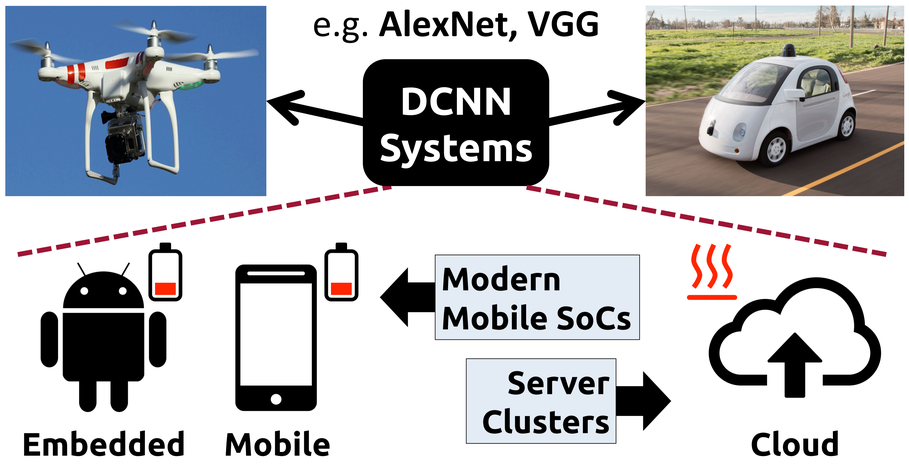
Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks

Yu-Hsin Chen[1], Tushar Krishna[1], Joel Emer[1,2], Vivienne Sze[1]

[1] Massachusetts Institute of Technology, Cambridge, MA [2] NVIDIA, Westford, MA

1. Motivation

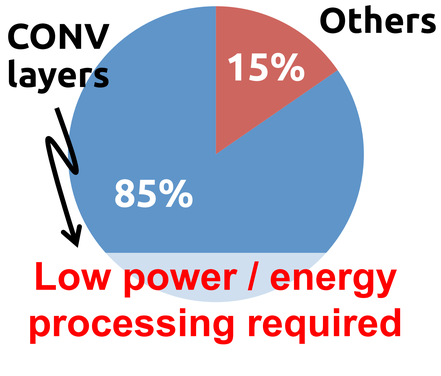
- Accelerators are required to run **state-of-the-art** CNNs at **high throughput** and **low power/energy**



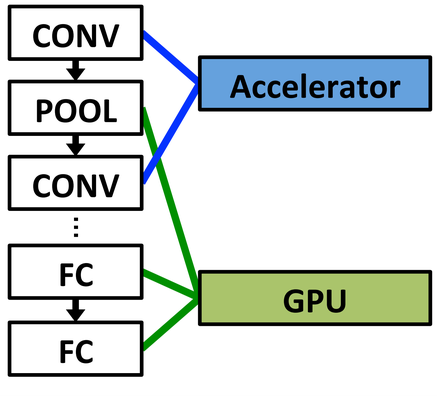
2. Key to High Efficiency

- Focus on the most **computation** and **energy** demanding block – **Convolutional (CONV) layers**

DCNN Energy Breakdown

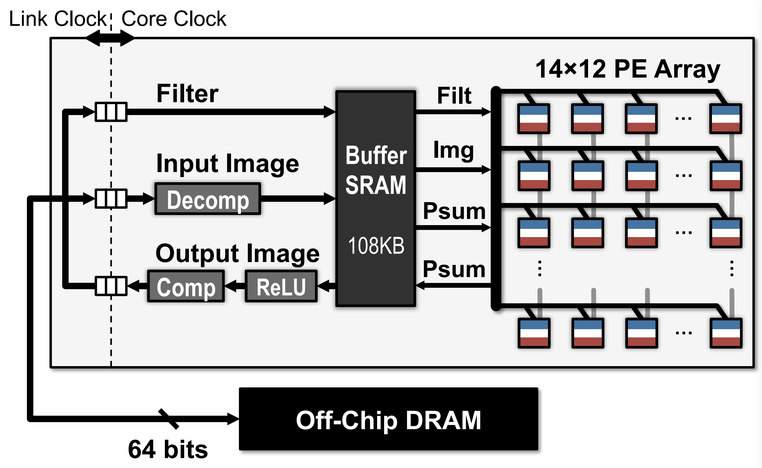


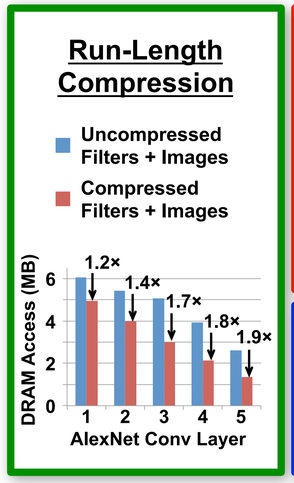
DCNN Processing Pipeline

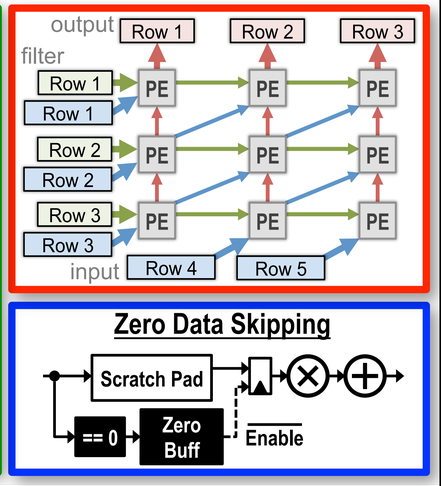


3. Eyeriss Architecture

- 168 PEs with **Reconfigurable** Dataflow Mapping to Maximize Data Reuse
- **Data Compression** (Reduce DRAM BW) and **Zero Skipping** (Reduce Processing Power)

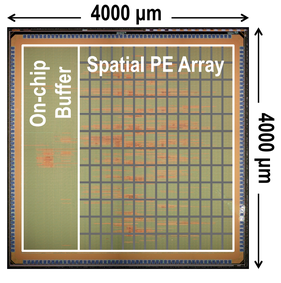


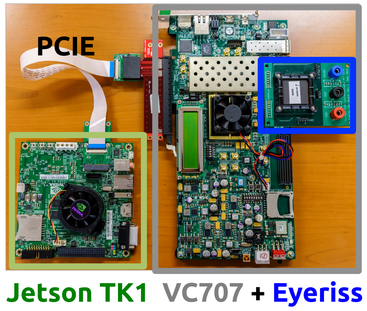




4. Implementation

Technology	TSMC 65nm LP 1P9M
Core Area	3.5mm×3.5mm
Gate Count	1852 kGates (NAND2)
On-Chip Buffer	108 KB
# of PEs	168
Scratch Pad / PE	0.5 KB
Supply Voltage	0.82 – 1.17 V
Core Frequency	100 – 250 MHz
Peak Performance	33.6 – 84.0 GOPS (2 OP = 1 MAC)
Word Bit-width	16-bit Fixed-Point
Filter Size*	1 – 32 [width] 1 – 12 [height]
# of Filters*	1 – 1024
# of Channels*	1 – 1024
Stride Range	1–12 [horizontal] 1, 2, 4 [vertical]





Jetson TK1 VC707 + Eyeriss

5. Benchmark / Demo

AlexNet Performance Benchmark

Batch Size = **4** (227×227 frame)
200 MHz Core / 60 MHz Link

Layer	Power (mW)	Latency (ms)	# of MAC (MOPs)	Active # of PEs (%)	Buffer Data Access (MB)	DRAM Data Access (MB)
1	332	20.9	422	154 (92%)	18.5	5.0
2	288	41.9	896	135 (80%)	77.6	4.0
3	266	23.6	598	156 (93%)	50.2	3.0
4	235	18.4	449	156 (93%)	37.4	2.1
5	236	10.5	299	156 (93%)	24.9	1.3
Total	**278**	**115.3**	**2663**	**148 (88%)**	**208.5**	**15.4**

Eyeriss Caffe System Demo

1000-Class Image Classification with AlexNet

Link: https://vimeo.com/154012013