專 題 期 末 報 告

Neural Network Based on Computation In-Memory with

Ferroelectric Material

指導學生: 吳峻陞

指導教授: 郭峻因 教授、趙天生 教授

目錄

設計動機

在機器學習 (Machine Learning, ML) 蓬勃發展的今天，許多領域在其的幫助下都有了顯著的成長和發展，因此，如何有效率的執行訓練 model 變成了現在學界的主要研究方向。目前，多數的研究重心是放在如何增強 model 本身的表現(performance)以及資料集的單一化(dataset regularization)上面，但是除了上述兩個軟體上的優化之外，硬體方面的研究，如: CUDA、硬體加速器(Hardware Accelerator)、記憶體內運算(Computation In-Memory, CIM)，也是近期相當有發展潛力的研究方向。

在上述的硬體優化中，CIM 是實質神經元(physical neurons)可以在記憶體內成功實現的架構，若邏輯閘運算能夠在記憶體中實現並載入訓練時所需要的權重(weight)，對於 model 在之後不論預測速度或是效能最佳化上(power consumption and optimization)都能有比現在更好的發揮。

為了解決權重儲存和 CIM 的實現問題，非揮發性(non-volatile)且運算耐受度(endurance)高的材料是迫切需要的，而鐵電材料(ferroelectric material)的電特性恰恰符合我們所要求的，考量到擴充性(extensibility)和量測的難度，我選擇了 FTJ(ferroelectric tunneling junction)材料和 FeFET(ferroelectric FET)作為架構的基底來發想電路設計，以成功消弭 CPU 和記憶體之間相當費時的存取(access)時間來優化 model 的效能及運算。

材料特性 & 應用

- Ferroelectric Tunneling Junction (FTJ)

FTJ 在不同的退火(annealing)的狀態下(圖 1.)會有不同的極化窗(polarization window)產生，當退火溫度為 700°C 時，window 的寬度足夠我們在狀態(0/1)判別時有夠大的電阻開關比(resistance on/off ratio)，在開時(state = 1)，FTJ 的電阻值會在大約 100M 歐姆，為低電阻態(low resistance state, LRS)，而在關時(state = 0)，電阻值會落在 400M 歐姆，為高電阻態(high resistance state, HRS)，當我們設定一個感測器的觸發條件在 LRS 和 HRS 之間，就可以區分 FTJ 目前的狀態並使用其為邏輯單元。

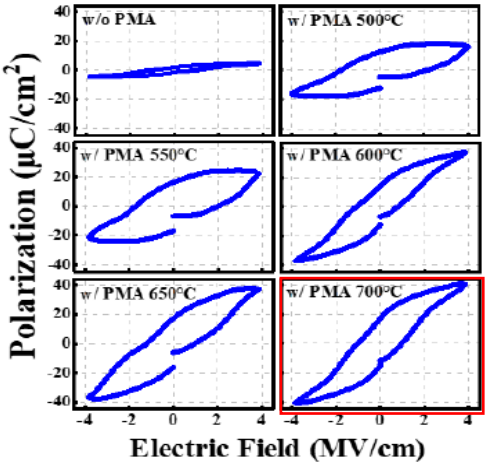

圖 1.

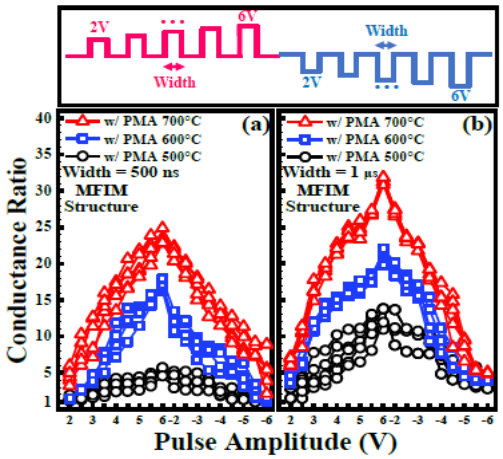

圖 2.

FTJ 還擁有隨電壓大小而改變晶體排列的特性，前述的 LRS 實為晶狀有序排列結構(crystalline structure)，而 HRS 為亂度無序排列結構(amorphous structure)，因此利用施加不同方向的電壓，我們可以操控 FTJ 現在的導電度(conductance)並得知其狀態，這樣一來，我們就可以讀&寫(read & write)由 FTJ 所構成的邏輯元件，並利用施加電壓的大小來控制得到我們想要的導電度(圖 2.)以符合整體電路設計。

了解其電特性之後，endurance 也是目標符合 CIM 設計重要的一環，利用 2~6V 的 pulse 重複寫(圖 3.)0.5~1.5V 的 pulse 重複讀(圖 4.)下，FTJ 的 on/off ratio 仍舊有顯著的差別。

此外，在加上適當的緩衝層(buffer layer, 這裡使用絕緣薄膜)後，電阻的變化會呈現如階梯狀的變化，在之後操作 FTJ 的熟練度上升之後，FTJ 可以由單一態(single state)轉變成為複合態(multi-state)，並進一步的提升電路中的

神經元效能。

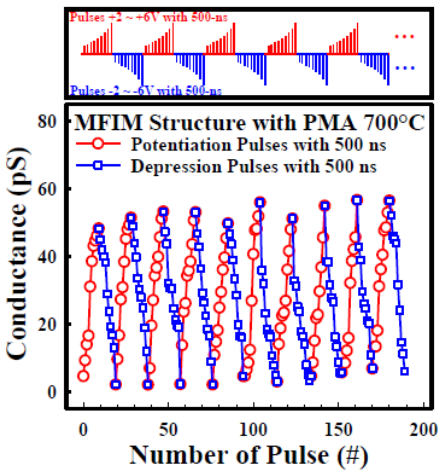

圖 3.

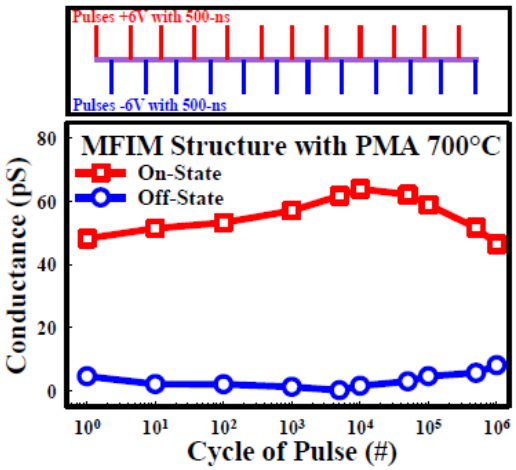

圖 4.

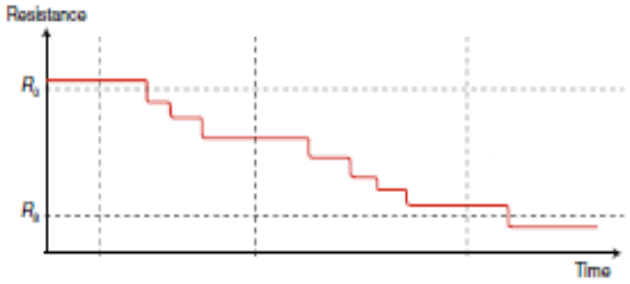

圖 5.

FTJ 除了電性上適合當作神經元之外，當重複寫入相同的 pattern 時，FTJ 會自己”學習”並對輸入產生快速反應，在圖 6.中可以看到，當我們重複輸入 pulse”F”10 次，FTJ 本身會加強反應(導電度相較上升)，我們可以利用此種特性，在訓練的時候先做 pretraining 的動作，讓記憶體(FTJ)對後續要輸入的資料產生一定程度的反應性後，訂定一個相對大的 on/off 標準，便可以同時消除可能誤差和提高準確度。

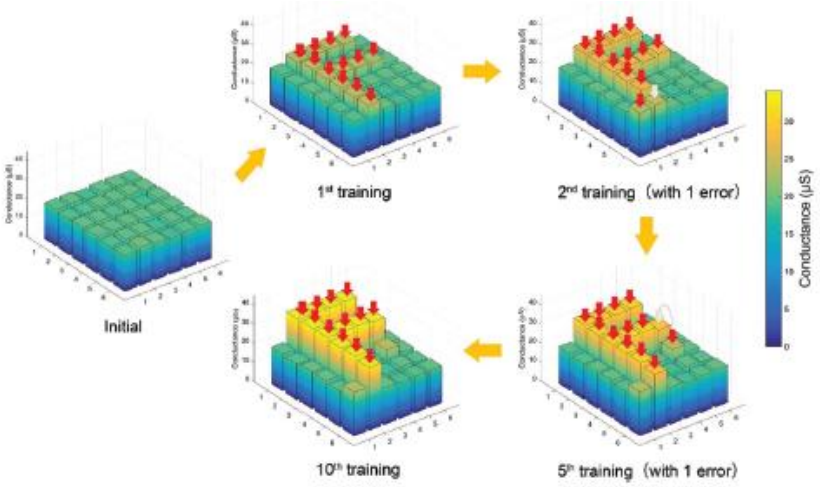

圖 6.

- Ferroelectric FET (FeFET)

FeFET 本身的製程設計和 MOSFET 略同，差別在於 FeFET 在閘極(gate)中將一部分的 polysilicon 替換為鐵電材料，讓 MOSFET 原本會因為漏電流(leakage current)而消失的態利用鐵電電容儲存(圖 7.)並保有原先 MOSFET 所具有的邏輯閘操作。

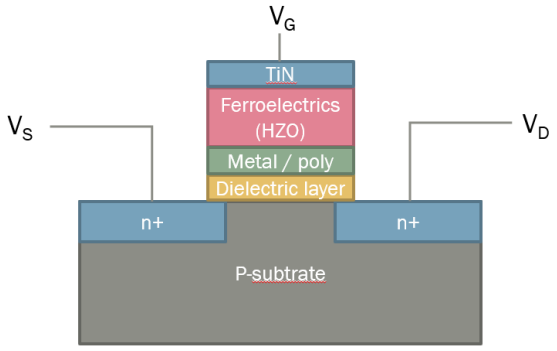

圖 7.

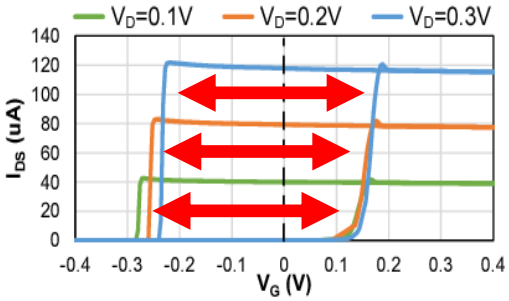

圖 8.

因為 gate 上多了鐵電電容，原先的 I-V curve 會被電容中的載子影響而有先飄移(drift)，此種飄移現象僅有在施加正向偏壓在 V_{GS} 上且具有如 FTJ 的覆寫性(可以負電壓 reset 鐵電材料的目前狀態)，因此可利用 I-V curve 上條件的滿足與否創造出和 FTJ 一樣的 conductance window，依此判別目前 FeFET 所儲存的狀態(0/1)。在圖 8.中紅色箭頭所示，施加不同的 V_{GS} 會使 window 大小有差異，讀取時施加 $V_{DS}<V_{th}$，若狀態為 1 的 FeFET 會表現出 LRS，擁有較大的 I_D; 反之，狀態為 0 時會表現出 HRS，擁有較小的 I_D(約 1/4 倍)。

設計邏輯架構

- Ferroeletric Tunneling Junction (FTJ)

圖 9.為[1]中所操作的 FTJ-MOSFET 邏輯電路，因為其所使用的 FTJ 讀寫操作時的覆寫導電度並不對稱，所以加上兩個穩流用的 NMOS 即可以把 FTJ 操作如 pseudo nmos 的電路設計，在現今 FTJ 的製程改良下，狀態改變(圖 2.)為對稱且具有覆寫性，在這種情況下可以嘗試將 NMOS 去除，以單純 FTJ 的電路設計下，來嘗試邏輯運算。

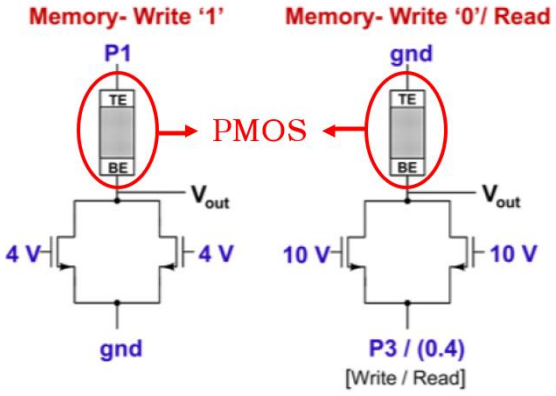

圖 9.

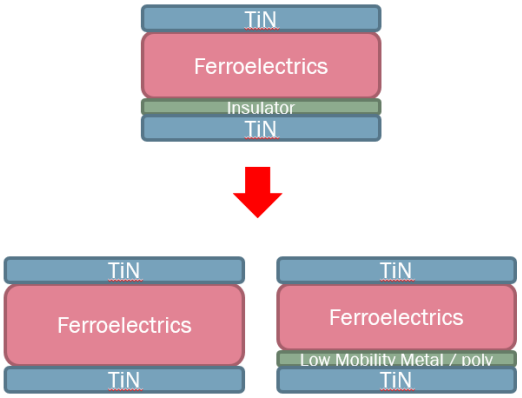

圖 10.

利用和圖 9.類似的概念，我嘗試將 FTJ 設計為 NMOS，因為 NMOS 本身可以提供基礎的 Inverter 運算並相對 PMOS 有更好的擴充性。首先，為了將邏輯運算以 FTJ units 實現，必須將 FTJ 的讀寫方向置反(前面所提供圖表及概念皆是同向讀寫)，因此，原先 FTJ 使用的 MFIM(Metal-Ferro-Insulator-Metal)架構必須被更改(圖 10.)，MFIM 架構限制於其絕緣層(insulator)的屏障(barrier)，電子通過的方向僅能是其中一個方向，此情況下只能消除這個 barrier 或是降低 barrier 的 energy gap，來讓電子可以雙向的通過，並依此實現 FTJ 的 Inverter 運算，這裡有兩個解決方法，一是去除絕緣層使 FTJ 的架構僅剩下 MFM，但此種情況可能會造成 FTJ 的電子被束縛(trap)在鐵電材料的兩側，而使得 FTJ 有大量的寄生載子產生的偏壓，使得材料穩定度大幅下降；另一種則是將絕緣層替換成低載子流動性的導體(low mobility conductor)或是 polysilicon，在大於啟動電壓的情況下(半導體的導通電壓)，保證電流經過並防止流經鐵電材料的電子被束縛，這樣一來就可以解決雙向導通的問題。

在成功實現讀取和寫入方向相反後，為了獲取更好的擴充性和方便後續的 scale up，可以將兩個 FTJ unit 以垂直方向堆疊，如此一來除了節省了平面

空間，也可以一併將 NAND(FTJ 串聯)運算放入基礎單位中，為之後的邏輯運算提供設計彈性(flexibility)(圖 12.)。

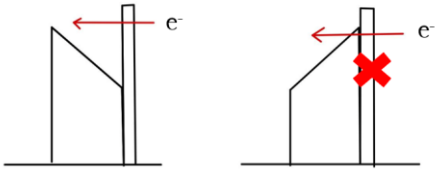

圖 11.

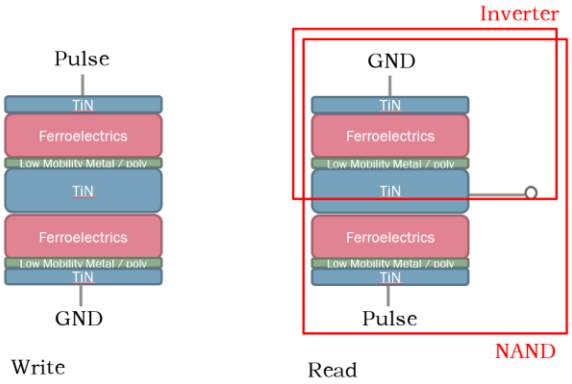

圖 12.

11.中左圖是當寫入偏壓和讀取偏壓相同時的壁壘屏障，中間區域是鐵電材料的經過寫入電壓改變後的 energy gap 示意圖，旁邊細長的矩形區域是絕緣層的 energy gap，從電子流動方向可以看到，電子僅須越過薄絕緣層就可以越過屏障到達電極位置，而如右圖寫入和讀取偏壓相反時，電子除了需要越過絕緣屏障以外，還需要越過鐵電材料自身的屏障區才能夠達到電極位置，所以若無法消除絕緣屏障或是降低越過其所需要的能量，讀取和寫入狀態無法是反向。

12.中左圖是寫入時的電壓施加狀況，右圖是讀取時的施加狀況，在讀取時可以利用中間電極層(TiN)來決定要讀取單一個 FTJ 的 Inverter 運算狀況或是兩個 FTJ 的 NAND 運算結果。

確立了邏輯運算的規則後，我嘗試利用 FTJ 來實現 SRAM 架構，見圖 13.，因為 FTJ 本身是 2port 元件，而非常見 MOSFET 為 3port 元件，在操作上需要額外的步驟才能實現 3port 的輸入，如常見 SRAM 中的 bit_line 和 word_line，在 FTJ 中必須要利用兩步驟操控，先是以 W 作為 input 對比 word_line 先寫入 pulse 至電路兩端，而後輸入反向但較小的讀取信號 R 至電路兩端後由 output 讀取相對應的改變量。

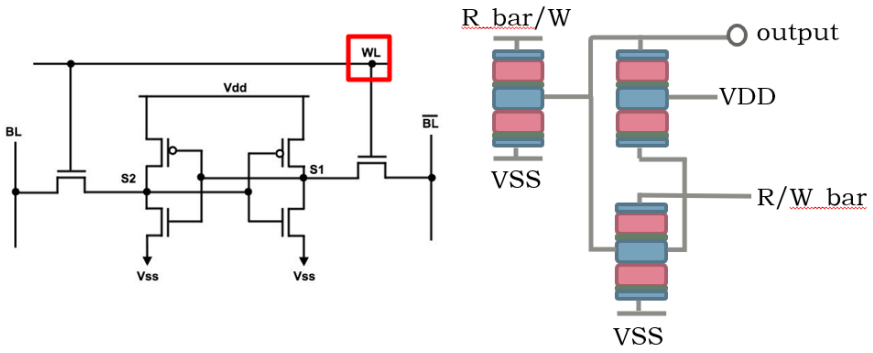

圖 13.

利用輸入方式不同，FTJ unit 可以當作 NMOS 也可以當作 PMOS，若 VDD 由 FTJ 的上下兩端輸入，可以以 NMOS 的邏輯運算看待，反之在 FTJ 中間電極輸入 VDD，FTJ 則會表現出 PMOS 的運算結果。

- Ferroelectric FET (FeFET)

為了使 FeFET 帶有資料，會先利用正負 2~6V 控制電壓施加在閘極上使鐵電材料有狀態(0/1)產生，而後 FeFET 會產生電容偏壓，使得在施加一讀取電壓(VDS<Vth)時會有不同的對應電流，利用測量電阻值的方法就可以讓 FeFET 是一個邏輯運算元，此外，為了不改變鐵電材料的狀態，讀取電壓(VDS)不能大於 1V，以免在讀取過程中誤改變了鐵電材料的排列狀態，也防止在大量讀取時會因此將現在狀態覆蓋掉(圖 14.)。 FeFET 的邏輯運算則和普通 MOSFET 一模一樣，一般的 NAND、NOR 的電路設計方式皆相同(圖 15.)，有利用後續複雜電路研究以及量測。

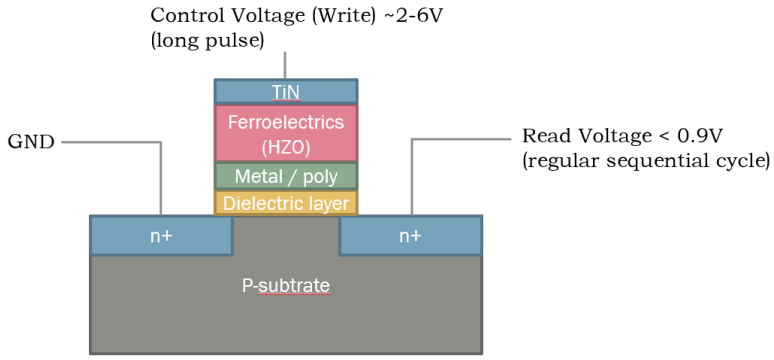

圖 14.

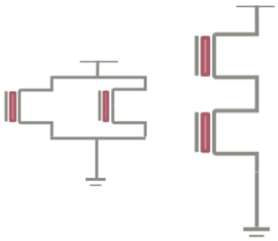

圖 15.

我試著也將 CIM 的概念套入 FeFET 中，在 Neural Network 中最重要的運算即為乘法，而乘法可以在 binary 的情況下可用 XNOR 得到，我嘗試用兩個 P-FeFET 和兩個 N-FeFET 製造 latch 並以兩步驟操作來得到 XNOR 的結果(圖 16.)。

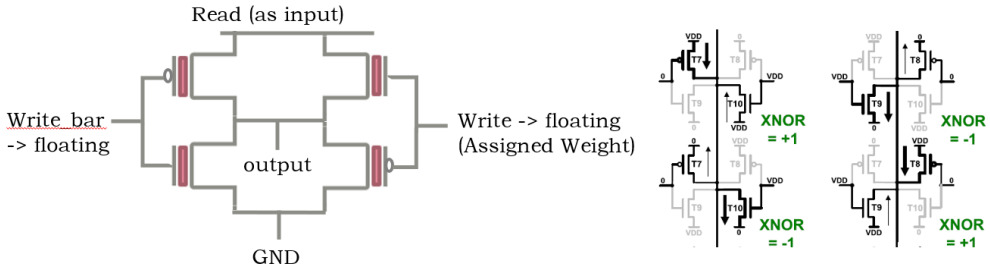

圖 16.

首先我將權重 (weight->write)以正反向輸入至兩端的 FeFET 對，在 weight 已經充分改變鐵電材料的狀態後，把這兩頭輸入端設定為 floating，以防止後續操作時鐵電材料因此而失去狀態，接著將需要運算的值(input)以 read 輸入至 FeFET 對，當權重(0：GND, 1：5V)和輸入值(0：Read < Vth,

1：Read > Vth)不同時，所開啟的 FeFET 也會有對應的改變，詳細的權重和輸入對應圖可見右上圖，以此架構為 Network 陣列的單位運算元，可以在訓練時的重複讀取上獲得相當顯著的成長，除此之外，運算後的結果亦可以存入後續 FeFET 陣列中，利用加減重組預測結果。

總結

在研究過程中可以明顯感受到無論是 FTJ 或是 FeFET 皆是發展 CIM 電路上相當有潛力的材料，FTJ 不只可以縮減大幅的製程面積，在讀寫上的速度還有可靠性都符合神經元網路所需要的，而 FeFET 在不耗損多餘效能且匹配現今製程技術的情況下，利用鐵電電容改善原先 MOSFET 令人即為詬病的漏電流，卻仍舊保有 MOSFET 的運算彈性。以 FTJ 或是 FeFET 為基底構想的 CIM 設計還有許多可以改善的地方，數據也有很多不足的東西，因為這兩種新穎材料還有其餘面向，如: noise、interference impacts、tolerance 等還未研究明瞭，我會利用寒假沒有課業壓力的期間試著將所構想的電路以 Spice 的方式呈現及驗證，利用得出數據一步一步進行調整，期望能在 5 月時完成大致研究，在此非常感謝兩位教授在這個學期的教導，我會努力完成這個研究的!

引述

1. Sandeep Kaur Kingra, Vivek Parmar, Che-Chia Chang, Boris Hudec, Tuo-Hung Hou, and Manan Suri: *SLIM: "Simultaneous Logic-in-Memory Computing Exploiting Bilayer Analog OxRAM Devices"*

2. Yi-Shan Kuo, Shen-Yang Lee, Chia-Chin Lee, Shou-Wei Li, and Tien-Sheng Chao: "*CMOS-Compatible Fabrication of Low-Power Ferroelectric Tunneling Junction for Neural Network Applications*"

3. FTJ & FeFET details supported by Yi-Shan Kuo

4. All other backgrounding papers in: https://reurl.cc/o99aKl