電機資訊學院 2020 作 BRAIN PLUS HAND 實作專題競賽

Hardware Implementation of Neural Network and IoT Security in ReRAM-based Memory

電阻式記憶體於神經網路與物聯網安全性的硬體應用

組別: EECS03　組長: 張傳佳　組員: 陳祈瑋、謝鈁淩

I. Abstract

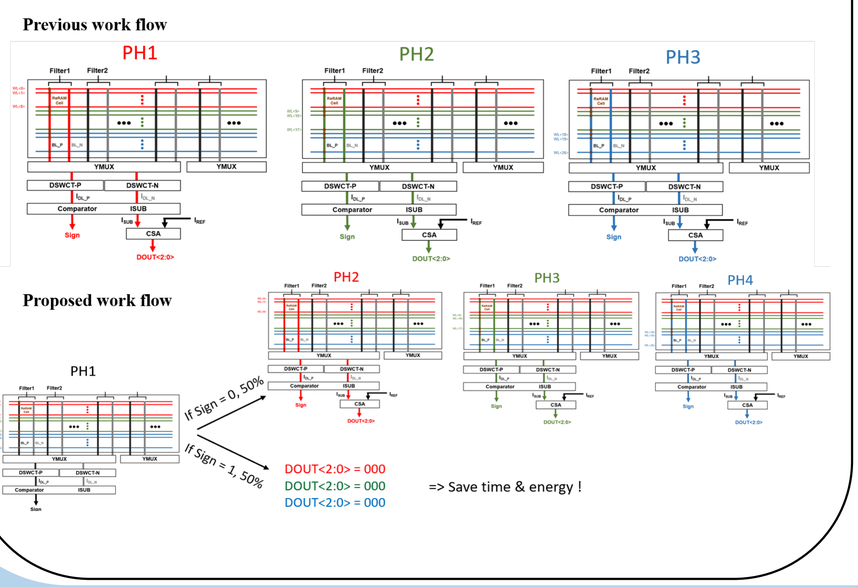
隨著深度學習的應用越來越廣泛，即時、穩定、安全的邊緣裝置運算越來越重要，例如自駕車、居家安全等。記憶體內運算(Computing in Memory)將大量的乘加計算(MAC)於記憶體內完成，解決了馮諾伊曼(Von Neumann)架構中速度慢與高耗能的障礙，將成為邊緣裝置運算的重要突破。此研究提出ReLU in Memory，比較於原架構[1]能將速度與耗能再降低30%以上。然而，在記憶體的資料可能會被駭客駭入並竊取資料。因此，此研究將硬體安防電路(Physical Unclonable Function, PUF)與記憶體內運算(CIM)結合解決此問題。PUF可以被視為硬體電路的指紋，它是利用製作晶圓的過程中每顆電晶體不會完全相同的特性，達到產生無法預測的密碼(key) 。

儘管如此，PUF電路的實作仍有些問題須被克服。此研究注重在兩重點: 一個是降低錯誤碼產生的比例(bit error rate, BER)，另一個是預防駭客會使用暴力破解法破解PUF的防範機制。

II. Implementation

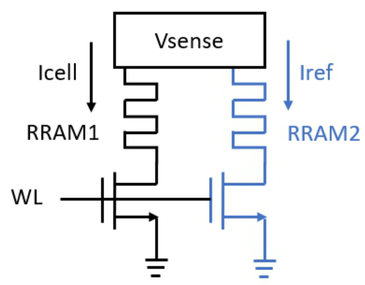
A. ReLU in Memory

此研究分析深度為三的CNN架構，原架構的流程分為三個階段，每階段處理9條WLs並感測出3-bit MACV再將三個值累加得到總值才送到Activation & Pooling function。由於目前ReLU被大量用於Activation function，此研究提出的流程增加第一階段處理27條WLs的正負電流比較。假設處理大量資料，結果為正或負的機率各為50%，若為正則接續原本的流程，若為負則直接輸出0節省了大量的時間與耗能。



B. ReRAM PUF

此研究目的想達成較低的錯誤碼比例(BER)和不穩定碼(unstable bit)的數量。理想情況下，比較兩個不同的cell來判斷PUF的輸出是0或1所以可以使BER趨近於0且得到inter-HD趨近於0.5[2]。另外，PUF也需要精準的sense amplifier來降低BER的產生，此研究使用電流感測放大器(current sense amplifier, CSA)，因為使用電容儲存採樣的電流產生的節點電壓所以也可以不受閾值電壓(VTH)因隨製程上的差異而變動的影響，以達到BER極低的情況。



Icell > Iref : response=1

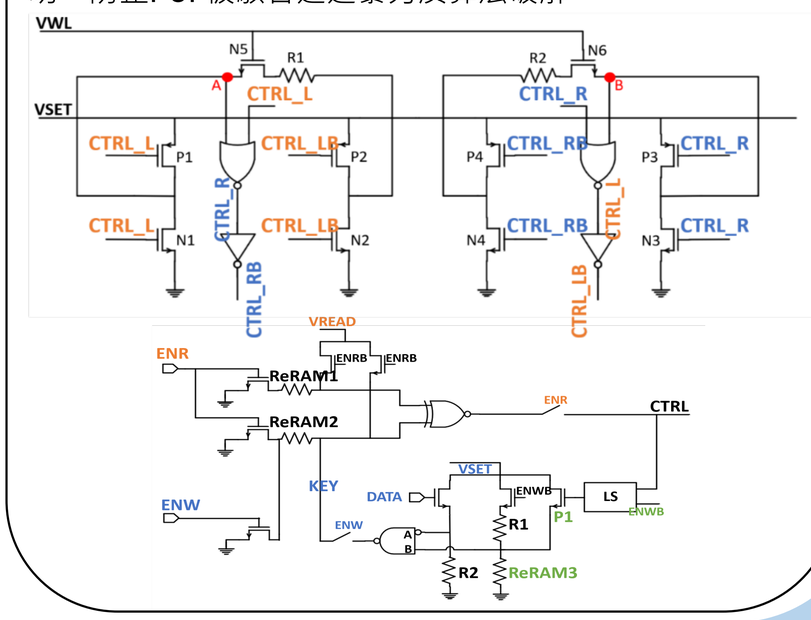
Icell < Iref : response=0

Set pulse : 55ns

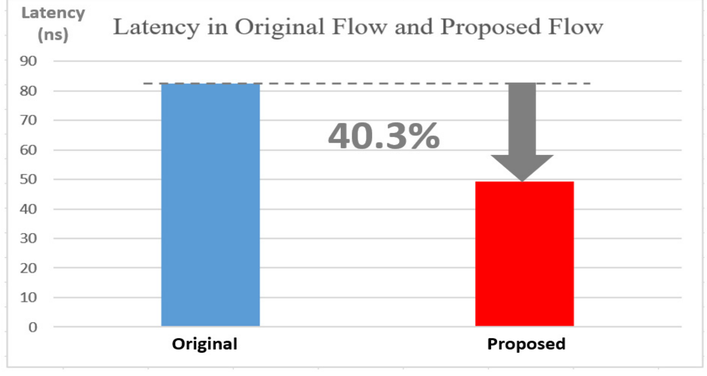
R ratio : 200K : 20K

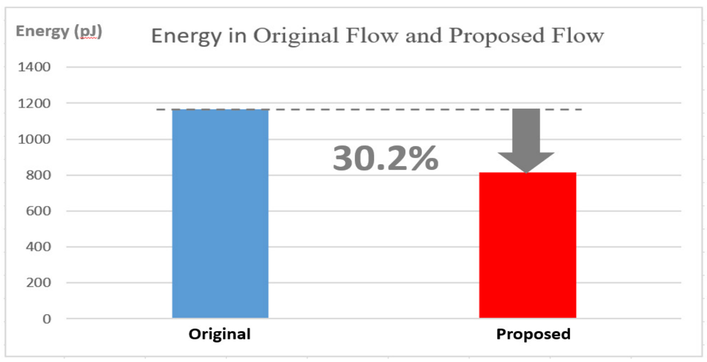
C. PUF Robustness

此研究加上兩項機制來強化ReRAM PUF的功能，第一項的做法是針對不同ReRAM在SET上的時間差異，當其中一顆ReRAM SET完成後，立即使另一顆ReRAM進行RESET，而完成SET的ReRAM則繼續SET，因此兩顆ReRAM的阻值差異會越趨明顯，進而大幅降低BER；而第二項的機制是當輸入的key錯誤次數超過一定程度時，即永遠無法再驗證成功，防止PUF被駭客透過暴力演算法破解。



III. Result & Conclusion





比較於原CIM架構能節省至40%的時間與30%的能源，進而克服邊緣裝置運算瓶頸。

Inter-HD of ReRAM PUF

corner / ℃	TT	FF	SS	SF	FS
25	0.493	0.493	0.487	0.493	0.446
75	0.5	0.514	0.488	0.497	0.504
125	0.475	0.489	0.511	0.516	0.482

BER before applying splitting resistance

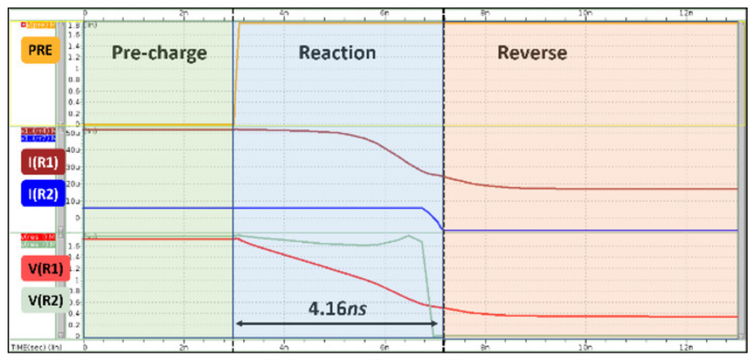
corner / ℃	TT	FF	SS	SF	FS
25	0%	0%	0.097%	21.09%	14.355%
75	4.101%	1.66%	1.172%	7.592%	15.657%
125	4.297%	0%	0.781%	6.355%	6.794%

BER after applying splitting resistance

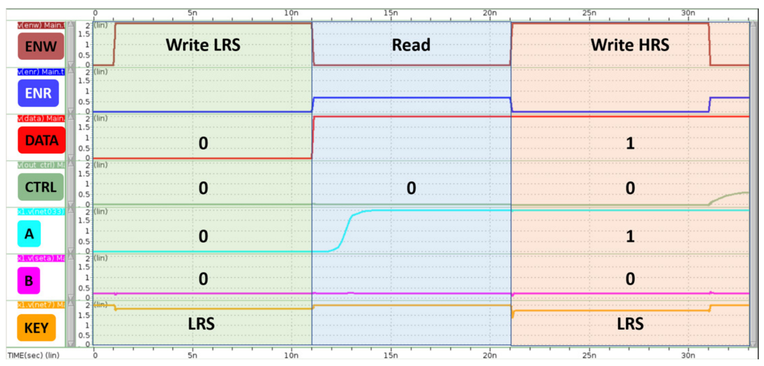
corner / ℃	TT	FF	SS	SF	FS
25	0%	0%	0.049%	11.523%	9.703%
75	4.199%	0.39%	0.244%	6.738%	13.77%
125	2.637%	0%	0%	6.055%	5.762%

經由比較兩個cell電流達到很低的BER藉由新增的阻值反寫功能又更降低且更不受溫度及corner影響。

Splitting Resistance



Prevent Brute Force



透過限制輸入PUF的key錯誤次數可以做到防止暴力破解法的機制，使PUF的按全性更加提升。

[1] Cheng-Xin Xue et al., "A 1Mb Multibit ReRAM Computing-In-Memory Macro with 14.6ns Parallel MAC Computing Time for CNN Based AI Edge Processors", International Solid-State Circuits Conference, 2019
[2] Yachuan Pang et al., "A Reconfigurable RRAM Physically Unclonable Function Utilizing Post-Process Randomness Source with < 6×10-6 Native Bit Error Rate", International Solid-State Circuits Conference, 2019