ETH zürich

Eyeriss v2: A Flexible Accelerator for Emerging Deep Neural Networks on Mobile Devices

Authors: Yu-Hsin Chen, Tien-Ju Yang, Joel S. Emer, Vivienne Sze

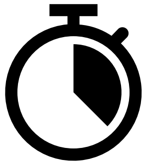
Presented by: Florian Mahlknecht

Energy Efficient Multimedia Systems Group

- Professor Vivienne Sze

Design Principles

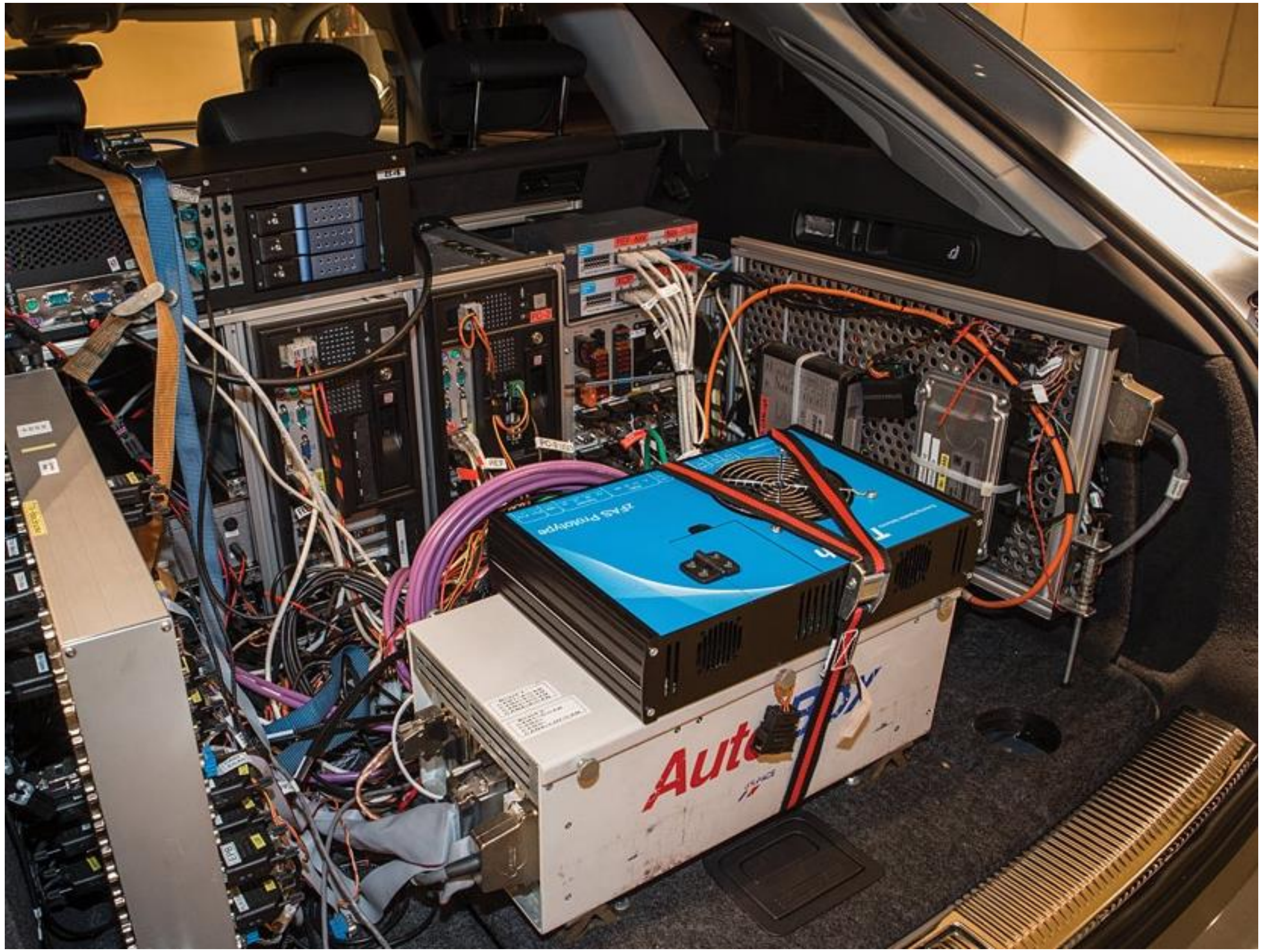
Efficiency



Latency

Flexibility

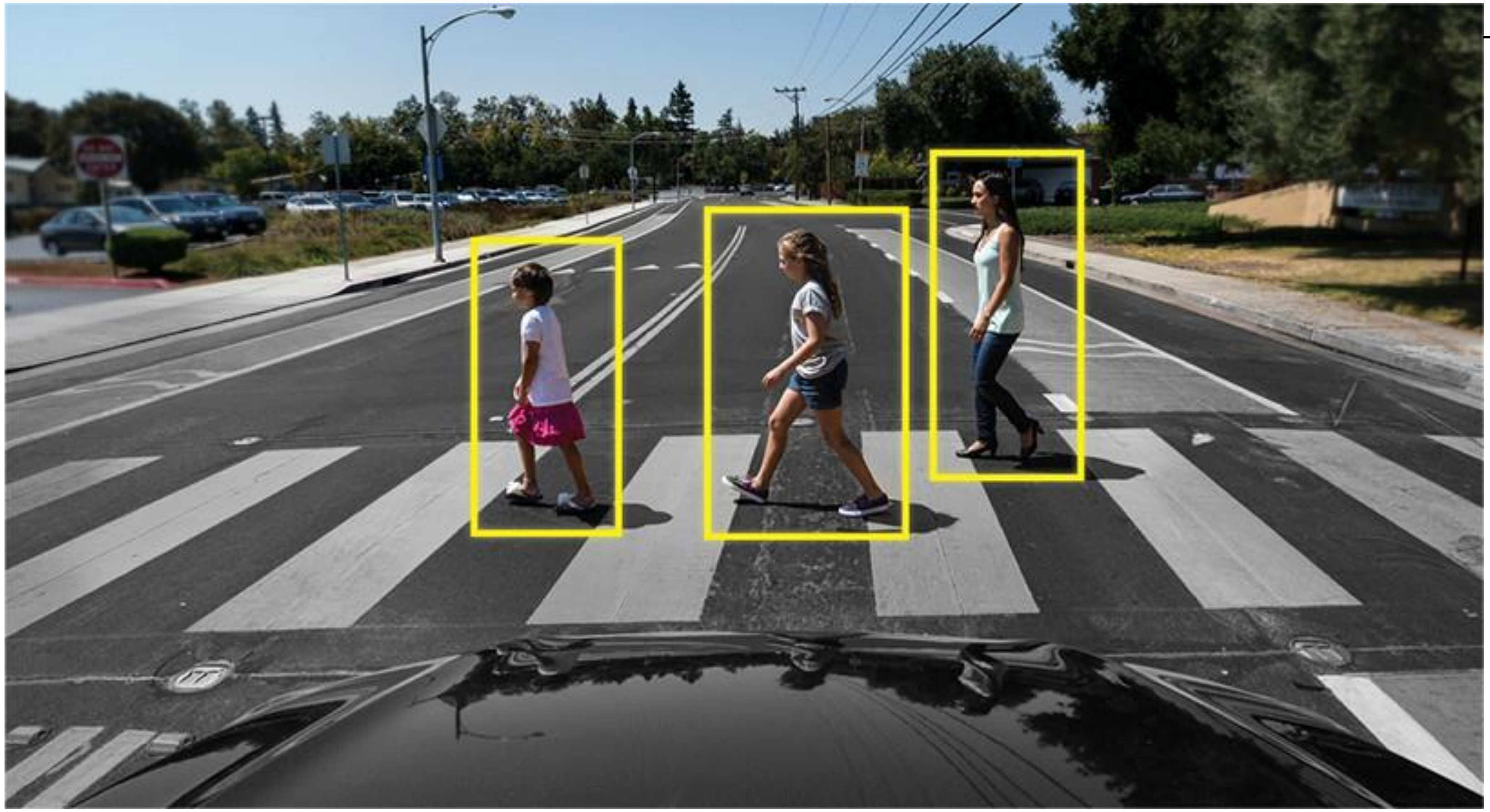
Motivation: Energy Efficiency



(slide credits: Prof. Sze, see Wired 02.06.2018)

- 6 GB of data every 30 seconds
- Avg. 2.5kW

Motivation: Latency



< 100ms

(Lin et al., 2018)

Motivation: Edge Processing

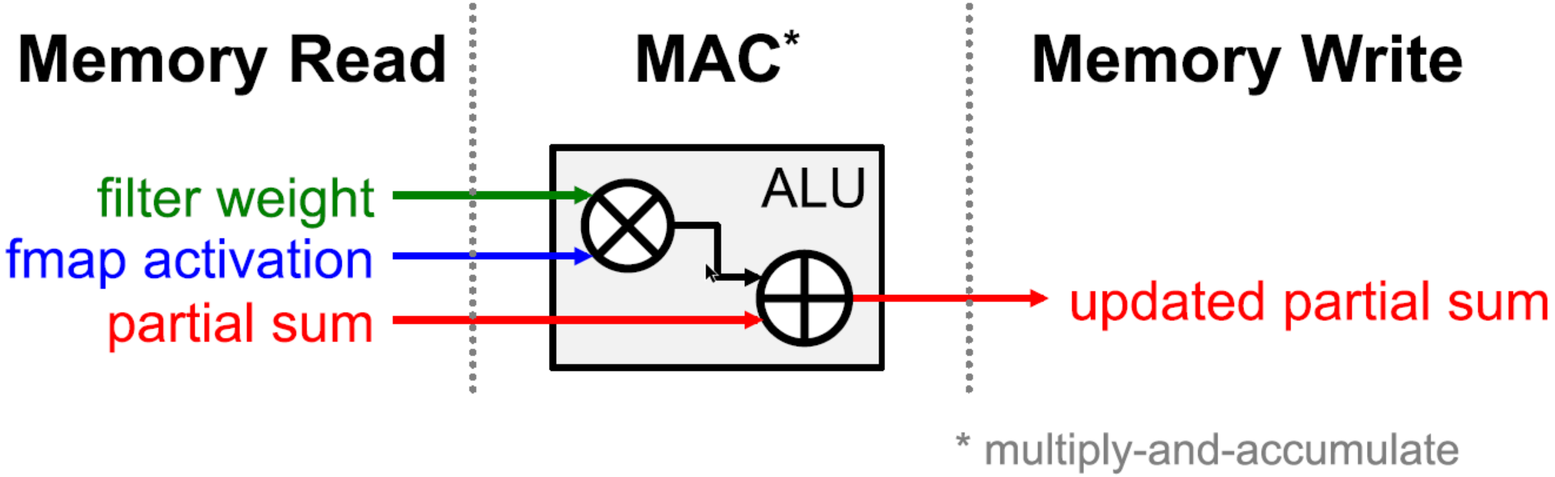
Motivation: Flexibility

Metrics	LeNet 5	AlexNet	Overfeat fast	VGG 16	GoogLeNet v1	ResNet 50
Top-5 error[†]	n/a	16.4	14.2	7.4	6.7	5.3
Top-5 error (single crop)[†]	n/a	19.8	17.0	8.8	10.7	7.0
Input Size	28×28	227×227	231×231	224×224	224×224	224×224
# of CONV Layers	2	5	5	13	57	53
Depth in # of CONV Layers	2	5	5	13	21	49
Filter Sizes	5	3,5,11	3,5,11	3	1,3,5,7	1,3,7
# of Channels	1, 20	3-256	3-1024	3-512	3-832	3-2048
# of Filters	20, 50	96-384	96-1024	64-512	16-384	64-2048
Stride	1	1,4	1,4	1	1,2	1,2
Weights	2.6k	2.3M	16M	14.7M	6.0M	23.5M
MACs	283k	666M	2.67G	15.3G	1.43G	3.86G
# of FC Layers	2	3	3	3	1	1
Filter Sizes	1,4	1,6	1,6,12	1,7	1	1
# of Channels	50, 500	256-4096	1024-4096	512-4096	1024	2048
# of Filters	10, 500	1000-4096	1000-4096	1000-4096	1000	1000
Weights	58k	58.6M	130M	124M	1M	2M
MACs	58k	58.6M	130M	124M	1M	2M
Total Weights	60k	61M	146M	138M	7M	25.5M
Total MACs	341k	724M	2.8G	15.5G	1.43G	3.9G
Pretrained Model Website	[56][‡]	[57, 58]	n/a	[57–59]	[57–59]	[57–59]

[†]Accuracy is Measured Based on Top-5 Error on ImageNet [14].
[‡]This Version of LeNet-5 has 431000 Weights for the Filters and Requires 2.3 million MACs Per Image, and Uses ReLU Rather Than Sigmoid.

(Sze et al., 2017)

Recall core operations

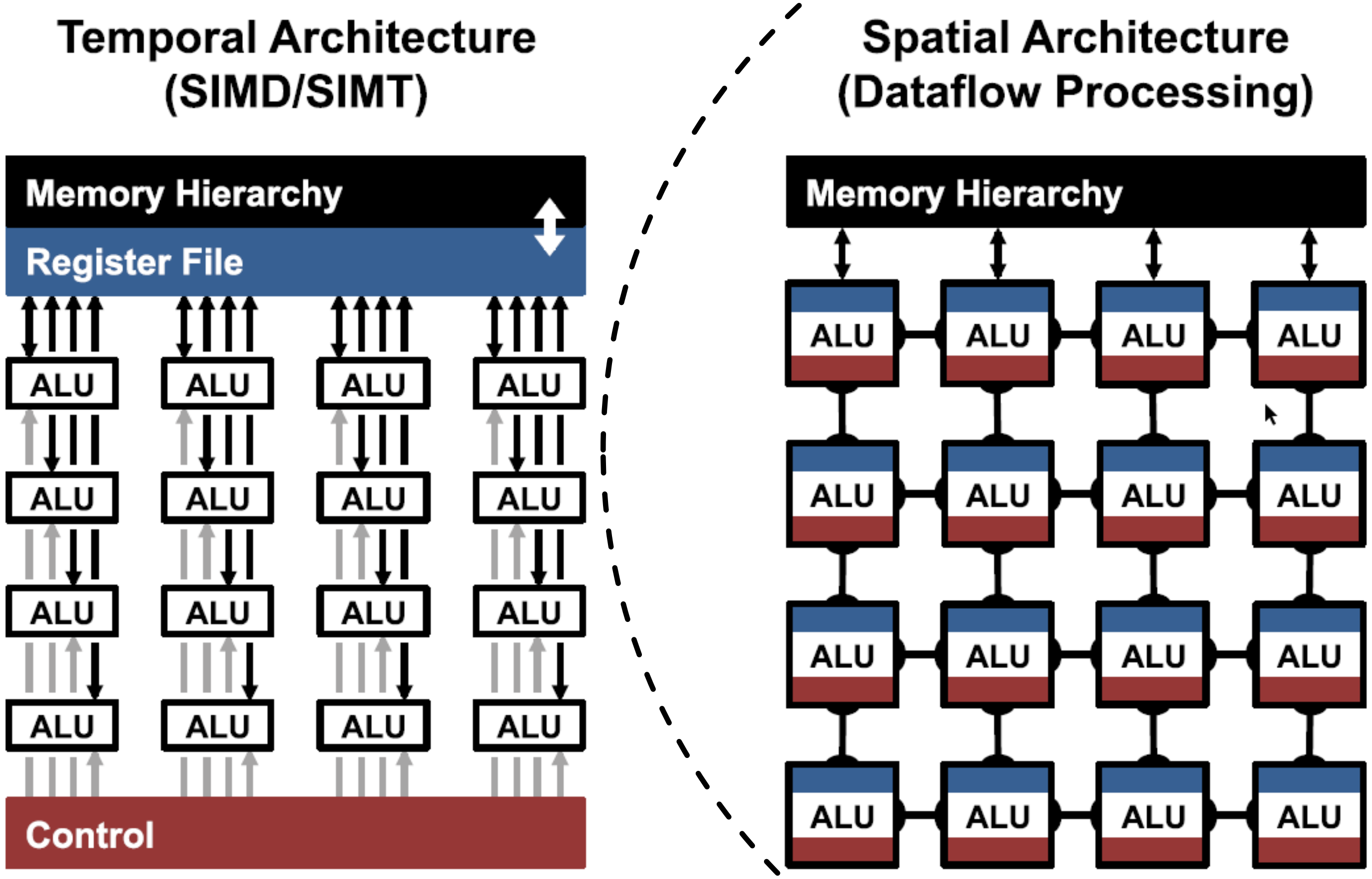


- 4 Memory transfers, 2 FLOP
- Parallelizable!

(Sze et al., 2017)

Architecture Overview

- GPU / CPU
- Vector / Thread
- Centralized control for ALUs
- data from memory

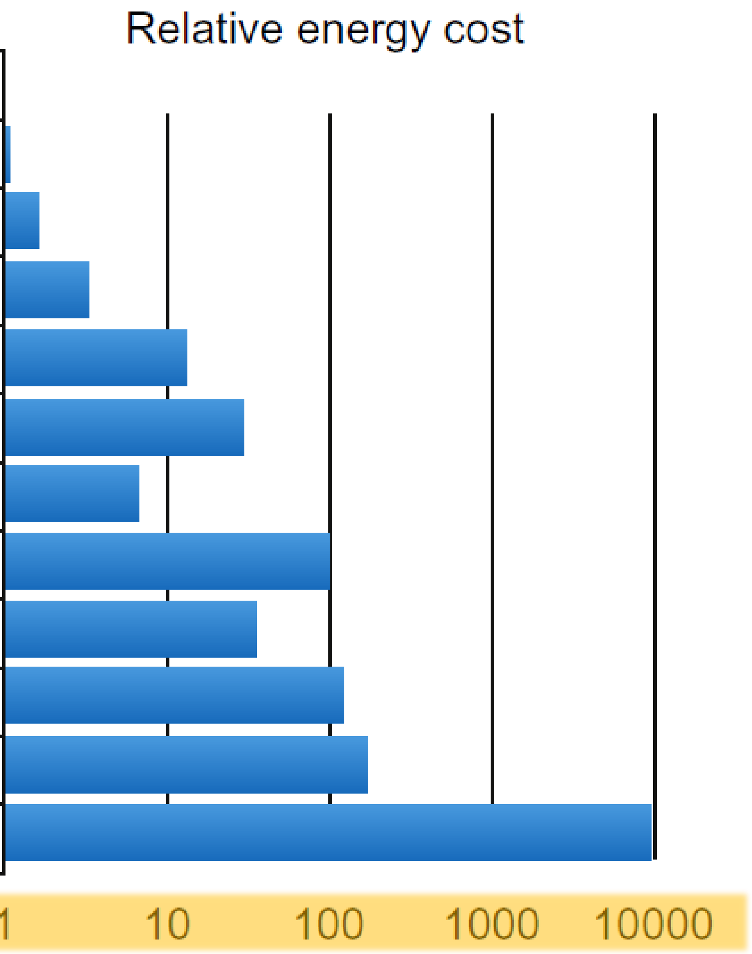


- Processing Engines
- Local memory
- Control logic

(Sze et al., 2017)

Memory access cost

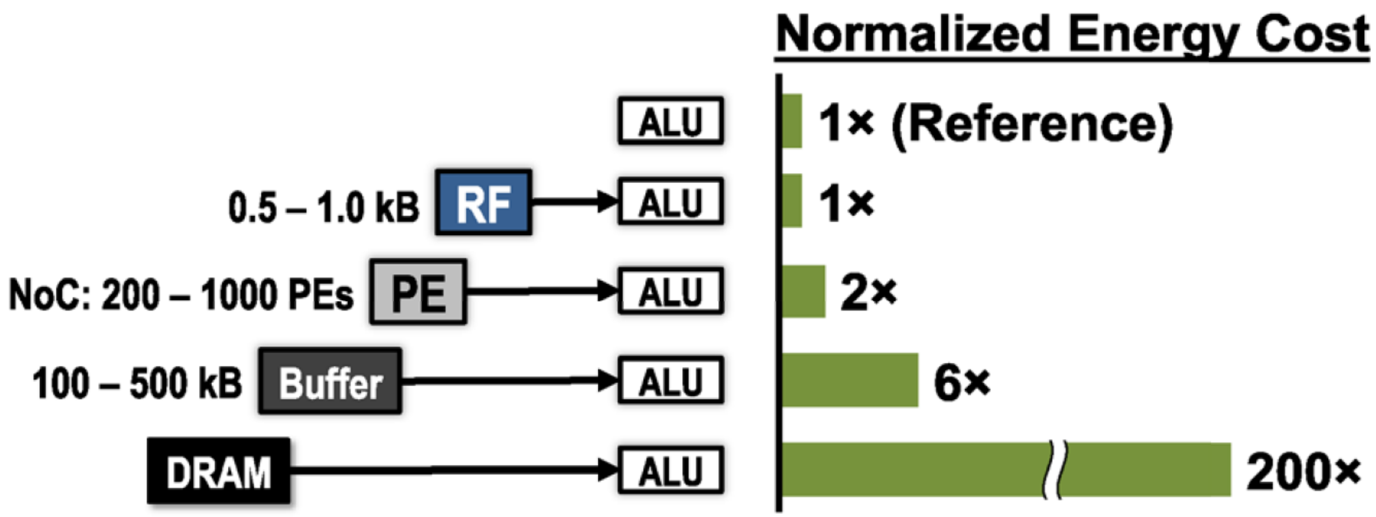
Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FB Add	0.4
32b FB Add	0.9
8b Mult	0.2
32b Mult	3.1
16b FB Mult	1.1
32b FB Mult	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640



- DRAM access **20’000 x** 8-bit addition

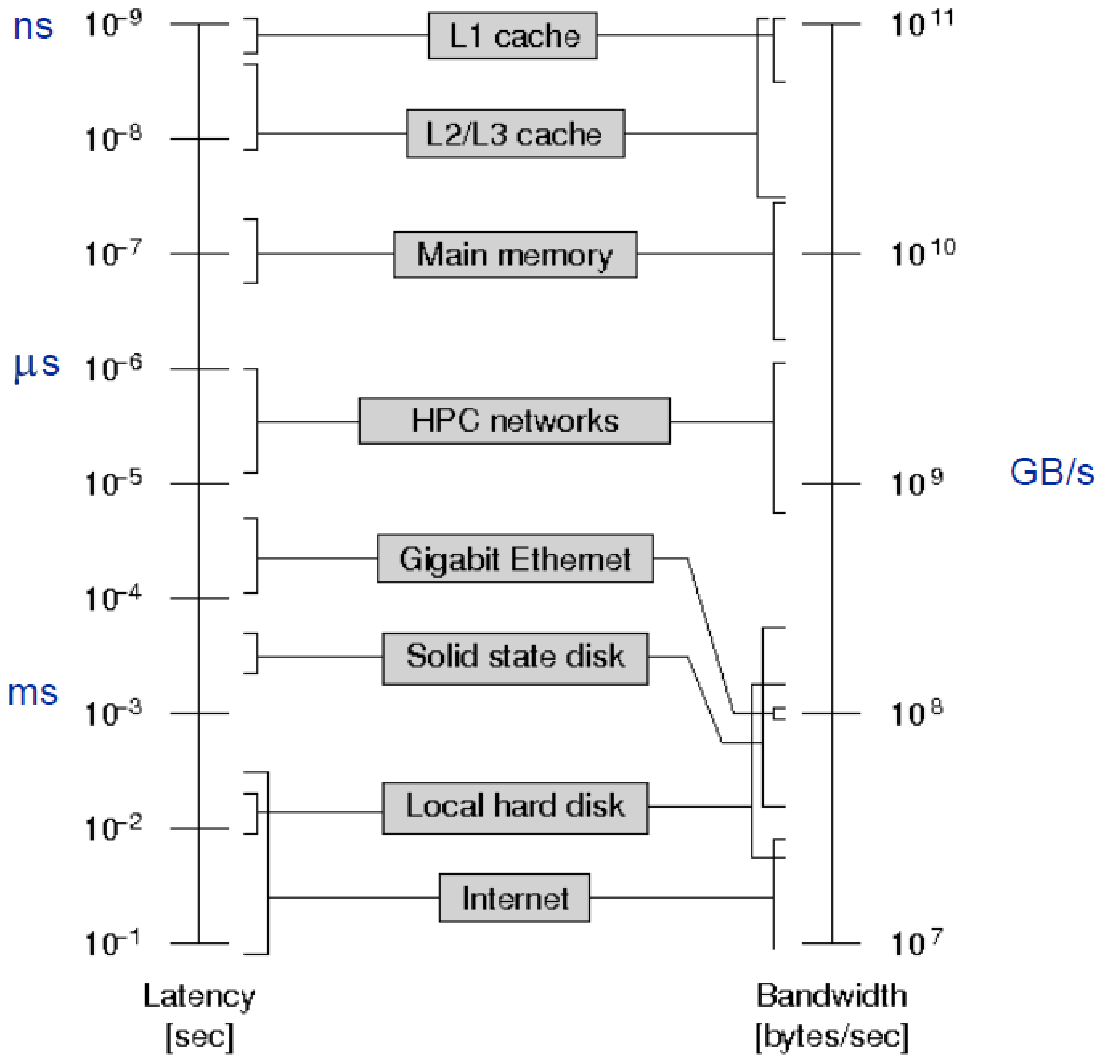
(Hennessy, 2019)

Memory access cost on Spatial Architecture



(Sze et al., 2017)

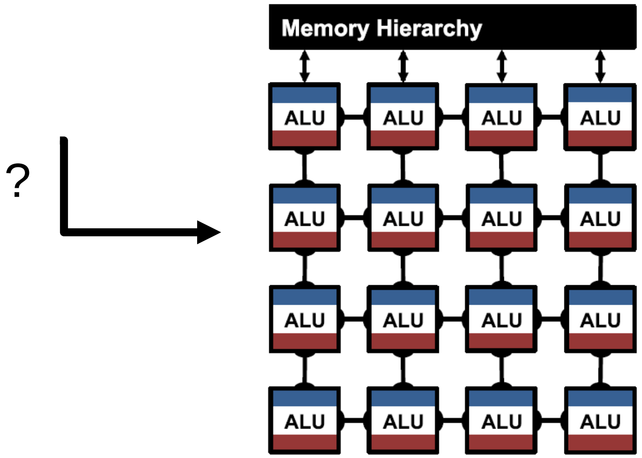
Memory access speed

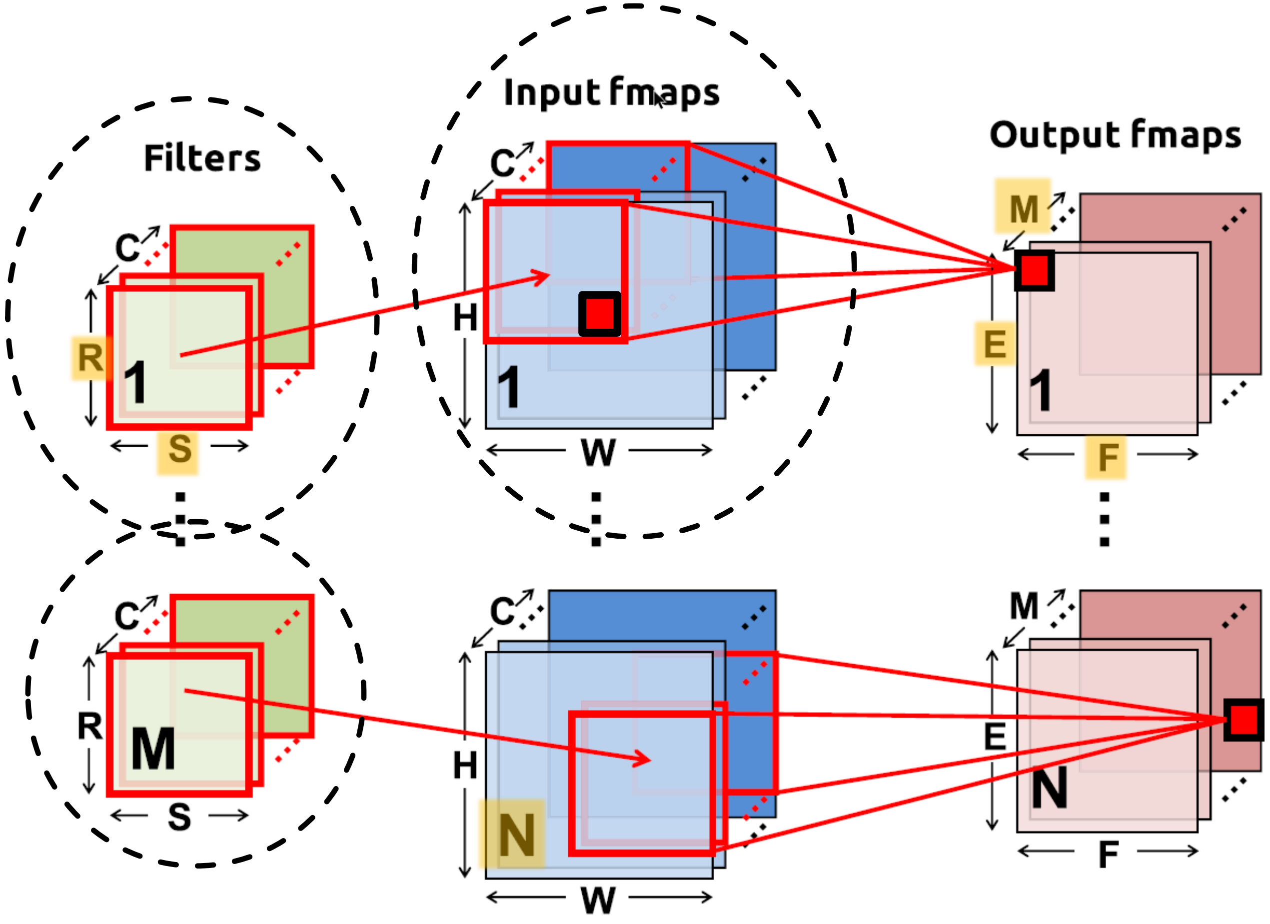


(slide credits: Prof. Koumoutsakos ETH)

Exploiting reuse opportunities

- Convolutional Reuse
- Filter Reuse
- Ifmap Reuse





(Chen et al., 2017)

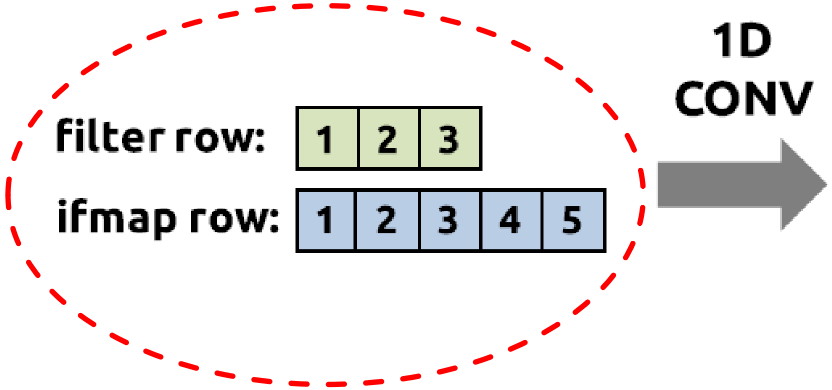
Row Stationary Dataflow

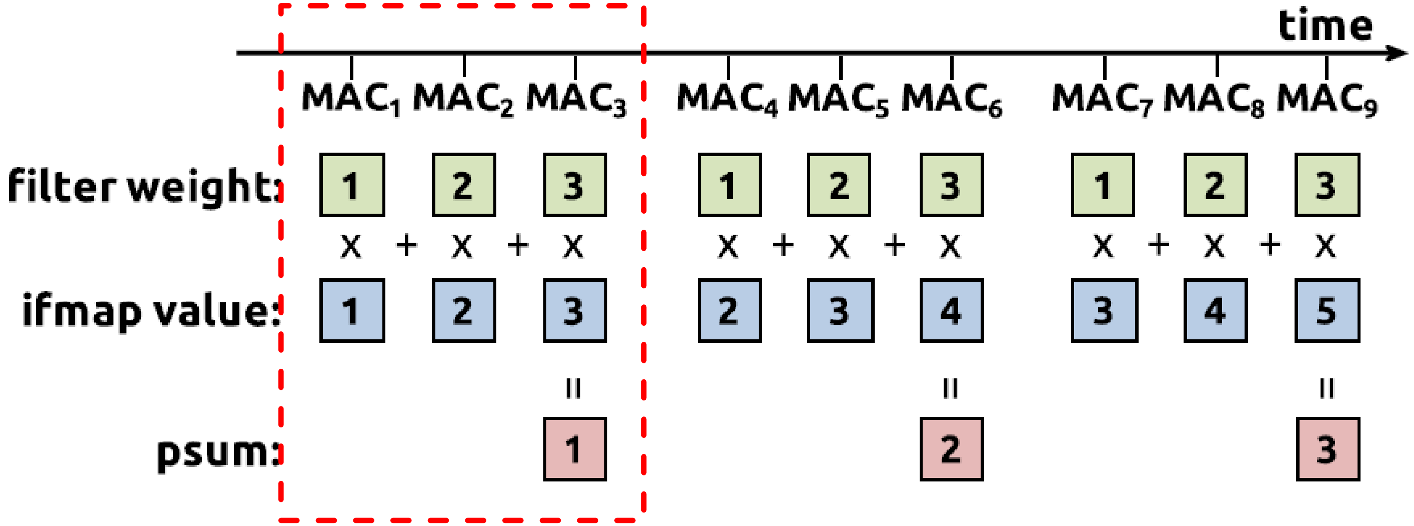
Split into 1D CONV *primitives*:

- 1 row of weights
- 1 row of ifmap

Map each *primitive* on 1 PE:

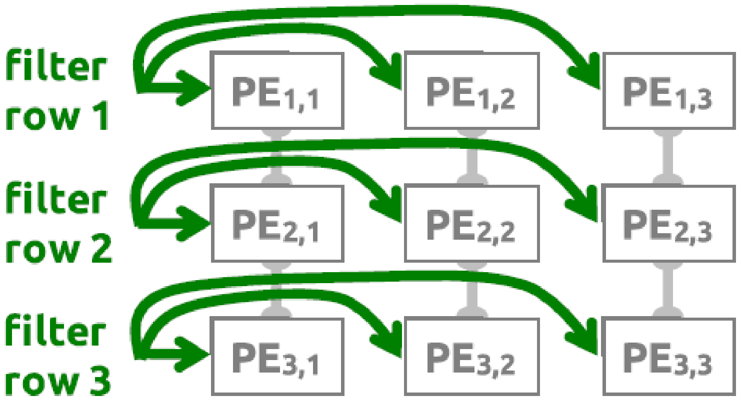
- Row pairs remain *stationary*:
  - psum and weights in local register
- Sliding window

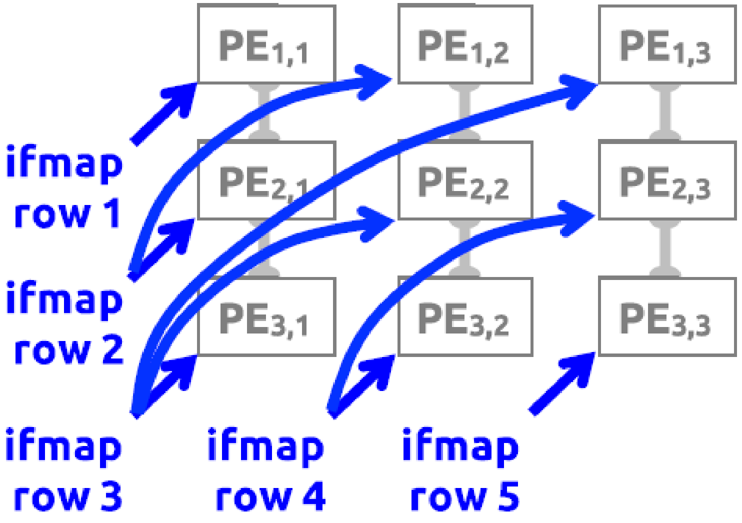


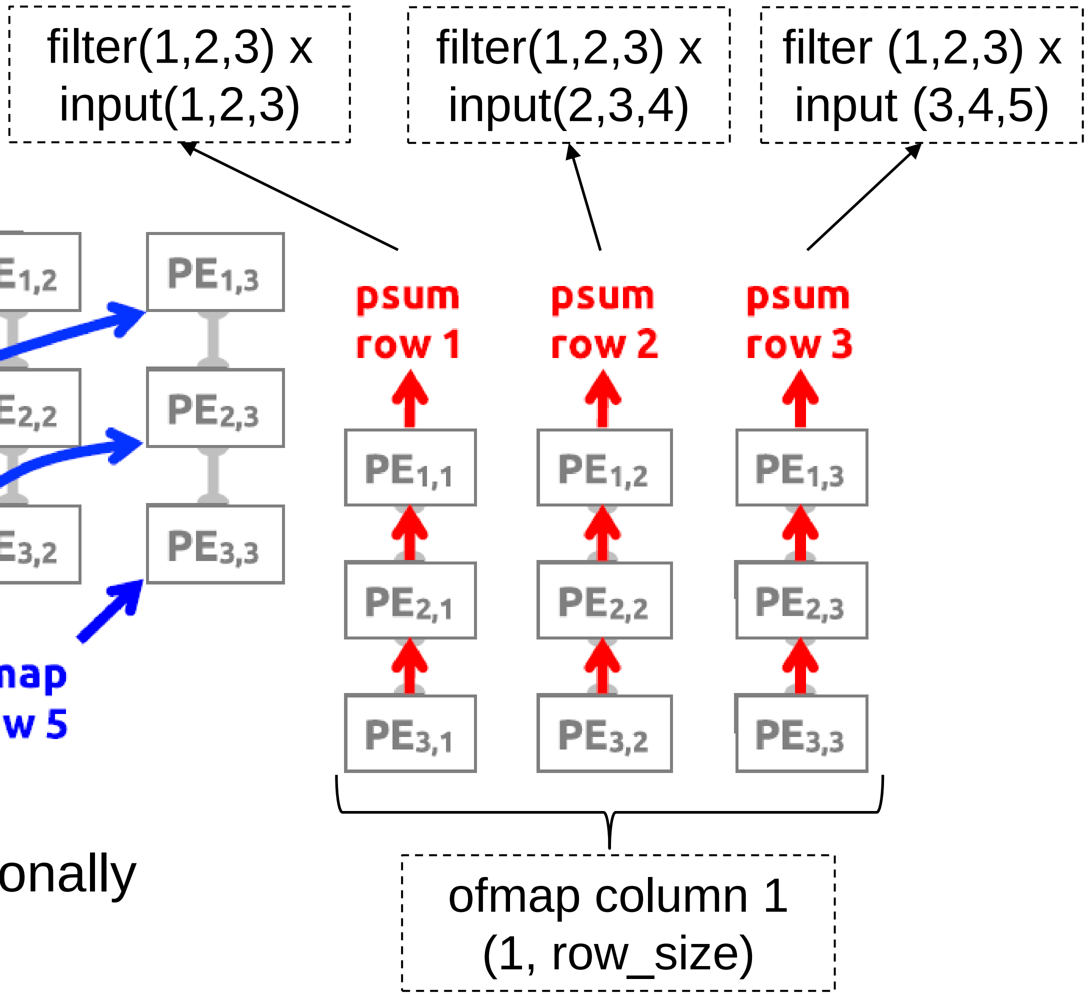


(Chen et al., 2017)

Row Stationary 2D convolution



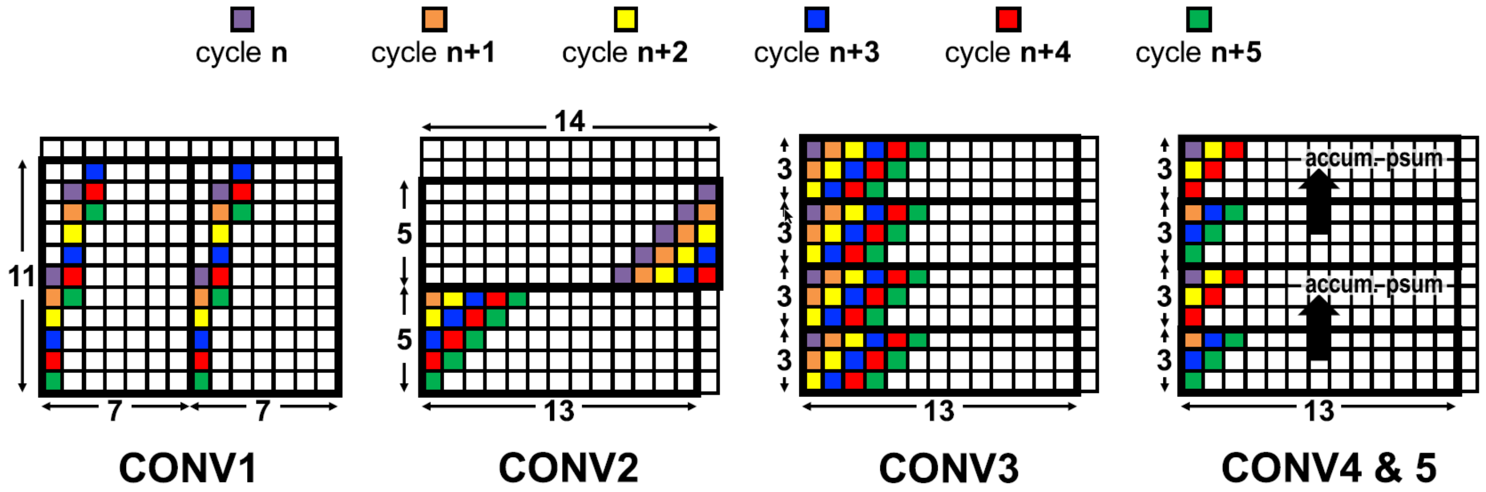


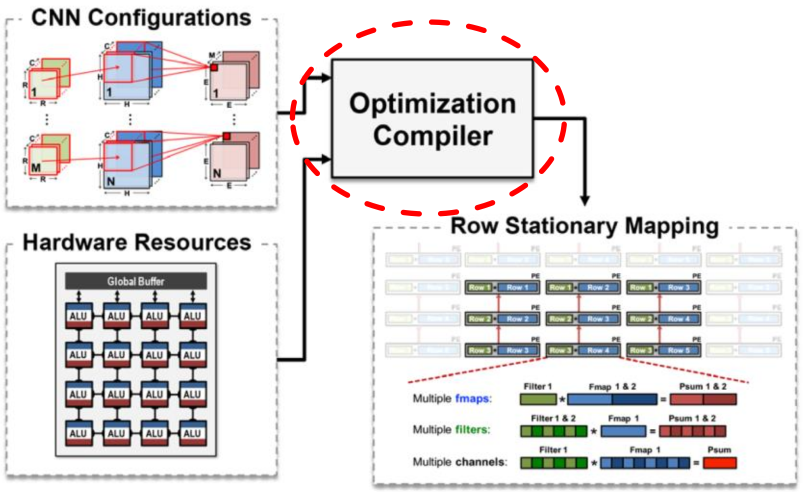


- Filter rows reused horizontally
- Ifmaps reused diagonally
- Psums accumulated vertically

(Chen et al., 2017)

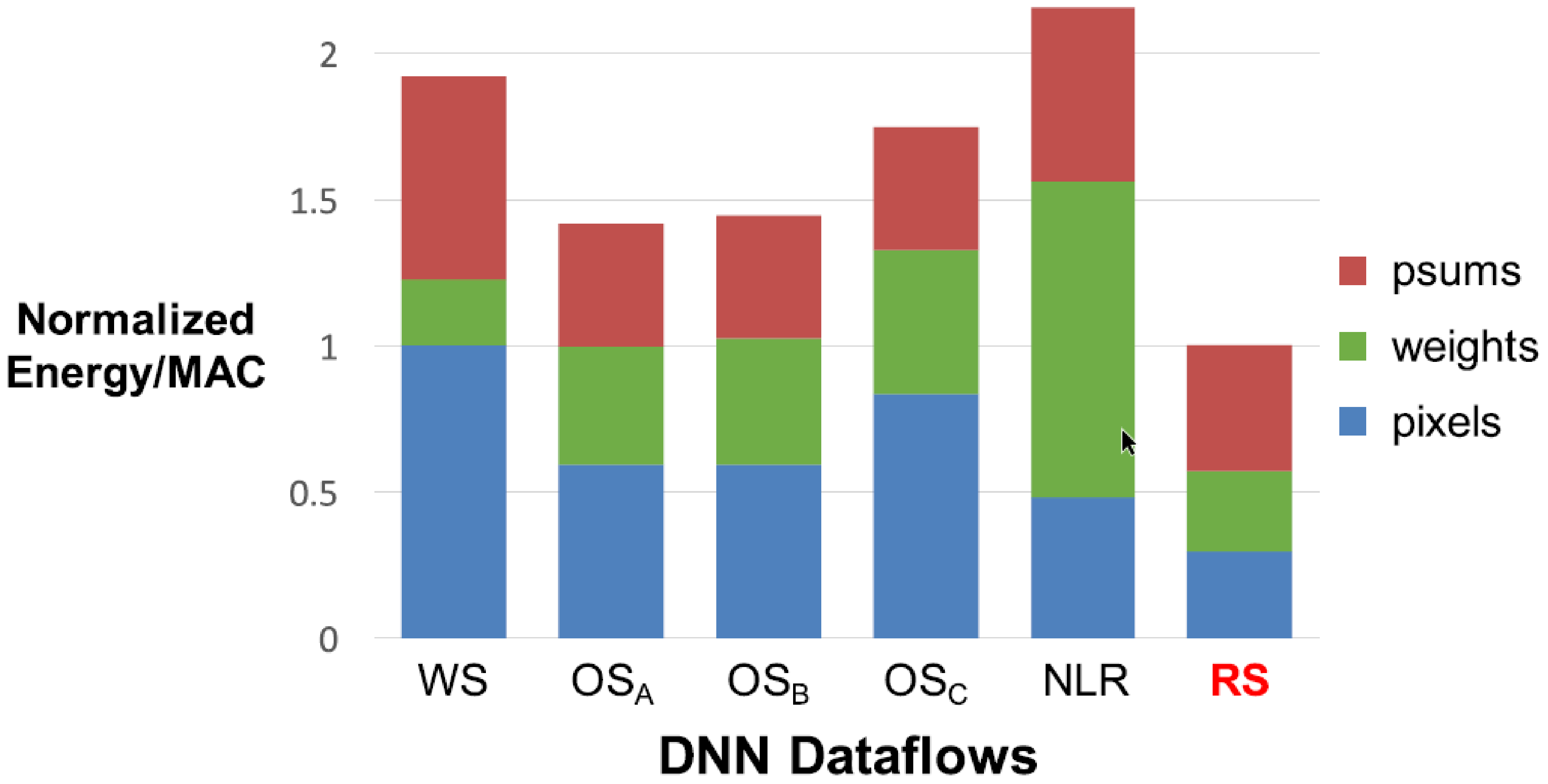
Alex Net example mapping





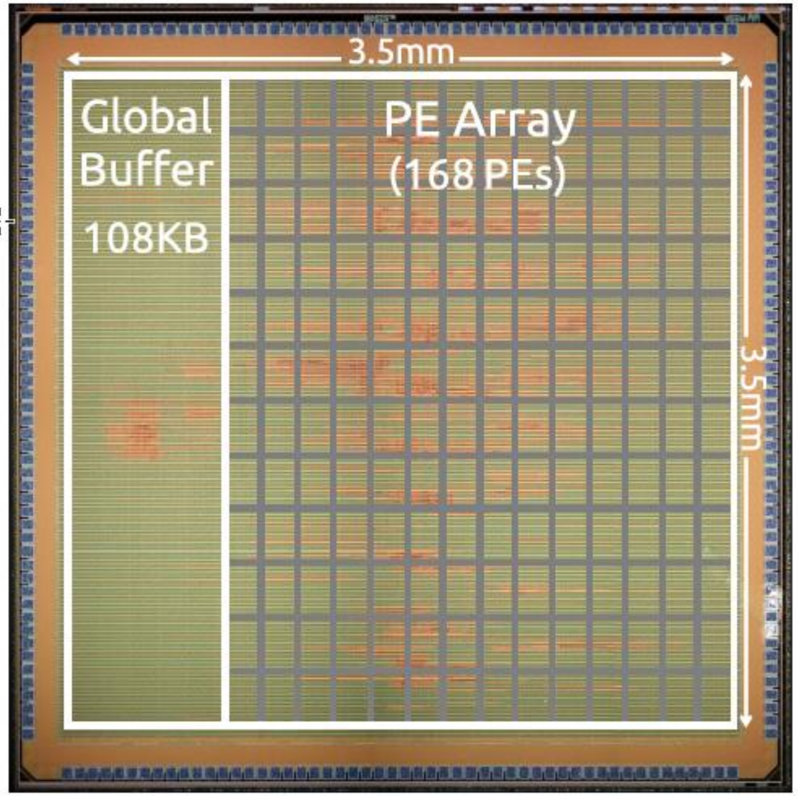
(Chen et al., 2017)

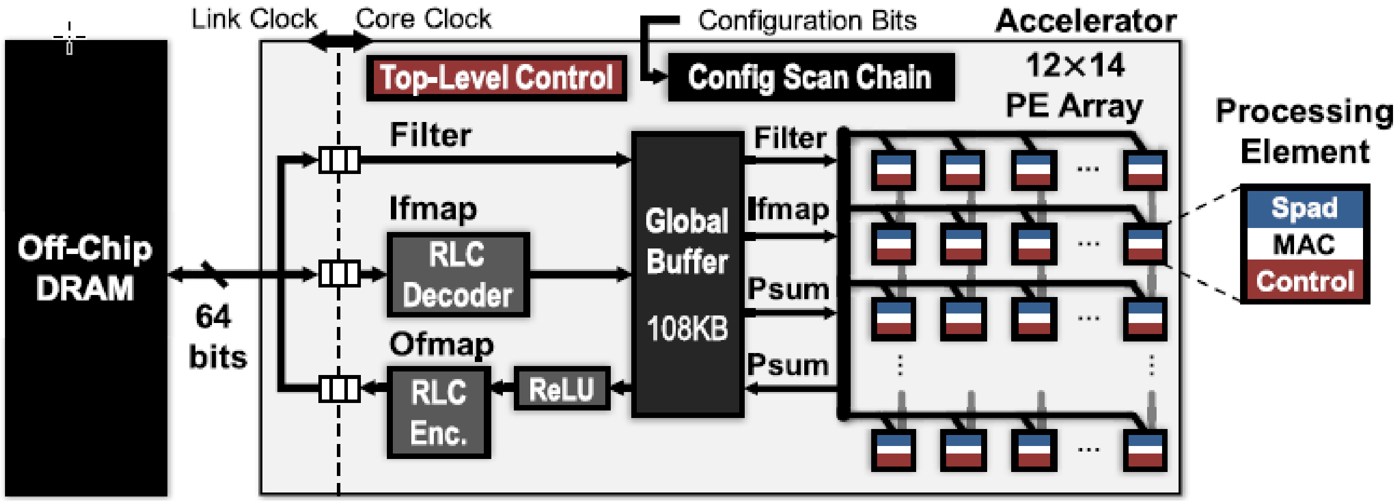
Row Stationary Dataflow



(Sze et al., 2017)

Eyeriss v1

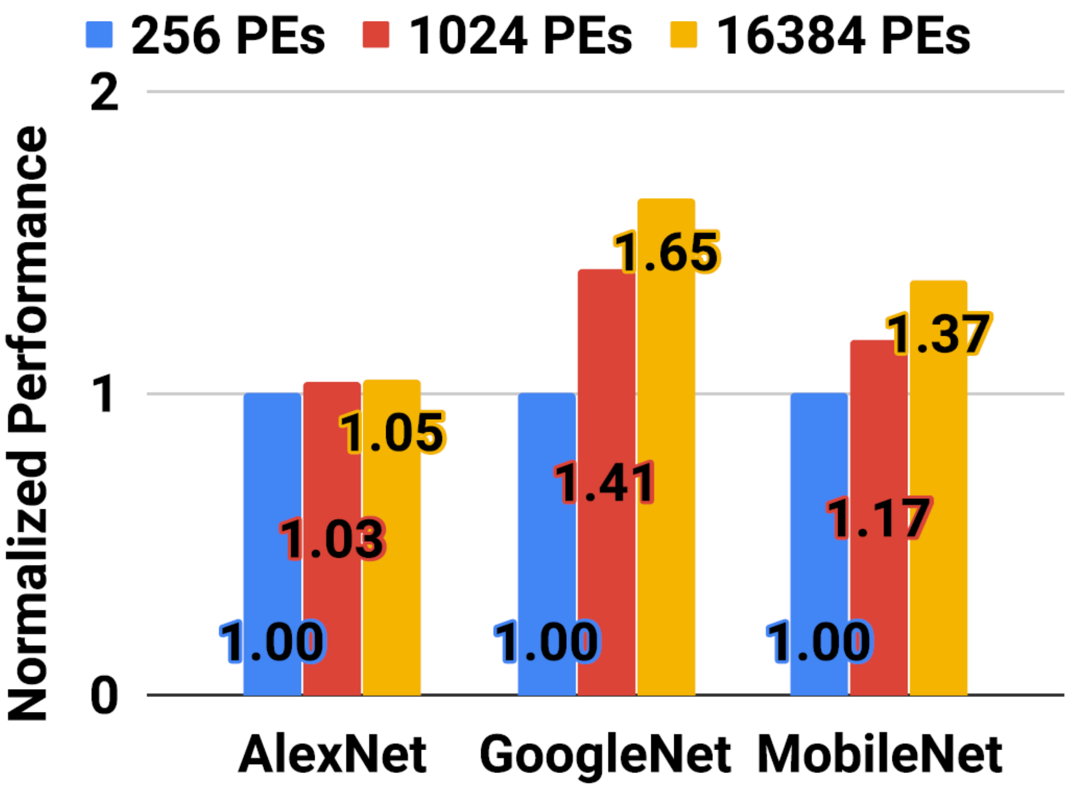




- layer by layer

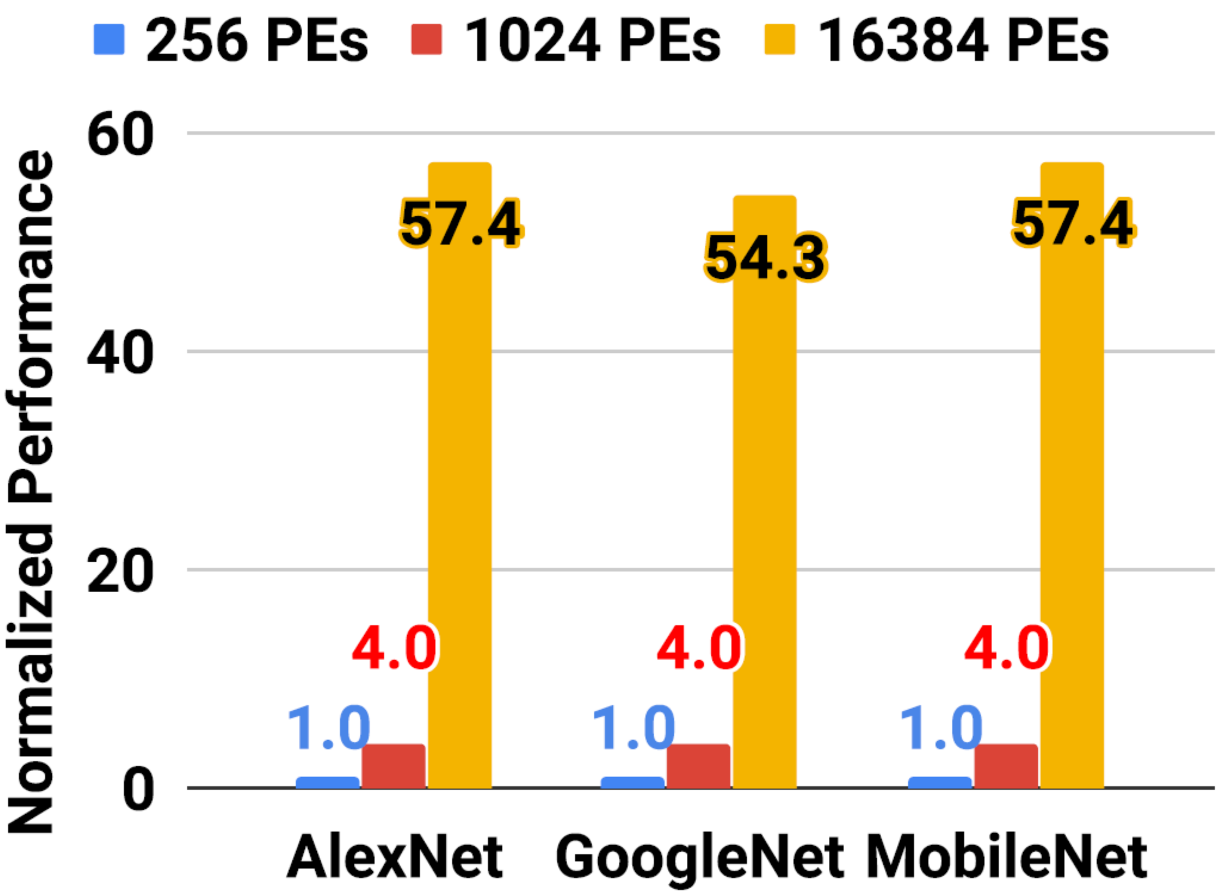
(Chen et al., 2017)

Scalability Eyeriss v1



(Chen et al., 2019)

Scalability Eyeriss v2

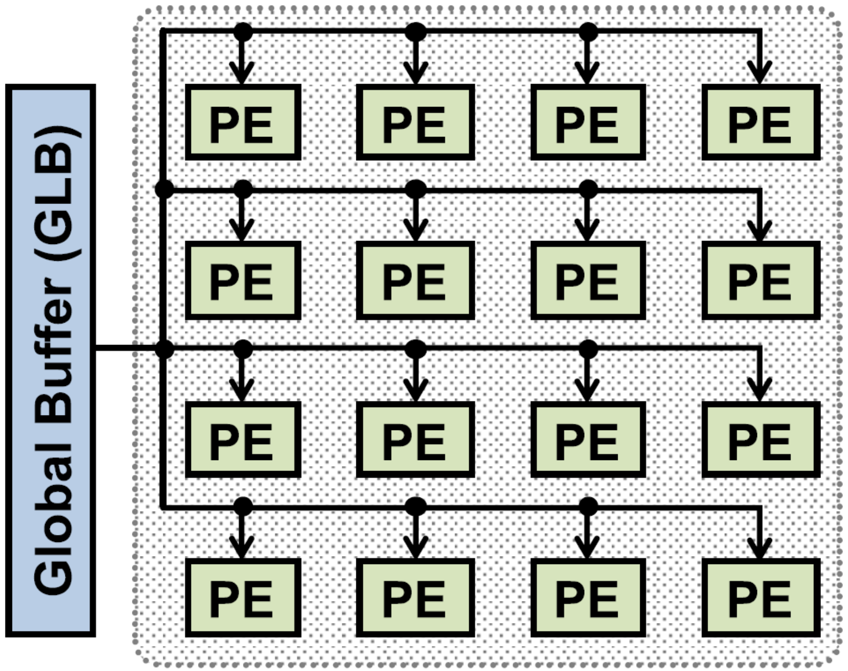


(Chen et al., 2019)

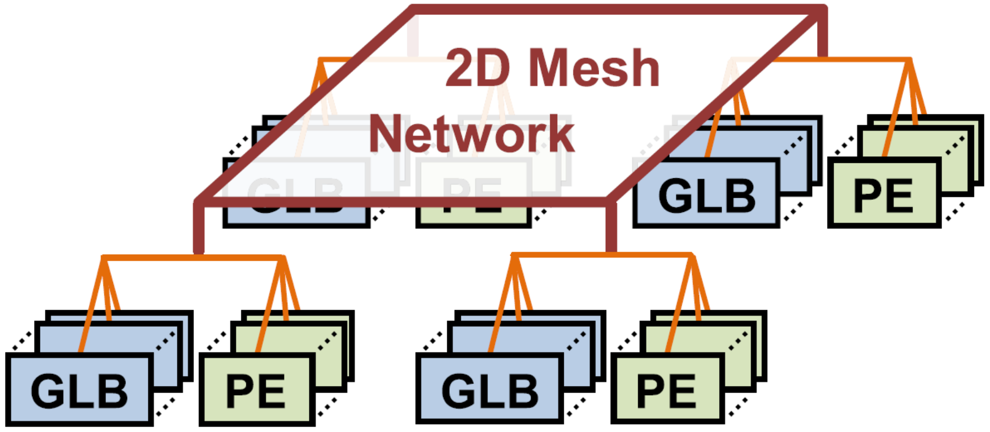
Design Principles Eyeriss v2

Efficiency Latency Flexibility Scalability

Hierarchical Mesh Network



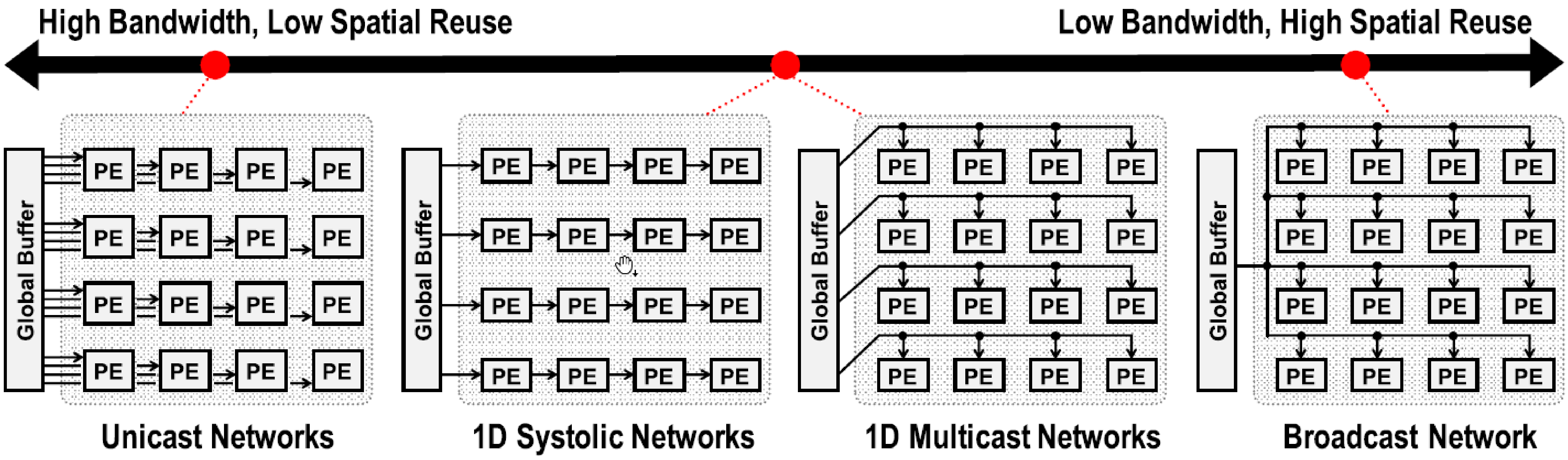
- Flat multicast network



- PEs and GLB grouped into clusters
- Hierarchical structure

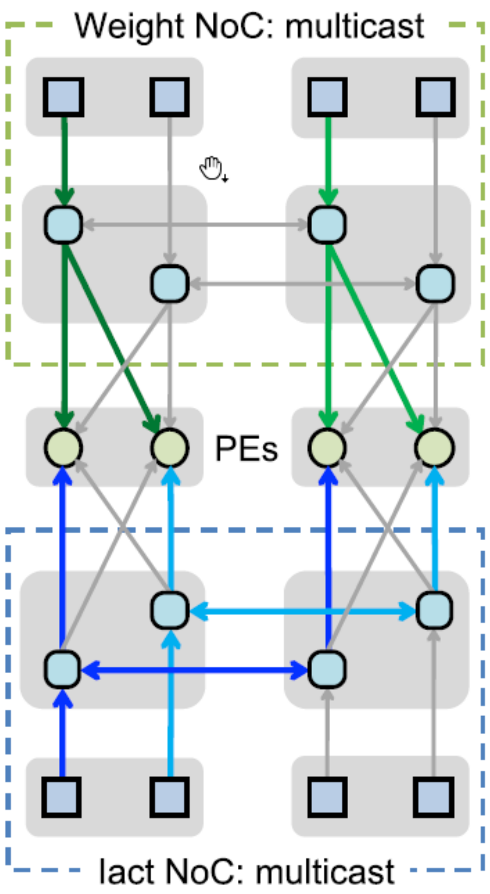
(Chen et al., 2019)

Why use a Mesh?

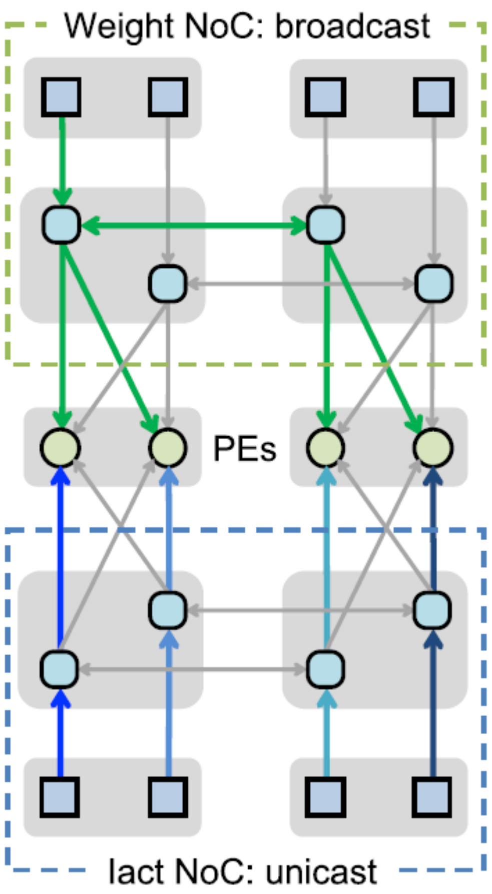


(Chen et al., 2019)

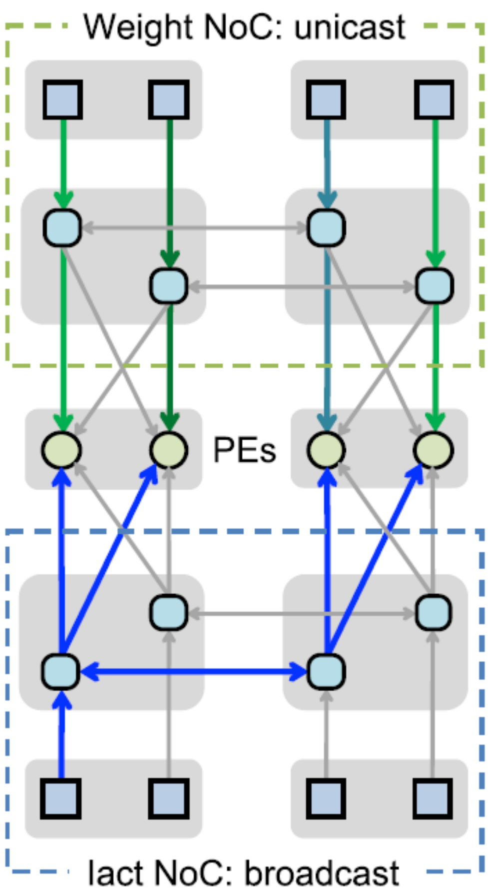
Mesh operation modes



- CONV layer



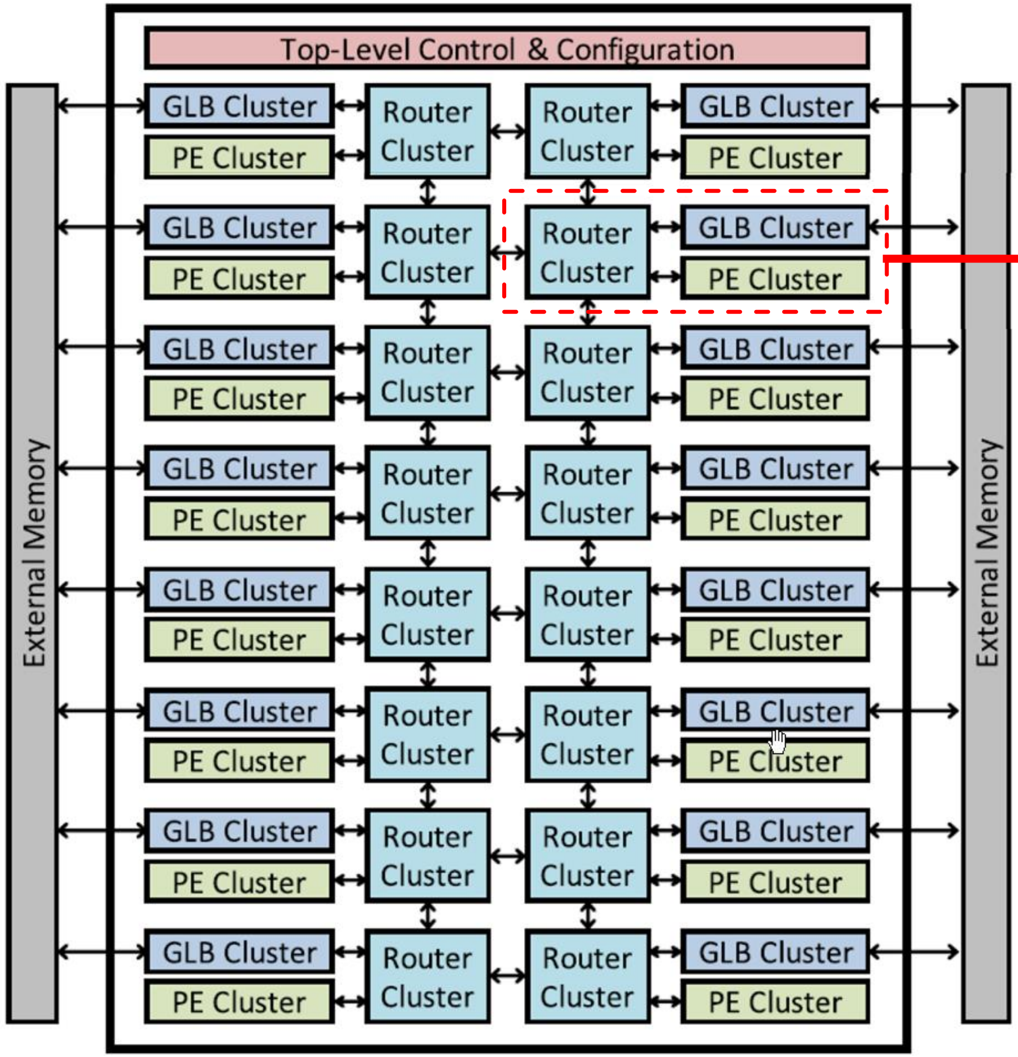
- DW CONV layer

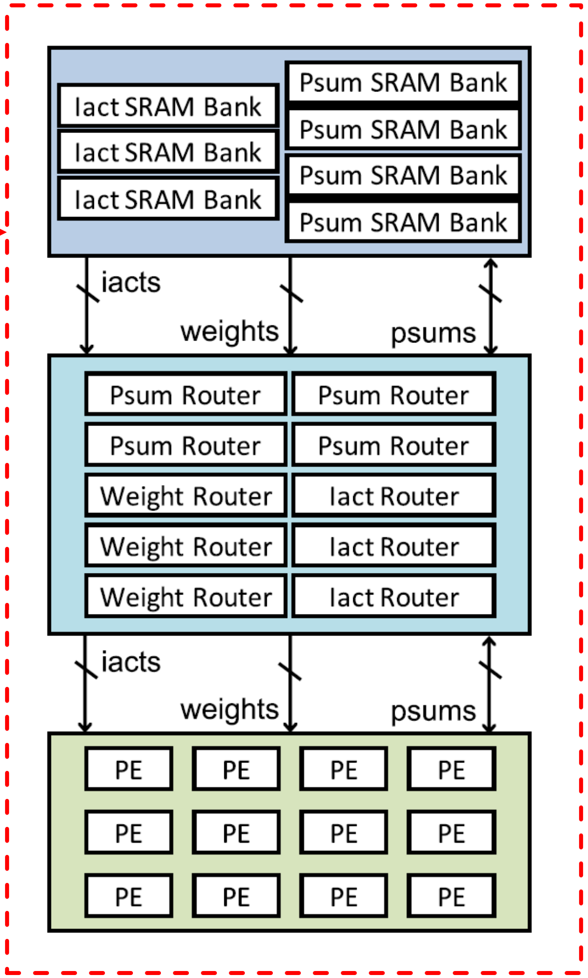


- FC layer

(Chen et al., 2019)

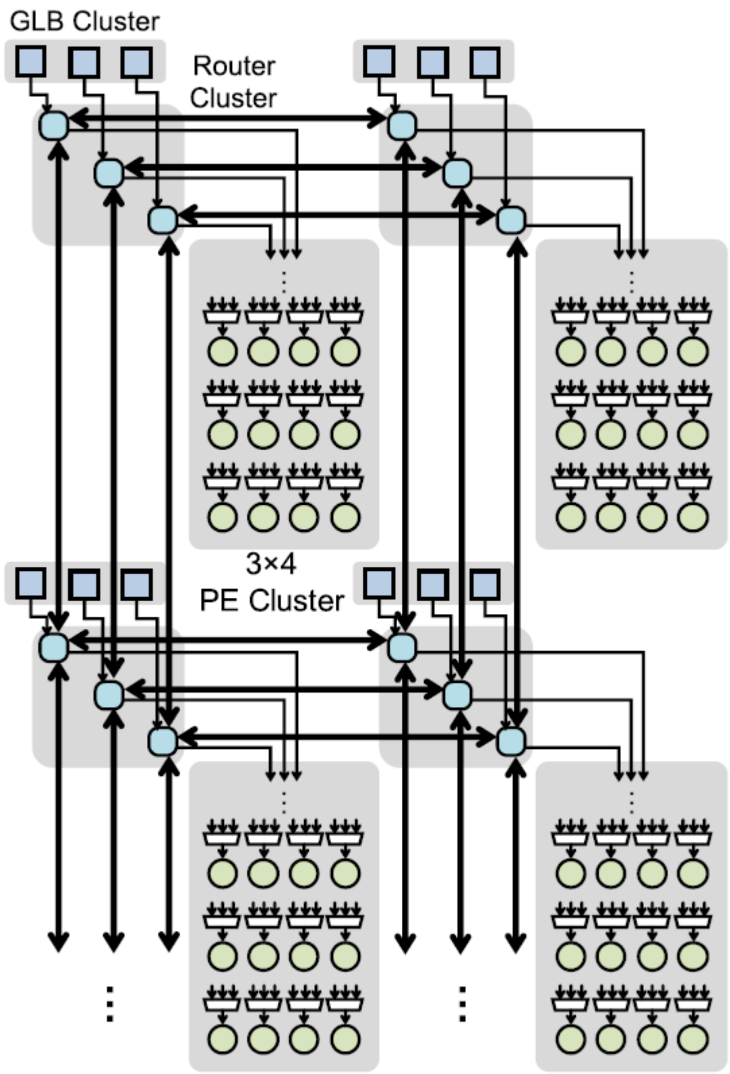
Eyeriss v2 Architecture





(Chen et al., 2019)

Network for input activations



(Chen et al., 2019)

Architecture Hierarchy

Hierarchy	# of Components
Cluster Array	8×2 PE clusters 8×2 GLB clusters 8×2 router clusters
PE cluster	3×4 PEs
GLB cluster	3×1.5 kB SRAM banks for iacts 4×1.875 kB SRAM banks for psums
router cluster	3 iact routers (4 src/dst ports, 24b/port) 3 weight routers (2 src/dst ports, 24b/port) 4 psum routers (3 src/dst ports, 40b/port)

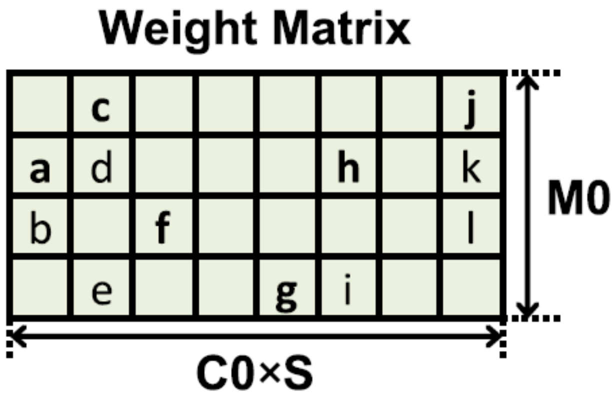
(Chen et al., 2019)

Specification

Technology	TSMC 65nm LP 1P9M
Gate Count (logic only)	2695k (NAND-2)
On-Chip SRAM	246 KB
Number of PEs	192
Global Buffer	192 KB (SRAM)
Scratch Pads (per PE)	weight addr: 14B (Reg) weight data: 288B (SRAM) iact addr: 4.5B (Reg) iact data: 24B (Reg) psum: 80B (Reg)
Clock Rate	200 MHz
Peak Throughput	153.6 GOPS
Arithmetic Precision	weights & iacts: 8b fixed-point psums: 20b fixed-point

(Chen et al., 2019)

Exploit sparsity



CSC Compressed Data:

data vector: {**a**, b, **c**, d, e, **f**, **g**, **h**, i, **j**, k, l}

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

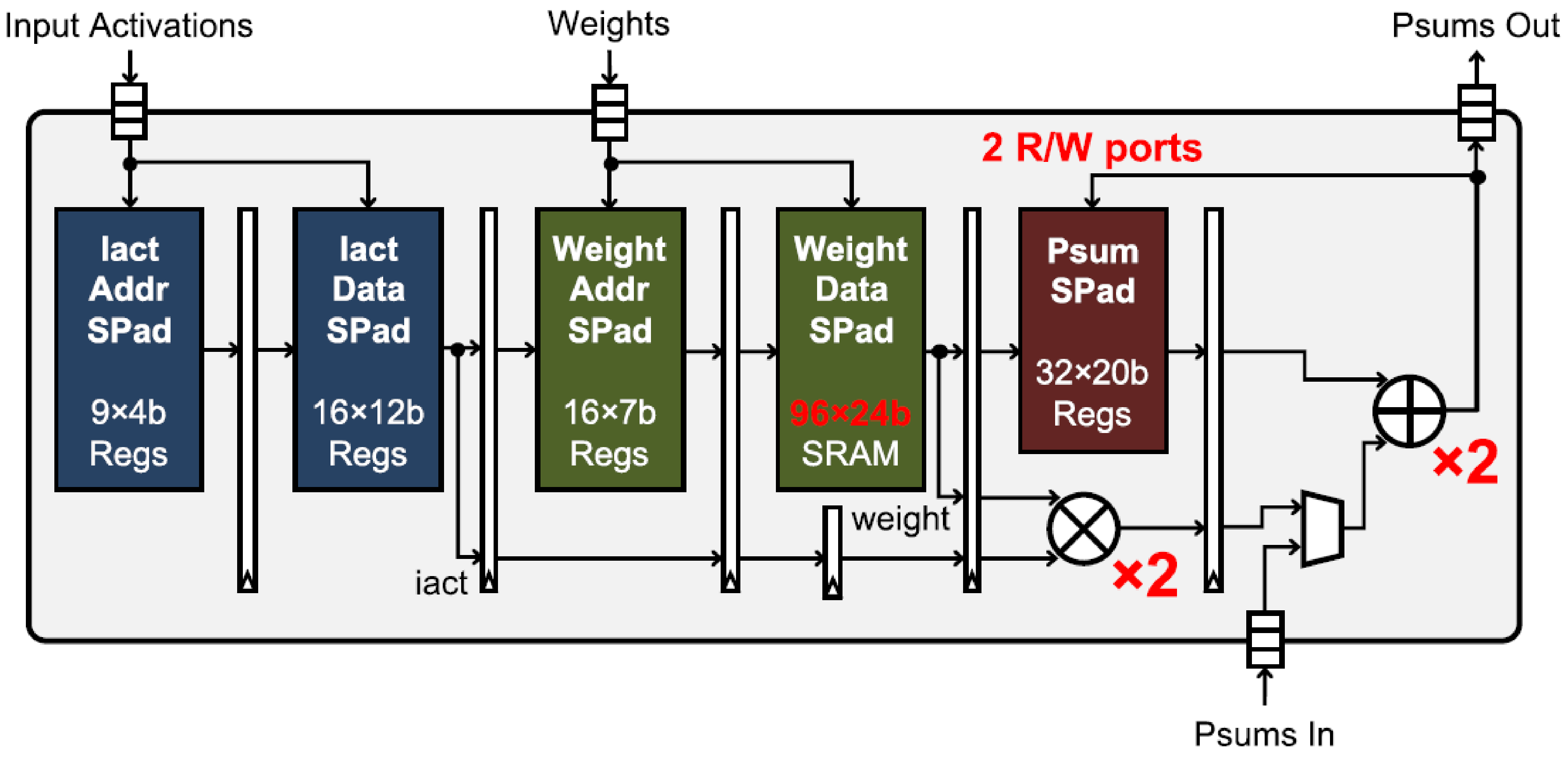
address vector: {0, 2, 5, 6, 6, 7, 9, 9, 12}

- *process* in CSC format

- *Latency* gain

(Chen et al., 2019)

Exploit sparsity



(Chen et al., 2019)

Results for Eyeriss v2

DNN	ImageNet Accuracy[1]	Nominal Num. of MACs	Inference/sec	Inference/J	GOPS/W	DRAM Acc. (MB)	PE Utilization[2]
AlexNet	80.43%	724.4M	102.1	174.8	253.2	71.9	100%
sparse AlexNet	79.56%	724.4M	278.7	664.6	962.9	22.3	100%
MobileNet	79.37%	49.2M	1282.1	1969.8	193.7	4.1	91.5%
sparse MobileNet	79.68%	49.2M	1470.6	2560.3	251.7	3.9	91.5%

[1] top-5 accuracy for the image classification task.
[2] measured in terms of number of utilized MAC datapaths; each PE has 2 MAC datapaths.

(Chen et al., 2019)

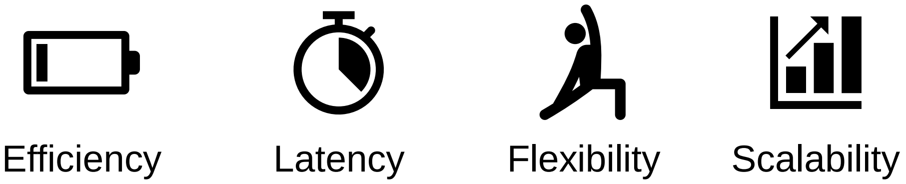
Comparison

	Eyeriss [33]	ENVISION [15]	Thinker [16]	UNPU [17]	This Work	
Technology	65nm	28nm	65nm	65nm	65nm	
Area	1176k gates (NAND-2)	1950k gates (NAND-2)	2950k gates (NAND-2)	4.0mm×4.0mm (Die Area)	2695k gates (NAND-2)	
On-chip SRAM (kB)	181.5	144	348	256	246	
Max Core Frequency	200 MHz	200 MHz	200 MHz	200 MHz	200 MHz	
Bit Precision	16b	4b/8b/16b	8b/16b	1b-16b	8b	
Num. of MACs	168 (16b)	512 (8b)	1024 (8b)	13824 (bit-serial)	384 (8b)	
DNN Model	AlexNet	AlexNet	AlexNet	AlexNet	sparse AlexNet	sparse MobileNet
Batch Size	4	N/A	15	N/A	1	1
Core Frequency (MHz)	200	200	200	200	200	200
Bit Precision	16b	N/A	adaptive	8b	8b	8b
Inference/sec (CONV only)	34.7	47	-	346	342.4	-
Inference/sec (Overall)	-	-	254.3	-	278.7	1470.6
Inference/J (CONV only)	124.8	1068.2	-	1097.5	743.4	-
Inference/J (Overall)	-	-	876.6	-	664.6	2560.3

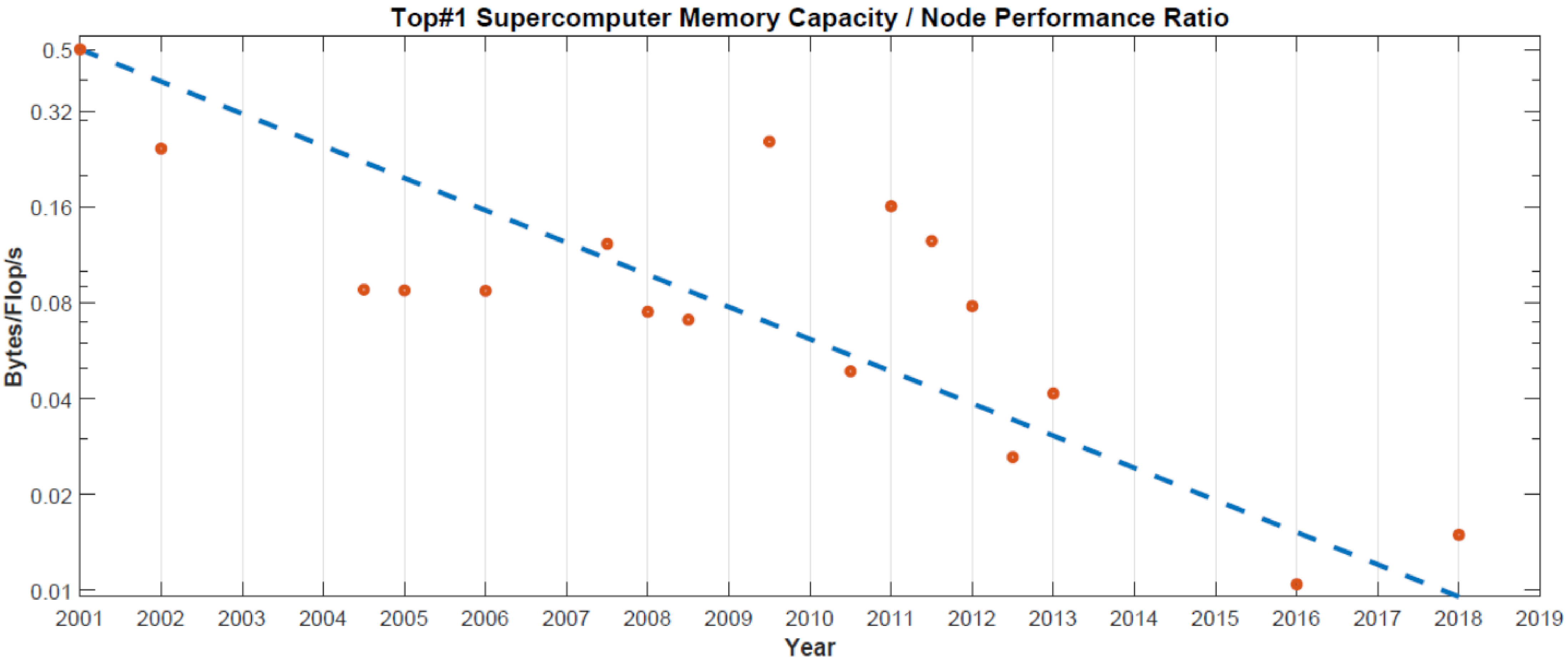
(Chen et al., 2019)

Conclusion

- > 10x improvement over v1
- processing sparse weights and iacts *in compressed domain*
- flexibility from high bandwidth to high data reuse, for filter shape variety
- extend with cache



Final comment



(slide credits: Dr. Sergio Martin ETH)

Final comment

References and image credits

1. Hennessy, J. L. *Computer architecture: a quantitative approach*. (Morgan Kaufmann Publishers, 2019).
2. Sze, V., Chen, Y.-H., Yang, T.-J. & Emer, J. S. Efficient Processing of Deep Neural Networks: A Tutorial and Survey. *Proc. IEEE* **105**, 2295–2329 (2017).
3. Chen, Y.-H., Yang, T.-J., Emer, J. S. & Sze, V. Eyeriss v2: A Flexible Accelerator for Emerging Deep Neural Networks on Mobile Devices. *IEEE J. Emerg. Sel. Topics Circuits Syst.* **9**, 292–308 (2019).
4. Chen, Y.-H., Krishna, T., Emer, J. S. & Sze, V. Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks. *IEEE J. Solid-State Circuits* **52**, 127–138 (2017).
5. Lin, S.-C. *et al.* The Architectural Implications of Autonomous Driving: Constraints and Acceleration. in *Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems* 751–766 (ACM, 2018). doi:10.1145/3173162.3173191.

Images: MIT EEMS Group (slide 2), Audi (slide 4), Nvidia (slide 5), WabisabiLearning (slide 6), iStock.com/VictoriaBar (slide 31)

Online video talks:
- slideslive.com
- youtu.be/WbLQqPw_n88

Q & A

Additional slides for interested readers

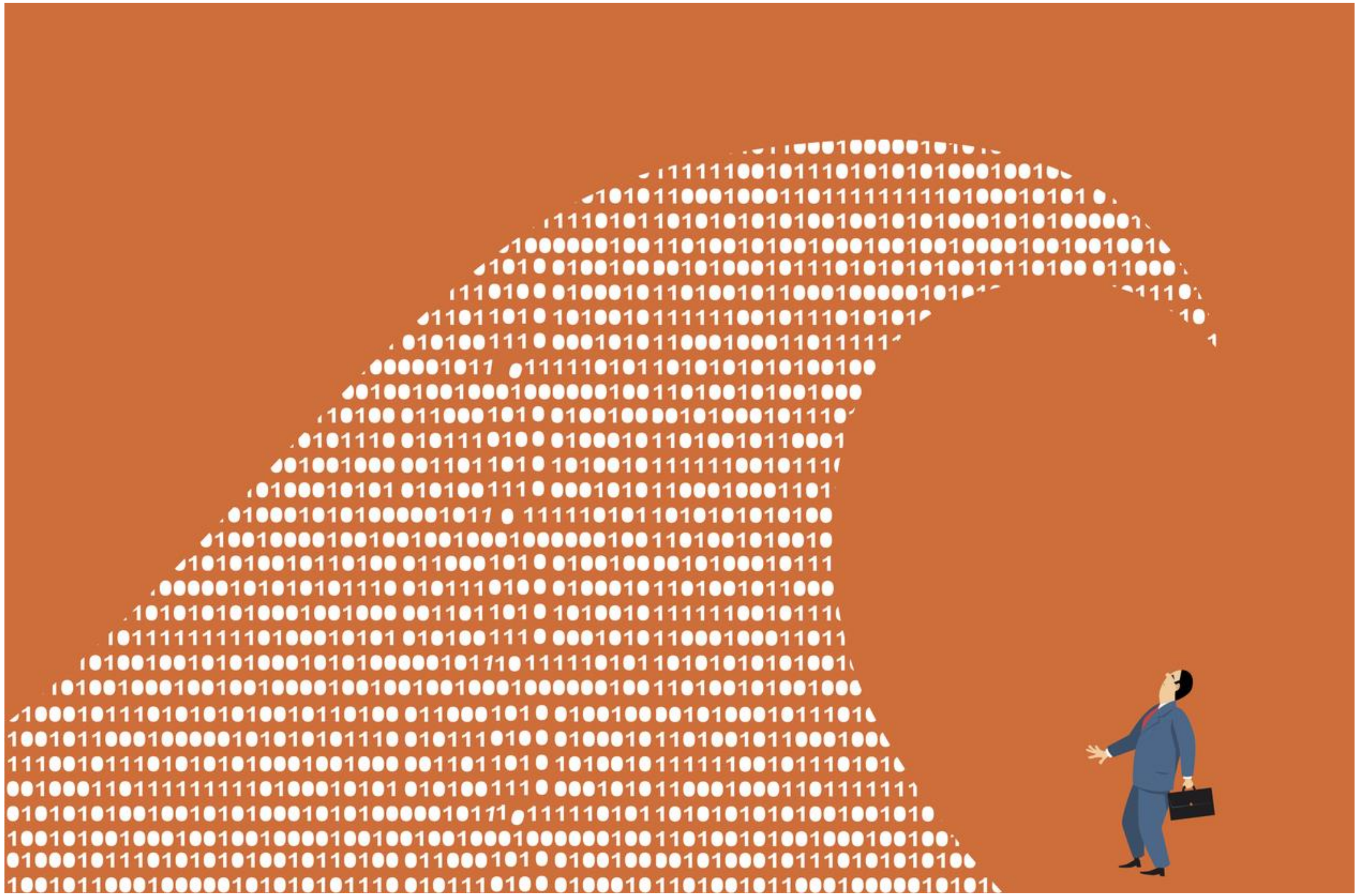
Energy Efficiency

Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000

Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

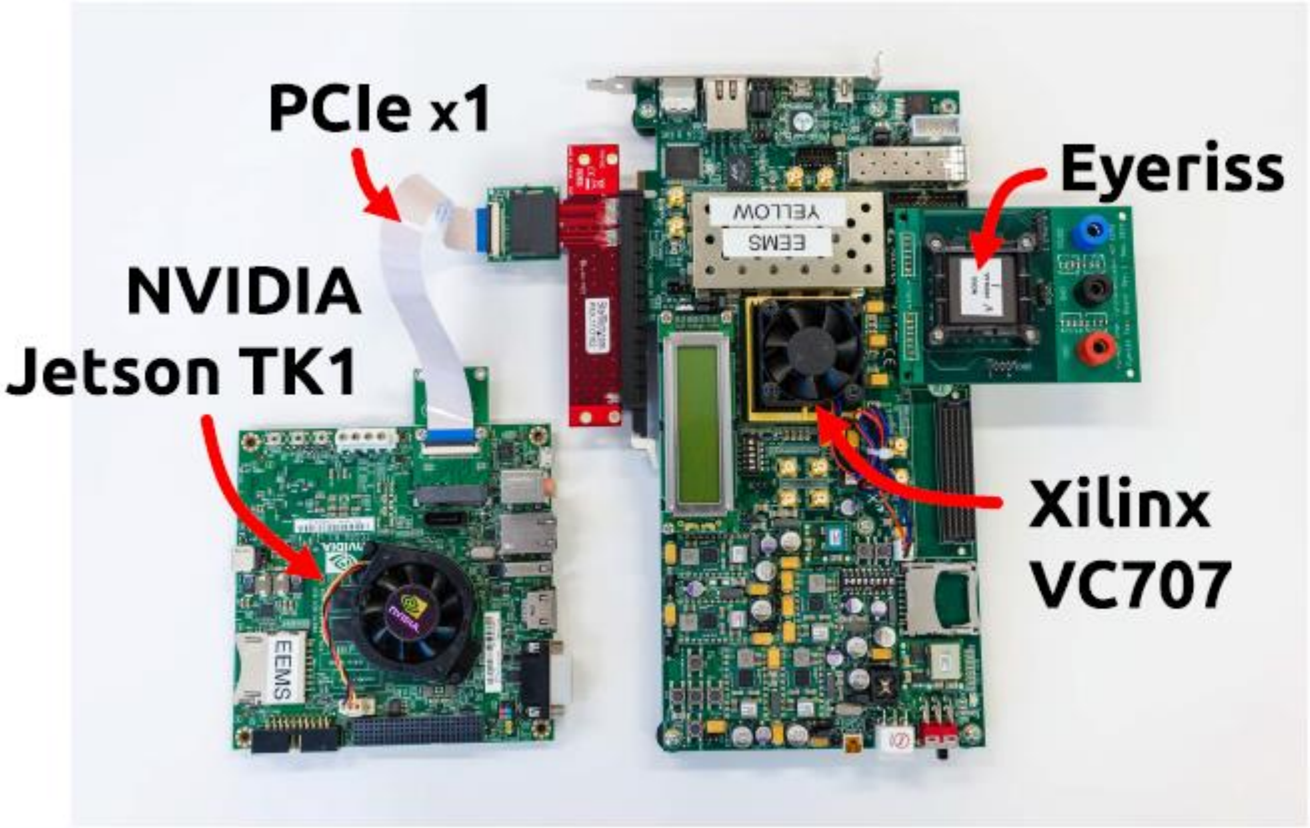
(Strubell et al., 2019)

Motivation: Flexibility, Software



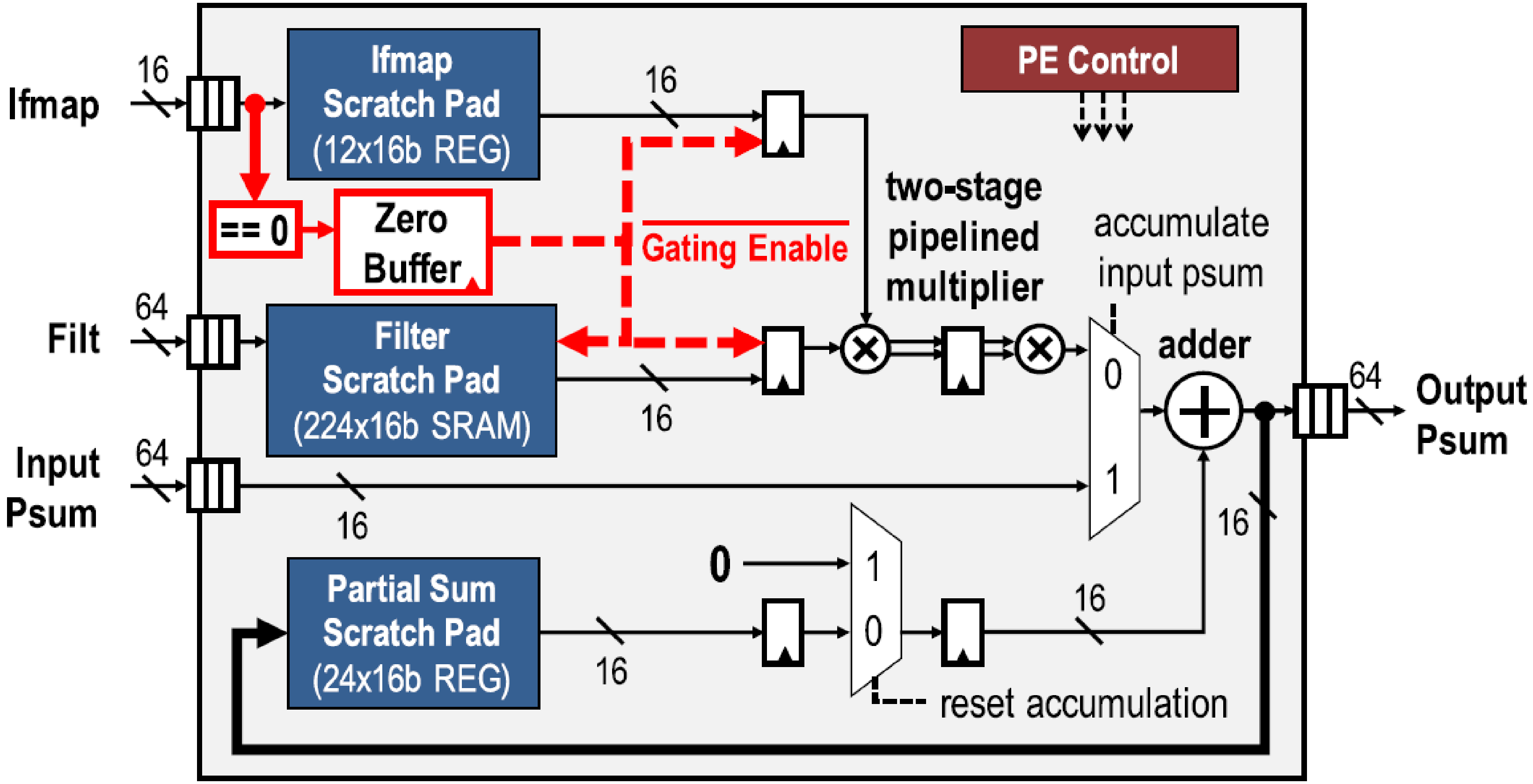
Eyeriss v1 Implementation

- Customized Coffee Framework run on NVIDIA development board
- Xlinix serves as PCI controller



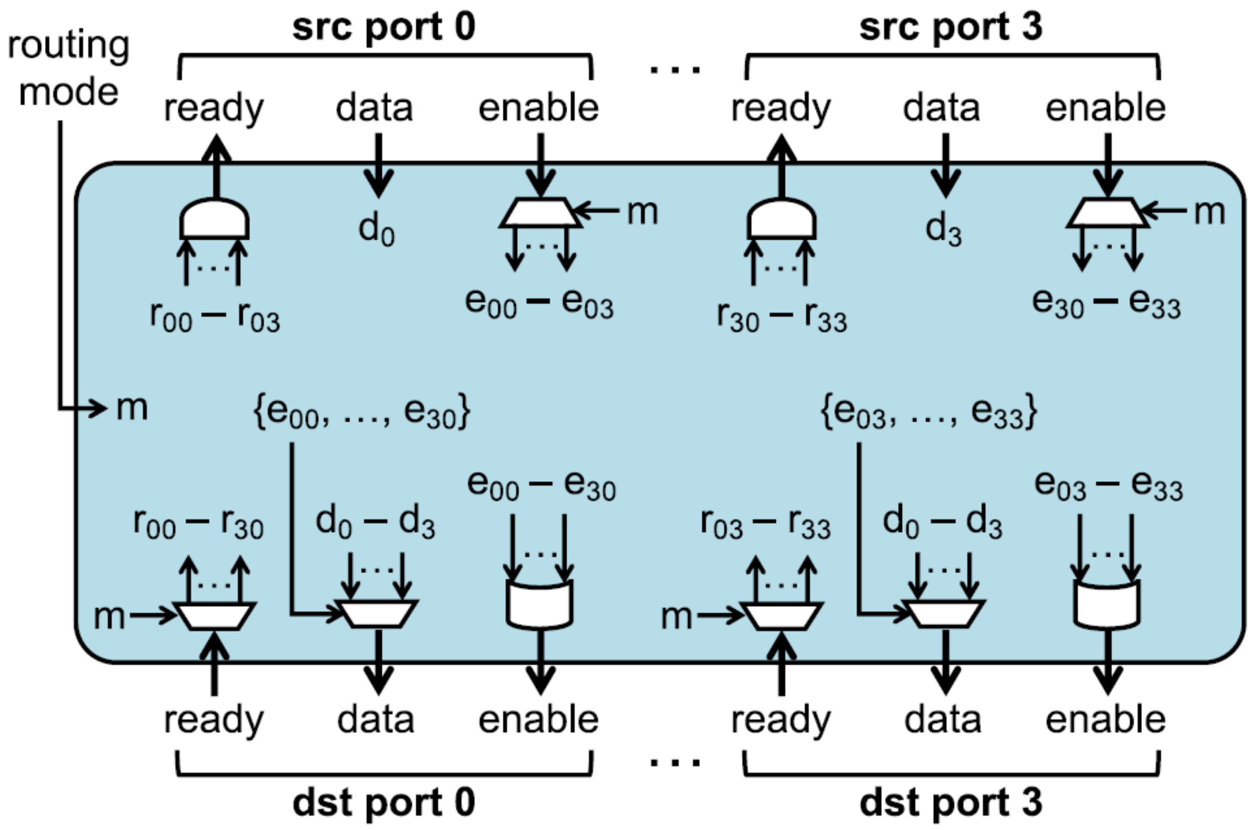
(Chen et al., 2017)

Eyeriss v1 PE architecture

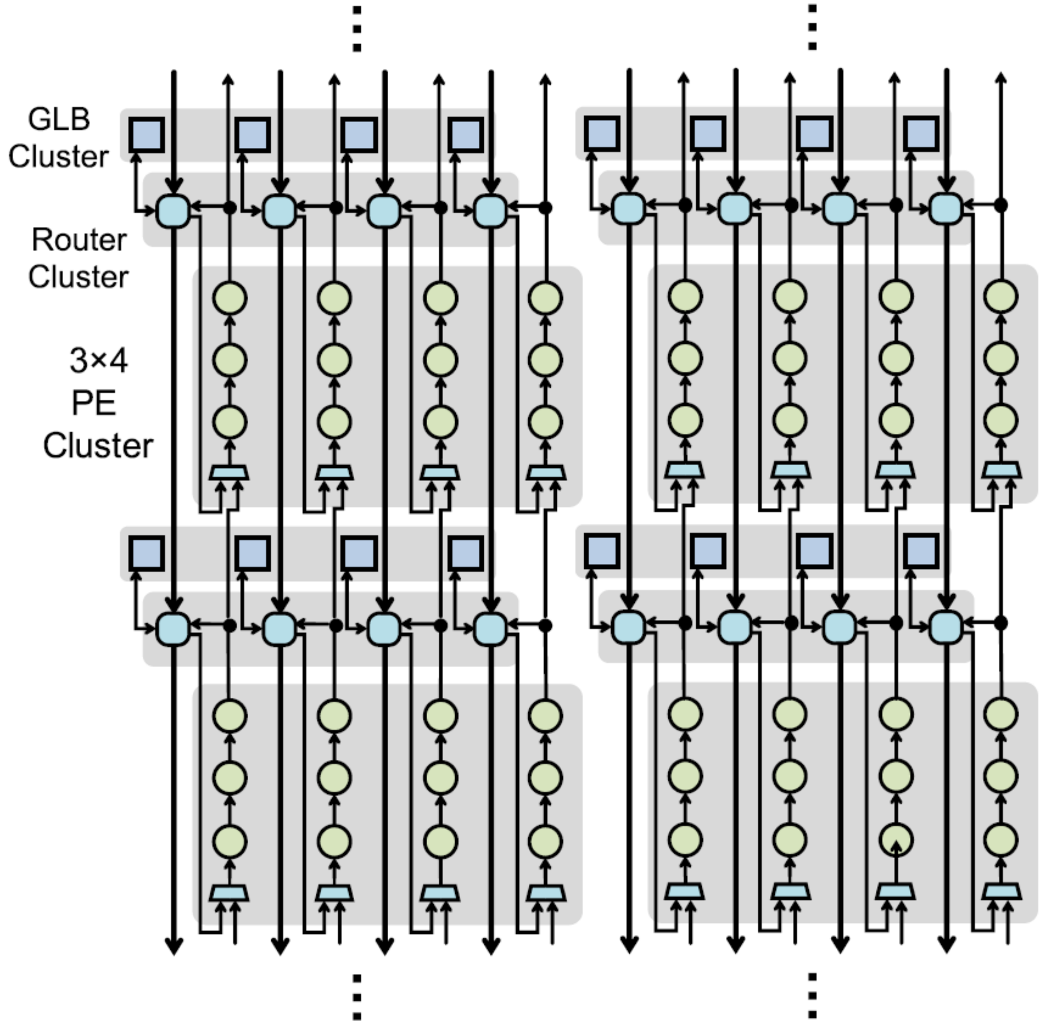


(Chen et al., 2017)

Router implementation details

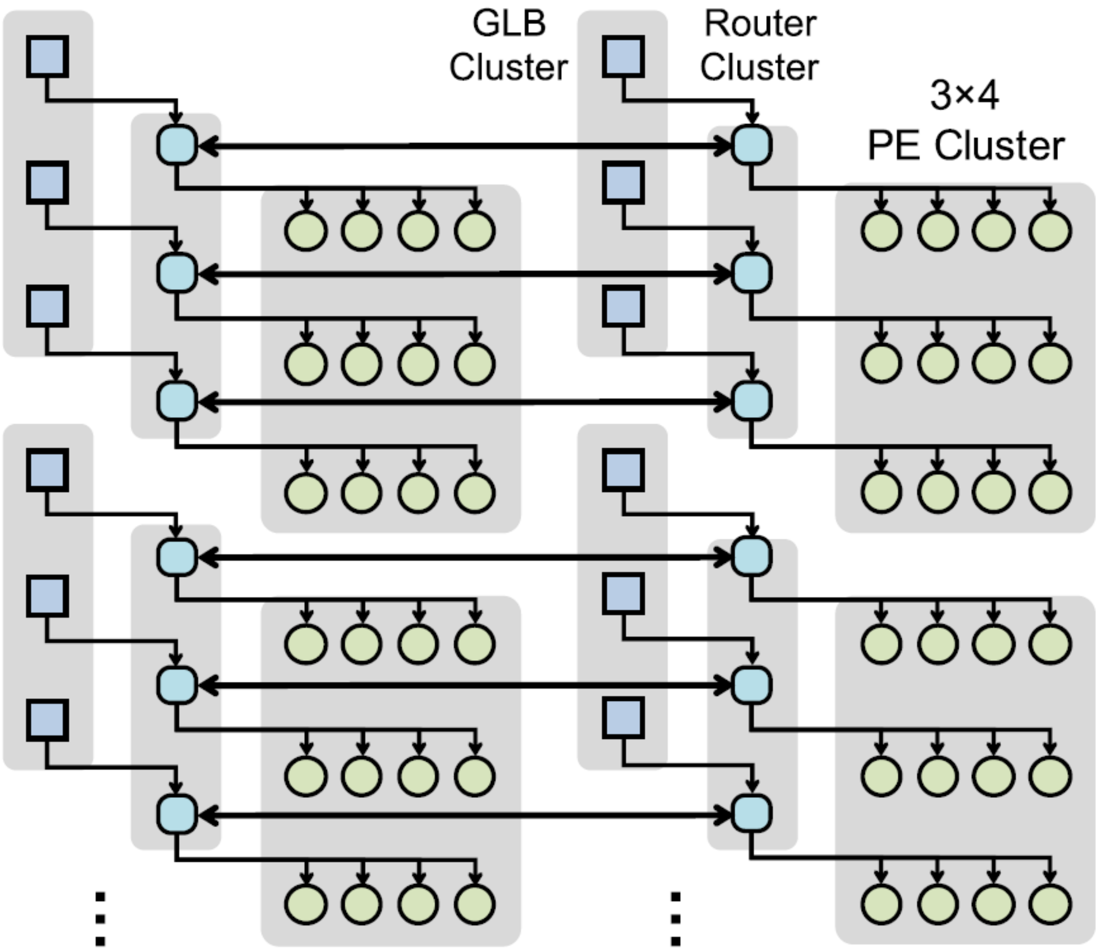


Network for psums



(Chen et al., 2019)

Network for weights



(Chen et al., 2019)