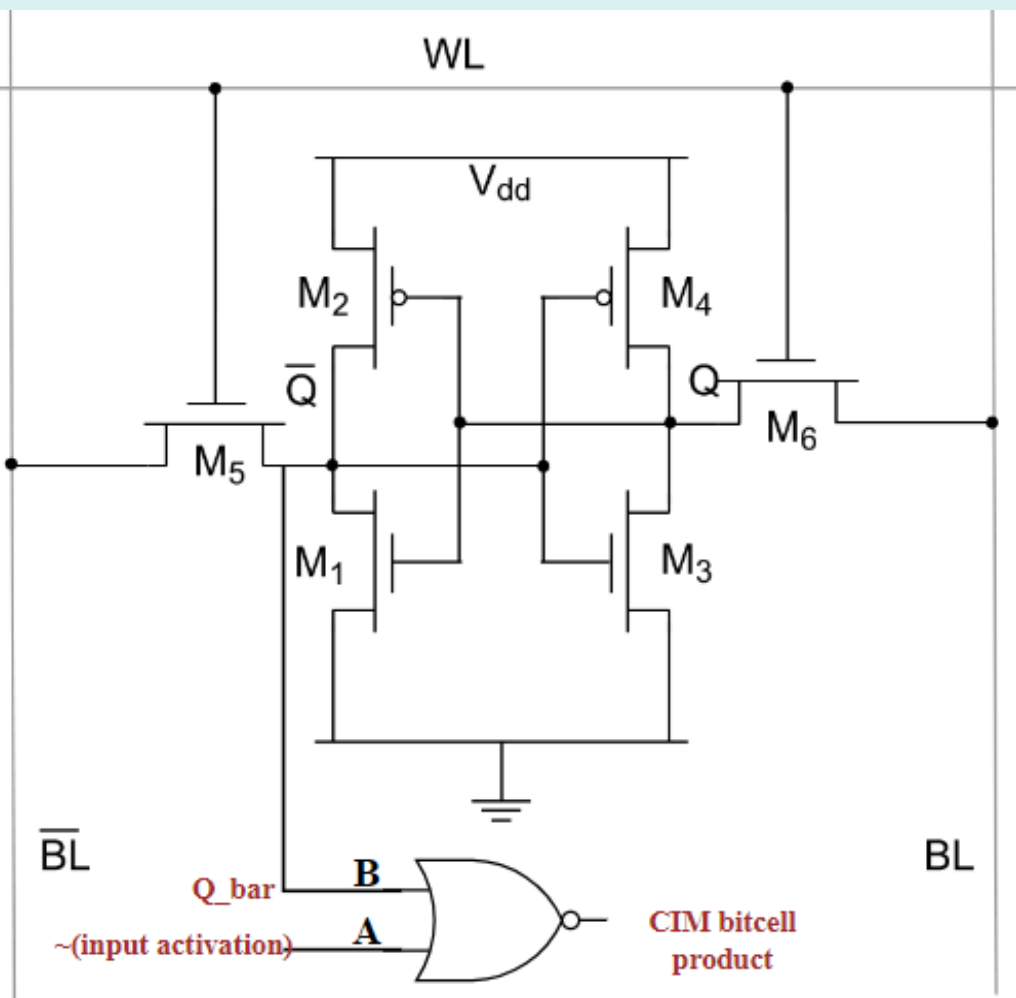
本專題研究參考2021年由TSMC團隊發表於ISSCC期刊的CIM電路設計，主要目標為透過6T SRAM與NOR gate建立一種提供卷積神經網路運算的新型記憶體內運算電路架構，實現記憶體內運算功能，提升電路運算效能與減少功率消耗，於邊緣環境中提供機器學習與深度學習的功能。

研究方法與設計

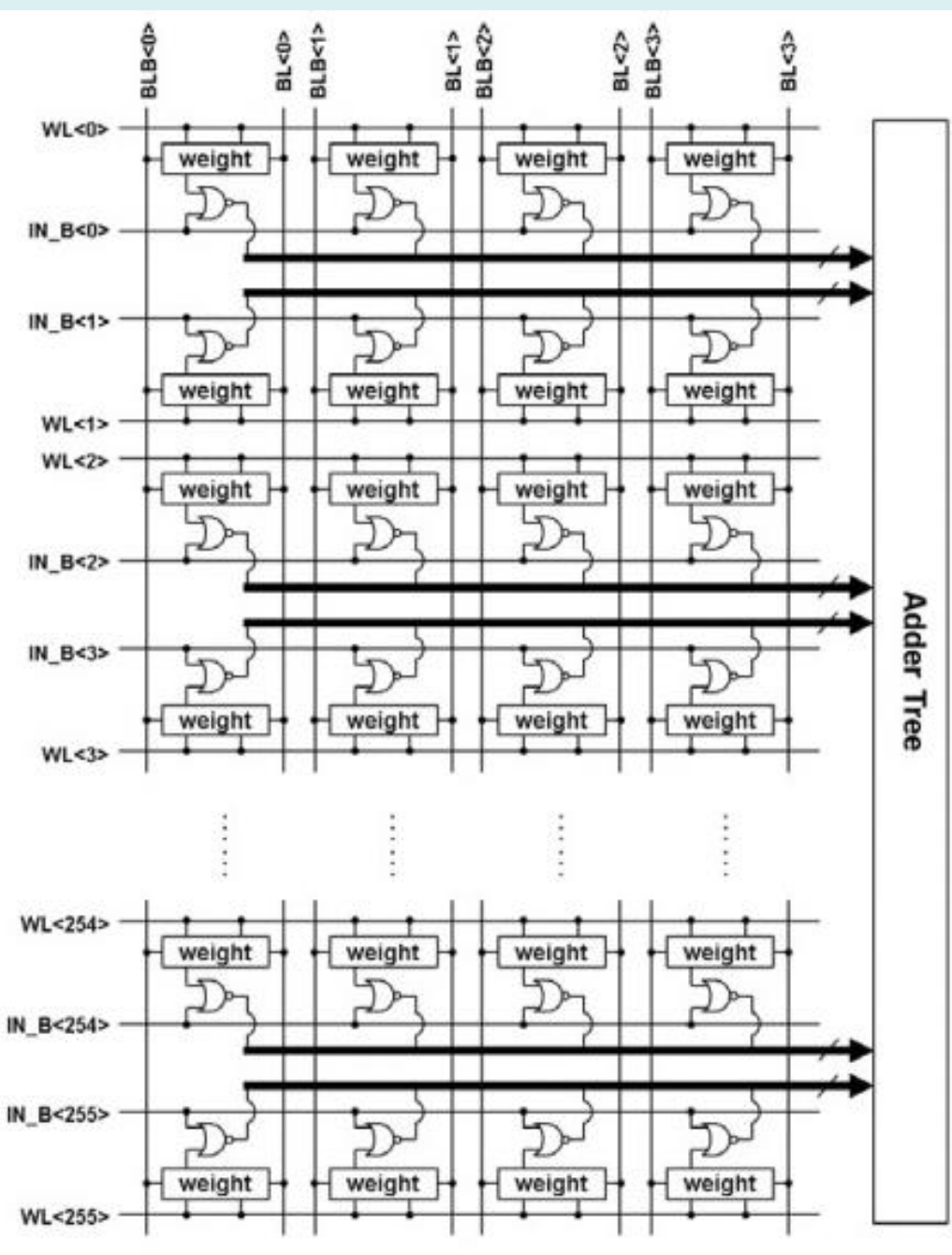
本專題實作的特色為bit-serial multiplication和並行加法樹的新架構。此CIM電路還支援可程式化的有號或無號輸入(1-8 bit)和可程式化的4種不同之有號或無號權重(4,8,12,16 bits)。

● sub-CIM unit :

6T SRAM作為儲存權重的記憶體單元，4T NOR 作為乘積運算單元。sub-CIM unit架構是由每列4個CIM bitcell，共256列組成，每塊sub-CIM unit會運算得出256組4 bits乘積。



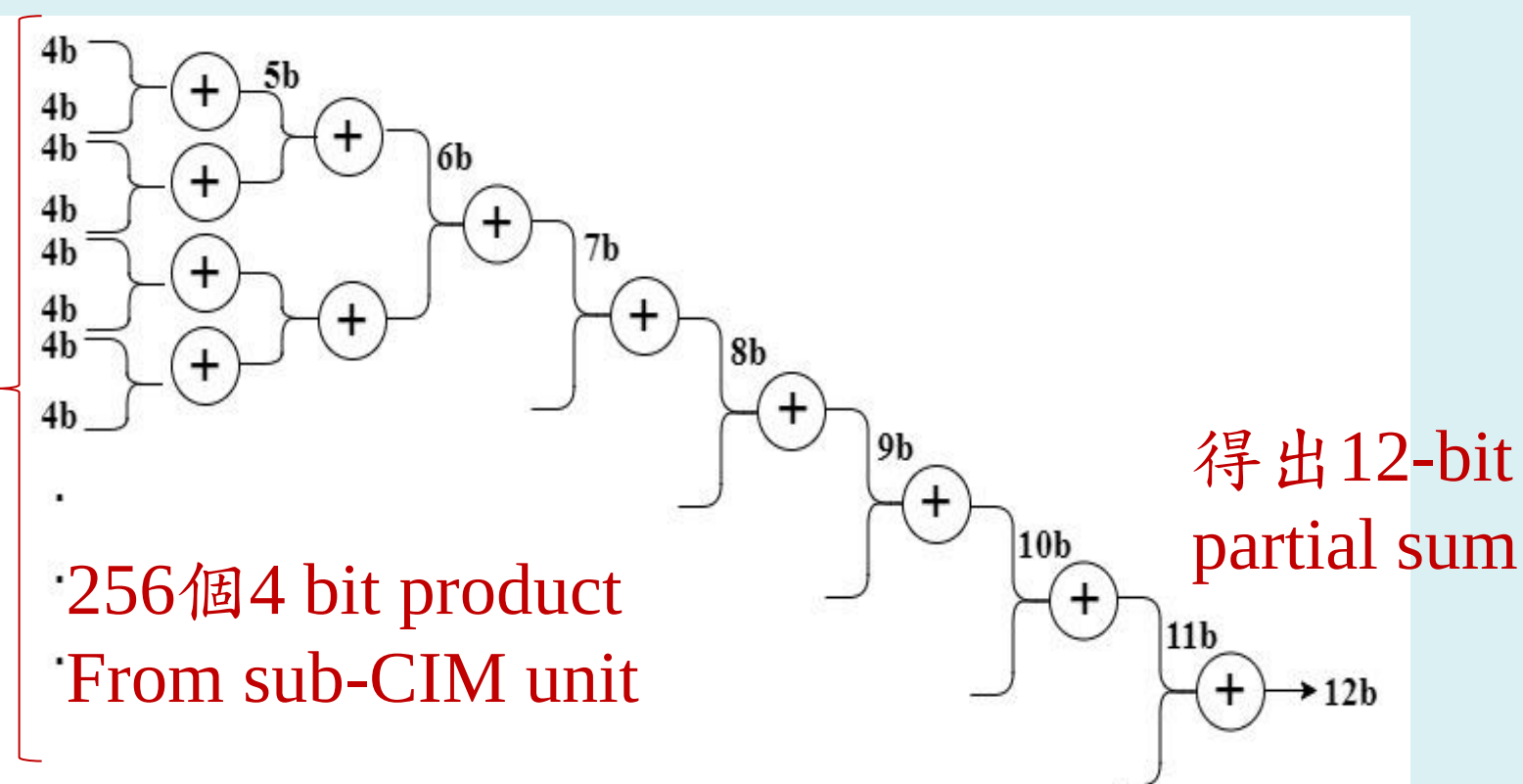
圖一、sub-CIM unit中的CIM bitcell架構



圖二、SRAM-based sub-CIM unit架構

● Adder tree :

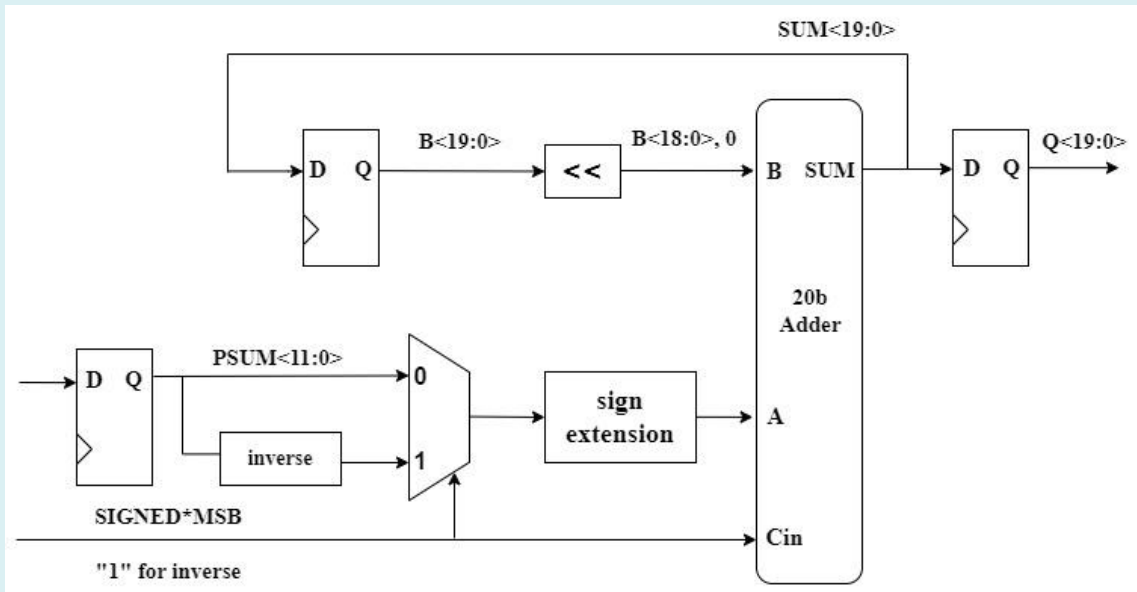
將sub- CIM unit運算輸出的乘積結果做初部累加。



圖三、Adder tree累加運算架構示意圖

● Accumulator :

將Adder tree的partial sum結果做移位相加的累加器。

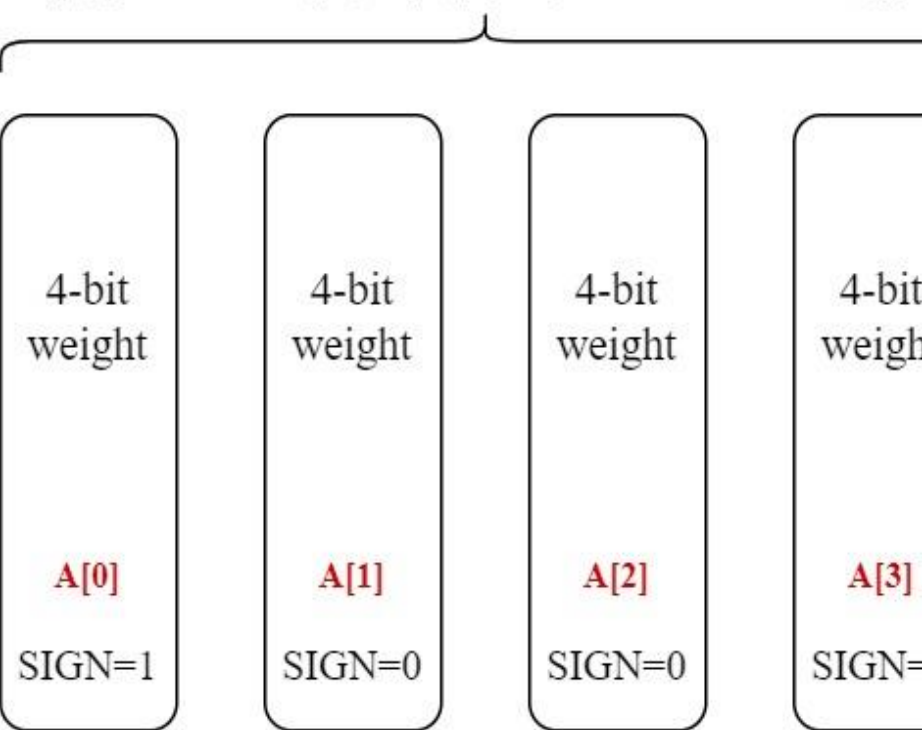


圖四、Accumulator電路架構圖

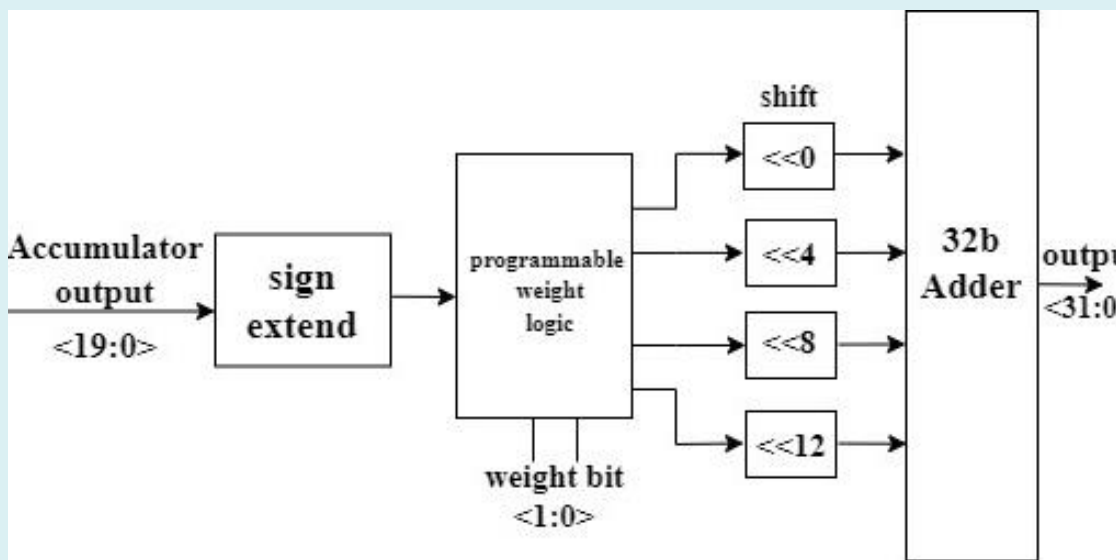
● programmable weight權重整合器 :

以多片sub-CIM unit組合來實現權重可程式化的功能。

有號、16-bit 權重架構(4片sub-CIM unit組成)



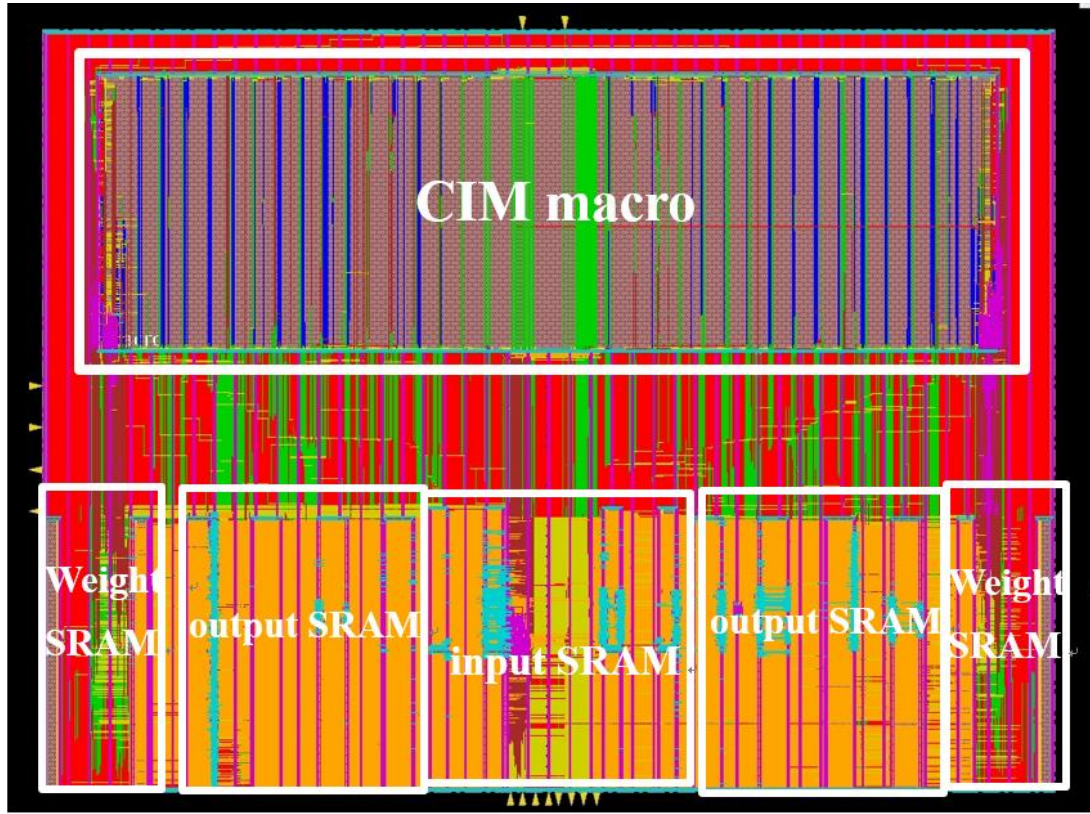
圖五、以多片sub-CIM unit組合來實現programmable weight示意圖



圖六、programmable weight權重整合器電路架構圖

實作結果

規格與功能	參考文獻	實際實現
input activation	1~8bit	1~8bit
weight bits	4, 8, 12, 16	4, 8, 12, 16
signed and unsigned	Yes	Yes
Dynamic Voltage scaling	Yes	No
Concurrent weight update function & MAC operation	Yes	Yes
SRAM test	Unknown	Yes
F_{max}	100MHz@0.8V	50MHz@0.8V
Tops/W	89	18.1(估算)
Tops/mm ²	16.3	1.54(估算)



圖七、實作結果圖

結論

本專題目前已完成22nm下CIM電路的設計和各子電路元件的post-layout simulation功能驗證和功耗測量、整體CIM電路含timing delay資訊的功能驗證。

未來嘗試實現參考文獻中的Adder tree架構(14T與28T FA交錯排列)來減少功耗，以及嘗試實現DVS功能來達成CIM電路閒置時降低供應電壓，優化電路功耗。

參考文獻

[1] Yu-Der Chih, Po-Hao Lee, Hidehiro Fujiwara, ... ,Tsung-Yung Jonathan Chang, "An 89TOPS/W and 16.3TOPS/mm² All-Digital SRAM-Based Full-Precision Compute-In Memory Macro in 22nm for Machine-Learning Edge Applications", ISSCC, p. 252-253, 2021.