深度神經網路FPGA加速器
Deep Neural Network FPGA Accelerator

組別：A186

組員姓名：廖一心、林書佑、陳信翰、吳定曉

指導教授：鄭桂忠 教授

Mentor：鄭揚翰、陳禹樵

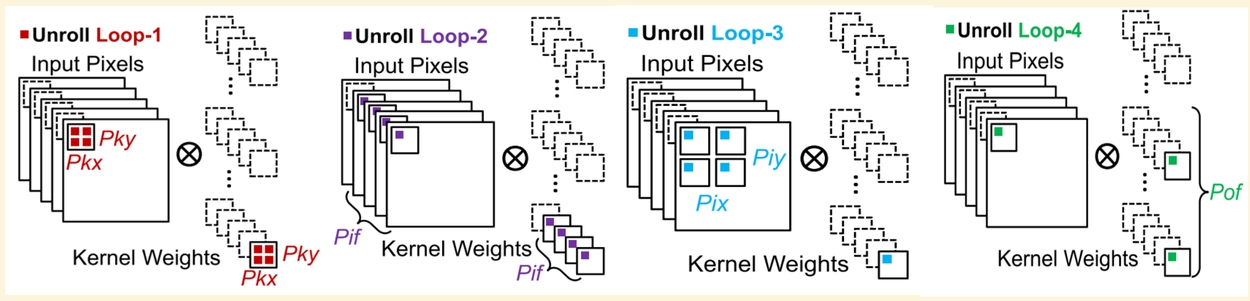
Abstract

本專題將在CNN架構的基礎下，討論幾種不同資料流(dataflow)，像是**Weight stationary**、**Output Stationary**等，以及**Processing Element(PE)**的擺放位置，還有不同卷積技巧所形成的架構。從討論中可以比較出各架構**定性化的優缺點**，而後以定量化的方法，進而可以得到各架構的表現，再實際量測其運算所需的**面積以及能耗**。

FPGA包含GPU處理器，較適合處理簡單、重複性高的工作，並搭配平行處理來大幅度的降低卷積的運算時間。在應用端，我們將討論後最合適的架構應用至**預測肺癌**的CNN model，並利用FPGA板來進行此模型中卷積的加速，以提高整個CNN model的運算速度，如此一來才能更添加用來預測肺癌之卷積模型的**即時性(Real time)**。

Background

1. **W**eight **S**tationary(**WS**): 讓權重固定在某些暫存器(Register)當中，藉由讓input流經這些有權重的暫存器，並將乘加後的結果(Partial Sum, psum)再往下一個PE傳。
2. **O**utput **S**tationary(**OS**): 將每一個輸入位置與權重的乘積以PE暫時固定，在下一個cycle會把下一個輸入位置與其相對應的權重做相乘，加上原本已經儲存在運算單元中的暫存值，更新儲存在運算單元中的psum。
1. Unroll Loop 1~4 (URL1~4): 如右圖。

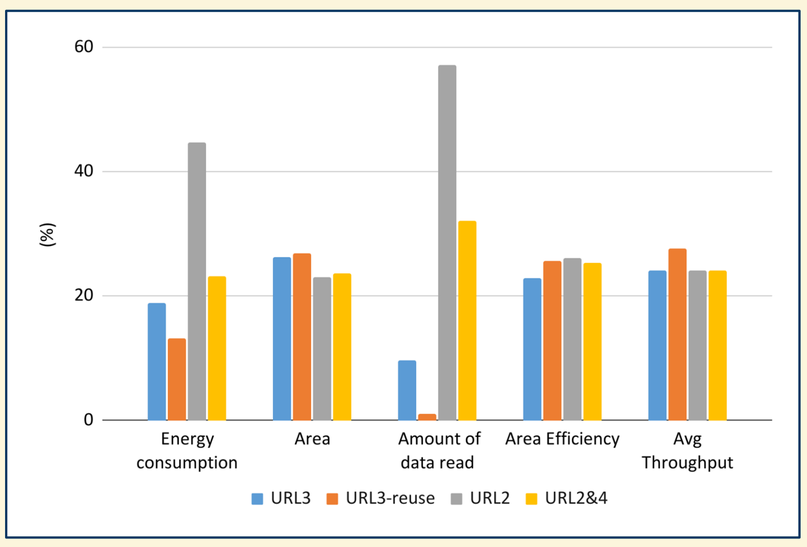


Research Method

在我們的架構當中，皆以**3×3×d×k**的**權重(weight)**以及**n×n×d**的**輸入(input)**作為計算參考，其中權重與輸入皆為4 bits，並且採取**5種指標**來比較架構差異與優缺，分別是：**平均吞吐量**(Average Throughput)、**面積效率**(Area efficiency：speed/area)、**讀取資料總量**、**計算能耗**、**面積**。

Result

下圖為我們透過定量化分析之後的結果，並將其標準化後，繪製出的長條圖。

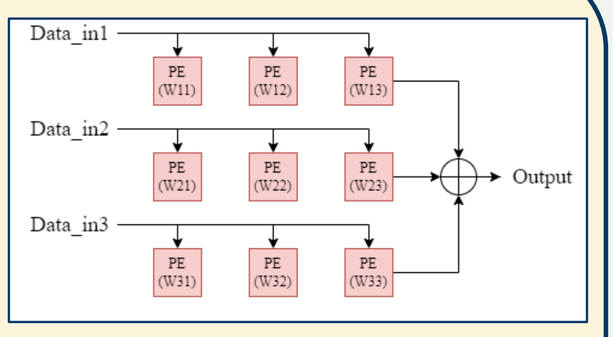


Implementation

◎ **URL3**(運算單元架構如右圖)

優點：Zero Padding Storage-減少了20% RAM的儲存量。Flexibility-只需讓PE9多算幾個週期即可完成與更大輸入大小的卷積。

缺點：Partial Sum很大。



◎ **URL3_reuse** (Data Flow設計如下圖(a))

優點：Data Reuse-大量減少記憶體的儲存空間，也可以減少讀取記憶體的次數。

缺點：Buffer空間浪費-當輸入層大小非4的倍數時，緩衝器(Buffer)必需額外補0才能維持硬體的正常運作。

◎ **URL2** (運算單元架構如下圖(b))

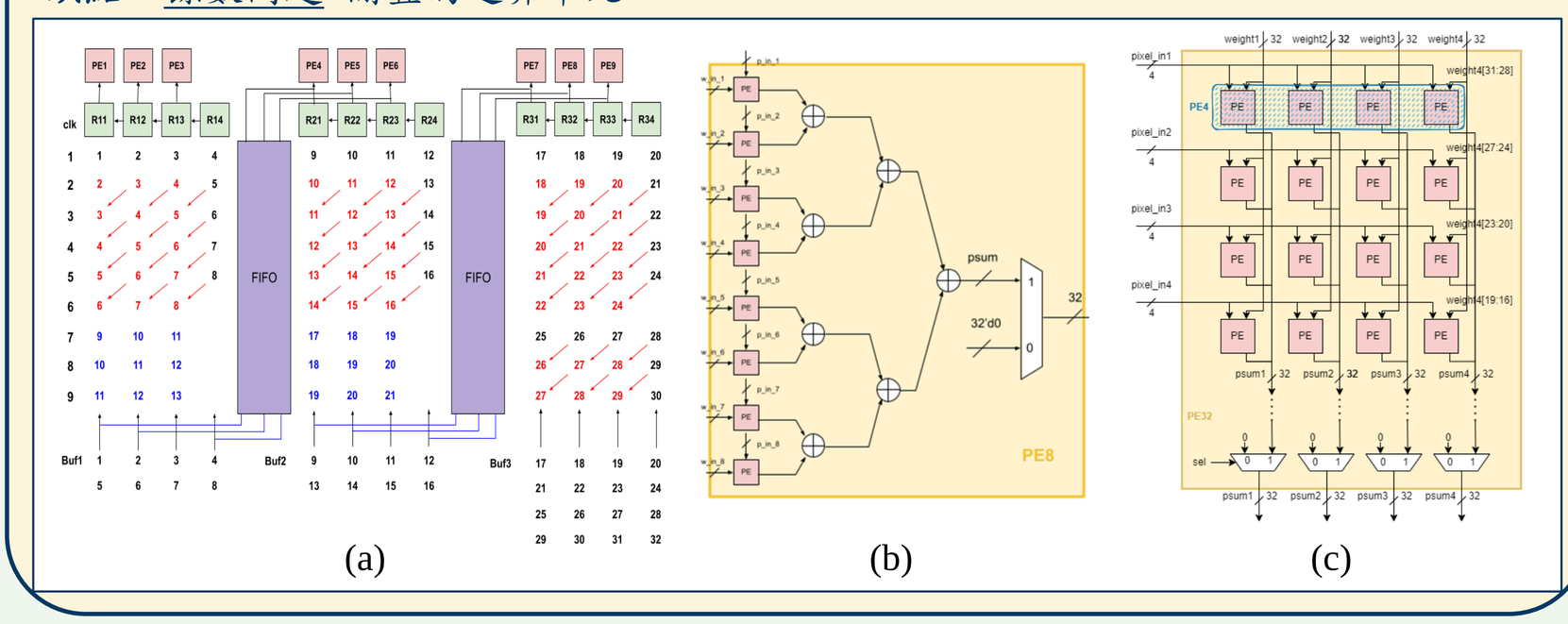
優點：架構小-減少面積並增加運算效率；Local Update-不須額外緩衝器存取psum。

缺點：No Data Reuse-造成更多延遲與能耗；吞吐量低-架構小所造成。

◎ **URL2&4** (運算單元架構如下圖(c))

優點：吞吐量大-在一個cycle中計算量大、平行度-放大Unroll Loop 2架構。

缺點：餘數問題-閒置的運算單元。



Application

◎ **應用方面**：

將一個CNN模型，放置FPGA上運算，進行加速，以達到及時性(Real Time)。

◎ **設計**：(系統架構之方塊圖如下)

1. 使用乒乓法(PingPong)，執行多層運算。
2. 特別設計Instruction Code，使硬體能計算不同輸入大小及不同運算層(全連接與卷積層)。

◎ **結果與驗證**：

在軟體上，運算總共花了50毫秒。加上硬體加速器後，運算花了1.3毫秒。

◎ **訓練CNN模型方面**：

利用軟體進行訓練(training)，並將在軟體得到的權重，以Dofera進行量化，用csv檔存取以便FPGA使用。

◎ **硬體實作方面**：

將已經設計好的架構包成IP，透過AXI4的協議，與FPGA的軟體端(PS)相連，便能以軟體控制我們的硬體設計。