National Tsing Hua University
CS5120 VLSI System Design

系級：電機系
姓名：陳力豪
學號：107061272

---

1. Functionality:

- Overall hardware architecture:

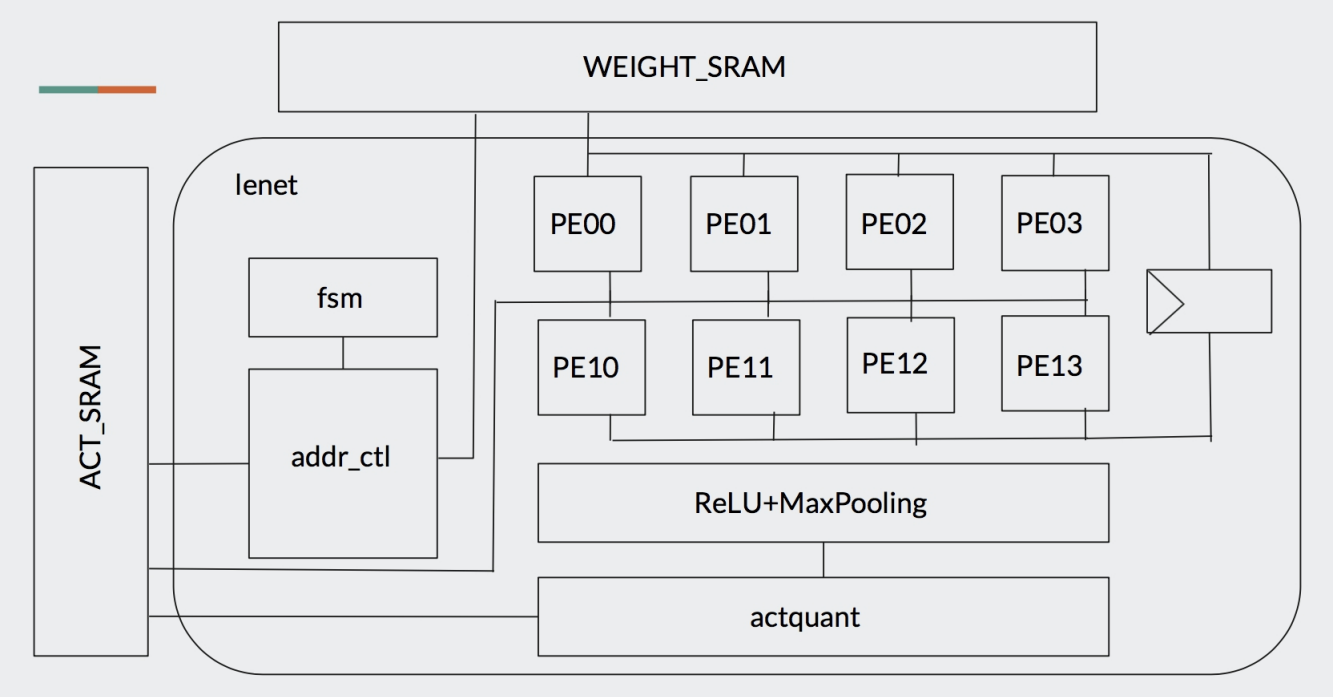
  在這次的架構中，我將 module 切割成五個部分來完成，分別是 1. lenet.v、2. fsm.v、3. addr_ctl.v、4. pe.v 以及 5. actquant.v，其中 lenet.v 是 top module。在 lenet.v 中我將其他的四個 sub-module 整合起來，主要做訊號的傳接、資料 ReLU 或 Maxpooling 以及處理寫入(寫出)訊號延遲處理。

  而不同於 HW2 的 LentModel 操作順序，由於 scale 必定為正的，故我將 acquant、ReLU 以及 Maxpooling 的順序調換。我先將 PE 出來的值進行 ReLU 並馬上做 Maxpooling，如此一來每次僅會有 2 筆資料需要經過 actquant。在我考量面積的情況下，選擇只開一顆 actquant，因此調換順序造成的好處就是我的 2 筆資料在 interval delay=1 的 pipeline 架構之下，僅需要 x+1 個 cycle 即可完成 2 筆資料的 quantization(x 為 actquant 內部所切的 stage 數目，也是第一筆資料完成 quantization 所需要的 cycle 數)。

  而 fsm.v 以及 addr_tcl.v 則是透過當下的 state 來決定特殊參數目前的值，例如 input_channel、output_channel 或者是操作 act_SRAM 以及 weight_SRAM 時所需要用到的 offset 值。並透過這些參數來判斷當下所需要讀（寫）的 sram_addr 為何以及操控 sram_wea 來控制目前需要持續讀入資料或者是已經計算完成要寫回 sram 中。

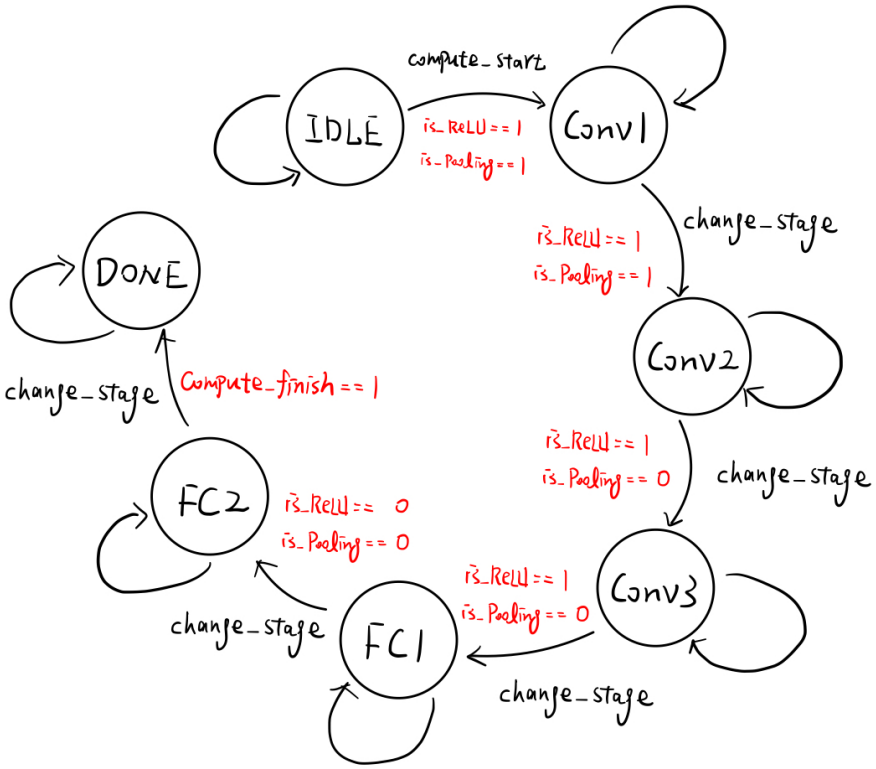
  而 pe.v 還有 actquant.v 則是 dataflow 的主要 module。在整個架構中我總共開了 8 個 PE 來「同時」處理 2 組 Maxpooling 所需要的 8 個 Convolution 運算，並且利用 Pipeline 的方式對資料進行點乘並累加。而 Fully Connected Layer 則可以使用其中一顆 PE 來運算，藉此重複使用硬體資源節省面積（其中 dataflow 更詳細的說明會在下方補充）。而 Convolution layer 經過 Maxpooling 後會產生 2 個 output，Fully Connected layer 則會產生 1 個 output，故我使用一顆 actquant 並且設計 interval dalay=1 的 Pipeline 結構即可處理。等到第二筆輸出資料從 actquant 送出後，即可準備將這兩筆(或一筆)資料寫回 SRAM。

- Block Diagram:



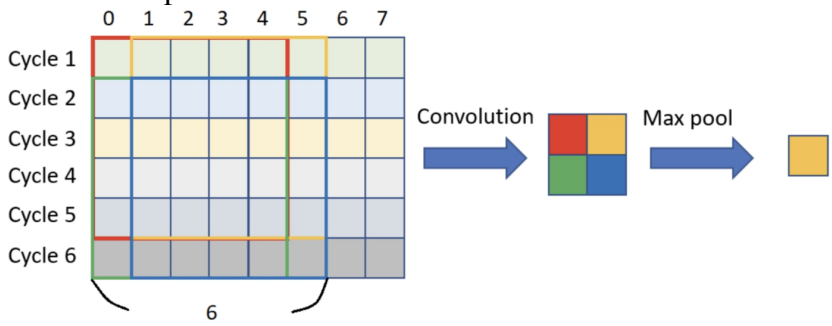
(Figure1.)

- FSM Diagram:



(Figure2.)

- Dataflow Explanation:



(Figure3.)

由於 act_SRAM 以及 Weight_SRAM 一次可以讀 2 組 address 的資料進來，故根據 Figure3. 所示，只要花 6 個 cycle 的時間即可完成 8 個 5x5 的 Convolution 運算。而我使用了 2 種 data reuse 的方式來減少資料的重複讀取：

1. Spatial Reuse：

由於除了須考慮第一個 cycle 下排的 PE 不需要使用該筆資料以及第 6 個 cycle 上排的 PE 不需使用該筆資料之外，其餘每個 Cycle 都可以將讀入的資料拿來做運算，故我透過「Broadcast」的方式，將讀入的資料同時傳給 8 顆 PE 做運算。

2. Temporal Reuse：

由於第一個 Cycle 會讀入 5x5 Kernel 的第一個 row，但是當下的資料並不能給下排的 PE 做運算，故我會在下一個 clk 敲起時，將當前輸入的 weight value 暫存起來，接著再把暫存住的 weight value 送入下排的 PE 做運算。故只要第二個 Cycle 之後，上下兩排 PE 就都可以開始進行 Pipeline 的點乘與累加。Weight 只需要一直讀入新的 value，不需要回頭重新讀前幾筆資料。

2. Result:

Item	Description
RTL simulation	PASS
Gate-level simulation	PASS
Gate-level simulation clock period	1.65 ns
Gate-level simulation clock latency	29962 cycles
Total cell area	45118.79 um^2

3. Others:

NONE