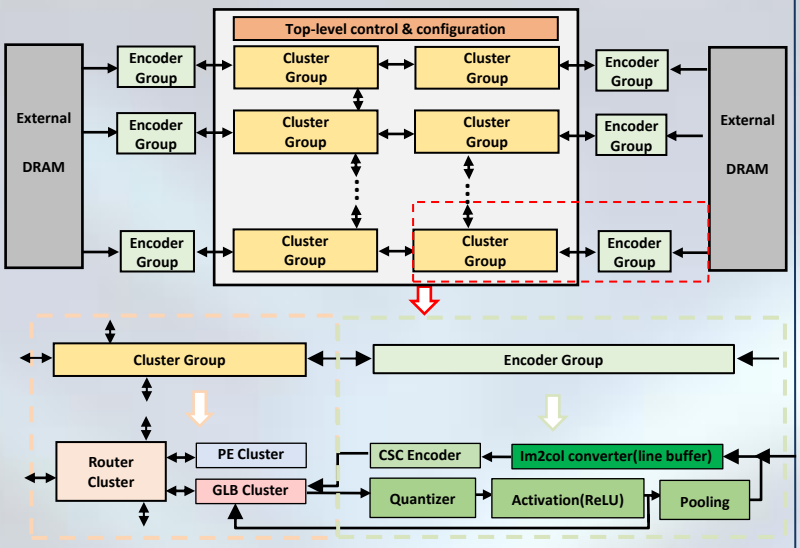
Implementation of a Flexible and Energy-Efficient Accelerator For Sparse Convolution Neural Network

摘要

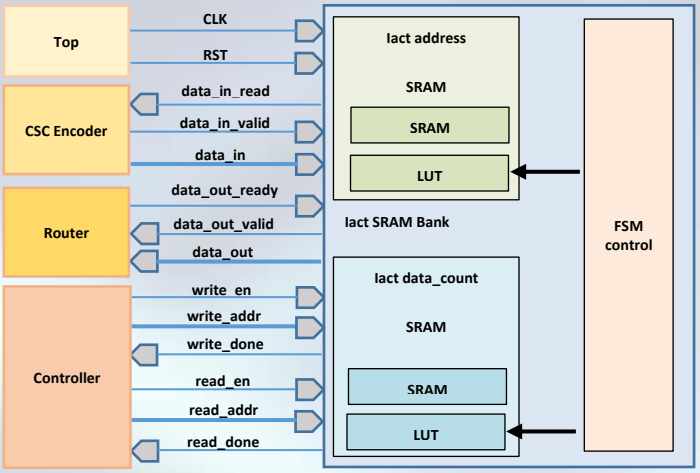
我們設計了一種基於 Eyeriss v2 架構的通用型深度神經網絡硬體加速器架構，結合了 im2col+GEMM 的數據重構模式，簡化了資料處理同時也增加了硬體架構的靈活性，並且在 NoC 層面引入了 Systolic Array 的設計，解決了原始 Hierarchical Mesh-NoC 架構中組合循環的問題，此外我們也在資料流的設計中對於 input feature map 和 weight 使用相同的硬體模組，並且在不同 layer 之間交換 input feature map 和 weight 的儲存位置以最大化地復用(reuse)資料。最後使用 OpenROAD，將我們設計的硬體加速器合成至 GDSII。在 NanGate 45nm CMOS 的製程下，系統推理 MobileNet (Convolution only)達到 1559.7 Inference/sec (batch size=1, Sparse ratio = 0.5)，此外也有將硬體實作於 FPGA 平台驗證及展示。

硬體系統架構

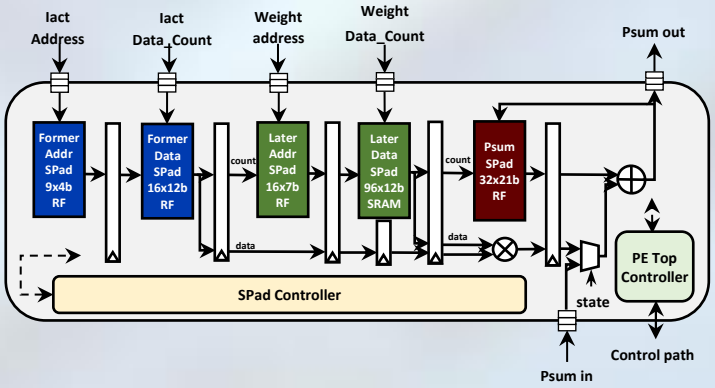
I. Top-level



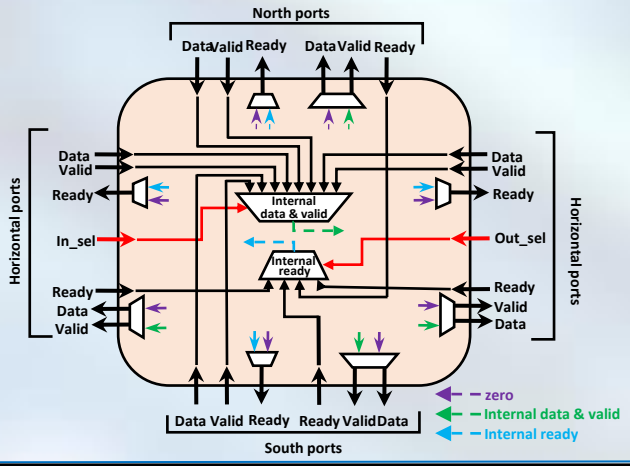
III. GLB Architecture



II. PE Architecture



IV. Router Architecture



Data Flow

	Data flow		SPad Stored data type	
Layer type	Inter-PE	Psum SPad	Former SPad	Later SPad
CONV	WS	Vertical Accumulate	Weight	Iact
DW-CONV	IS		Iact	Weight
FC	IS		Iact	Weight

CSC Data Format

	3		5	7
	4			8
1			6	9
				a
2				b

CSC Compress data:

Address vector : {2, 31, 4, 6, 11, 0}

count vector : {2, 4, 0, 1, 0, 2, 0, 1, 2, 3, 4, 0}

dara vector : {1, 2, 3, 4, 5, 6, 7, 8, 9, a, b, 0}

實作成果

ASIC

Specifications	
Area	1.53mm*1.53mm
Gate count(logic only)	1512k (NAND-2)
Total power	241.2mW
On-Chip SRAM	225KB
Number of PE	192
Scratch Pads (per PE)	weight addr: 14B(RF) weight data: 144B(SRAM) iact addr: 4.5B(RF) iact data: 30B(RF) psum: 84B(RF)
Clock Rate	200 MHz
Peak Throughput	76.8GOPS
Arithmetic Precision	weights & iacts: 8b fix-point psum: 21b fix-point

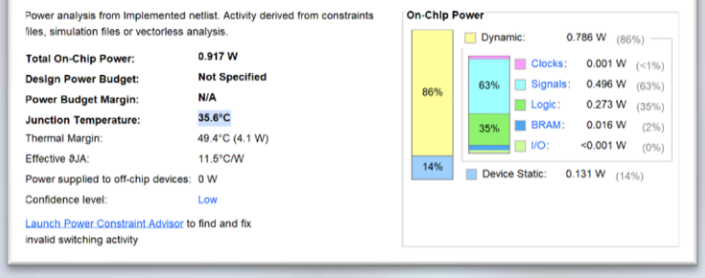
Performance			
DNN	**Metrics**	**Eyeriss v2**	**This work**
Sparse AlexNet	Nominal Num. of MACs	724.4M	
	Inference/sec	278.7	335.6
	DRAM acc.	22.3 MB	13.0 MB
	PE Utilization[2]	100%	100%
	Sparse Ratio	Not mention	0.9
Sparse MobileNet	Nominal Num. of MACs	49.2M	
	Inference/sec	1470.6	1559.7
	DRAM acc.	3.9 MB	2.2 MB
	PE Utilization	91.5%	100%
	Sparse Ratio	Not mention	0.5

FPGA (PYNQ-Z2)

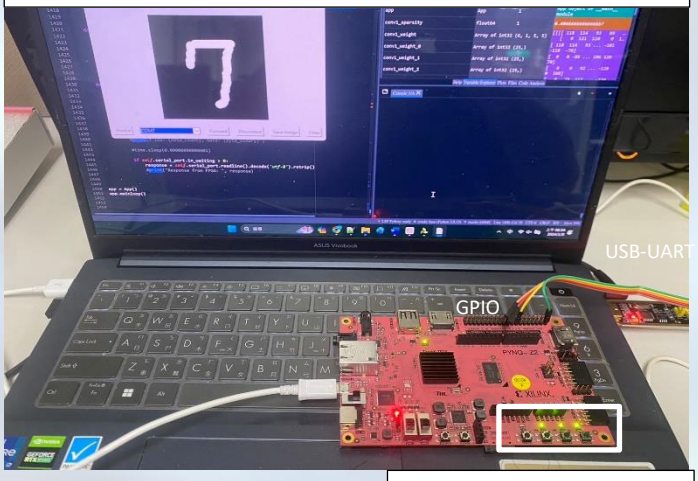
Resource Utilization

Resource	Utilization	Available	Utilization %
LUT	39577	53200	74.39
LUTRAM	339	17400	1.95
FF	51819	106400	48.70
BRAM	134.50	140	96.07
IO	7	125	5.60

Power(125MHz)



DEMO—Handwritten digit recognition



LED:0111=7 in binary

結論

我們改良了 Eyeriss v2 的架構，引入 im2col+GEMM 的資料重構方式，擴增 Encoder 讓整體系統更加靈活且泛用，也提出了在 PE 中交換資料類型的儲存模式，此外，我們在 GLB 中使用 0 來當作 data stream 的結尾偵測，且配有 LUT 來記錄 data stream 的起始位址，這樣的作法雖然增加了一些儲存空間，但就不需要額外的控制邏輯與 Top-level 的配置來控制資料讀取的數量，當有需要更改參數配置時也不需要調整上層控制的邏輯，也讓整體的擴展更加靈活。

我們透過 OpenROAD 將 RTL 合成至 GDSII，在 NanGate 45nm 的製程下，layout 的面積為 2.34 mm^2，並且只使用 1512k 個 NAND-2 來實現，相對 Eyeriss v2 少了 0.57X，在相同 MAC 數下 AlexNet 和 MobileNet 的推理速度分別可以達到 335.6 以及 1559.7Inference/sec 這相對於 Eyeriss v2 分別快了 1.2X 和 1.06X，且 power 只有 241.2mW，這比 Eyeriss v2 在最低功耗的神經網路(Sparse AlexNet)上運算還要低了 0.42X。