

DNN accelerator: The hardware design and power analysis of RS, WS, and OS

DNN加速器：RS與WS、OS的硬體設計與能耗分析

指導老師：黃朝宗

組別：B156

組員：李宇洲、王傑立、王靖淳

摘要

深度學習運算過程會需要使用到大量的資料，包括image、filter weight、bias等且儲存在記憶體中，多次存取記憶體會提高energy consumption，因此出現幾種dataflow提高energy efficiency，例如**WS** (weight stationary)、**OS** (output stationary)和**RS** (row stationary)等。其中在Efficient Processing of Deep Neural Networks: A Tutorial and Survey這篇paper中表示RS的energy efficiency最好，所以我們想知道是否在一些情況下，其他dataflow的energy efficiency反而會隨著應用、model、convolution layer、image size等的不同而比RS還要好？

Style Transfer

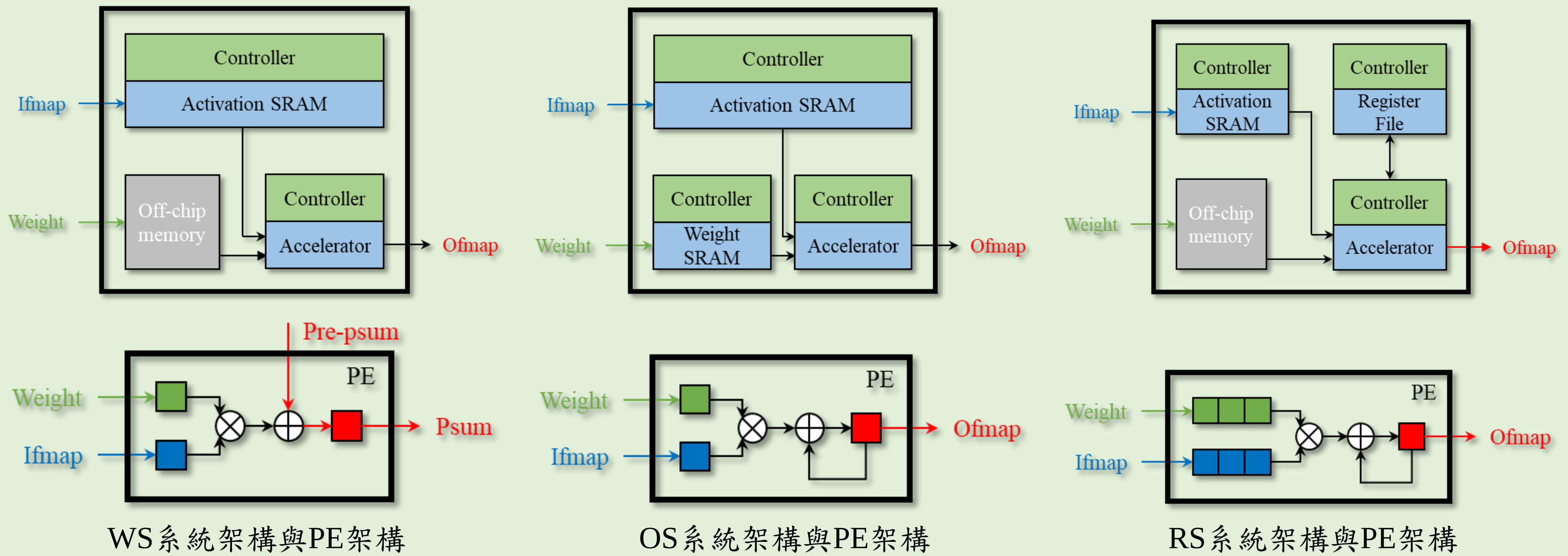


實驗方法

我們希望能模擬深度學習在實際應用中的運算情況，因此選擇**style transfer**作為加速器的應用，使用的model為**VGG19**。我們的研究方法為，先訂定規格，例如圖片大小、FPS、clock rate、PE (processing element) 數量等。接著設計三種dataflow的硬體架構，完成RTL code及測試code的正確性，合成RTL，利用pre-layout gate-level simulation的waveform做精準的power estimation。最後我們會探討在不同情況下，使用哪一種dataflow可以有較好的performance。

Dataflows

- Weight Stationary (WS)**：WS是一種重複利用 weight 的 dataflow，將weight固定在每一個PE中，activation則是會從 global buffer傳遞給每一個PE。每經過一個PE，partial sum 就會更新(累加)一次。
- Output Stationary (OS)**：OS的主要設計理念為減少partial sum的存取。DNN加速器每經過一個MAC，就必須更新一次partial sum，因此將partial sum固定存在PE中的register內，就不需要至global buffer存取partial sum。
- Row Stationary (RS)**：RS是Eyeriss神經網路加速器中所使用的一種dataflow，其設計理念是希望能最大化weight和 activation在PE array中的使用次數。



問題說明

我們需要處理的問題是要如何讓資料在加速器中的PE array中流動，且根據不同的dataflow也會有不同的架構設計，例如SRAM的大小及throughput、PE的結構、控制器等等。另外，為了確保比較上的公平性，我們必須準備對應的措施，例如input image大小多少、PE array的大小多少等。

實驗結果

我們的實驗結果顯示，在此規格上，若energy最重要就選OS，若area最重要就選RS。

	WS	OS	RS
Timing	4 ns		
Area	9,488,890 μm^2	11,359,573 μm^2	9,425,788 μm^2
Total cycle	4,512,512	4,600,005	3,747,732
Energy cost	3.383768 mJ	1.805017 mJ	2.05629 mJ

