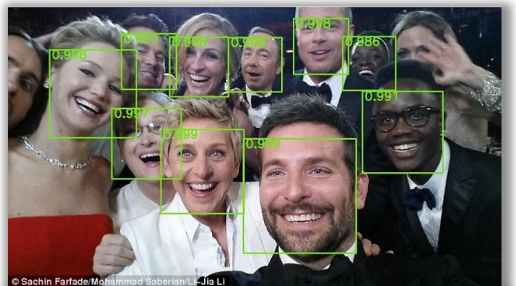
Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks

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

[1] MIT [2] NVIDIA

Future of Deep Learning



Recognition

Self-Driving Cars

AI

Computation

- **DCNN**
  15k – 300k OP/Px
- **DPM**
  0.1k – 0.5k OP/Px

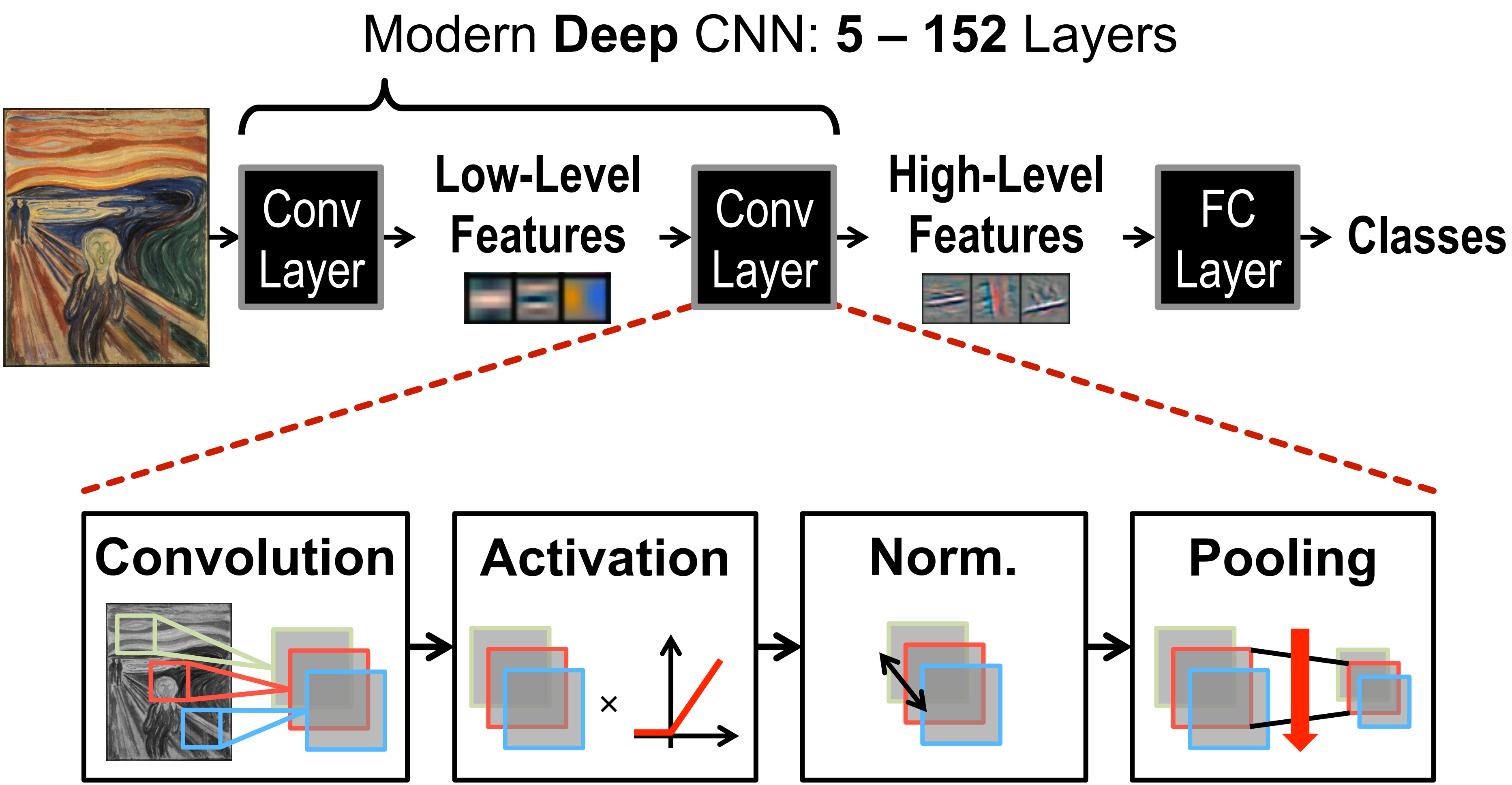
Accuracy

DCNN Accelerator is Crucial

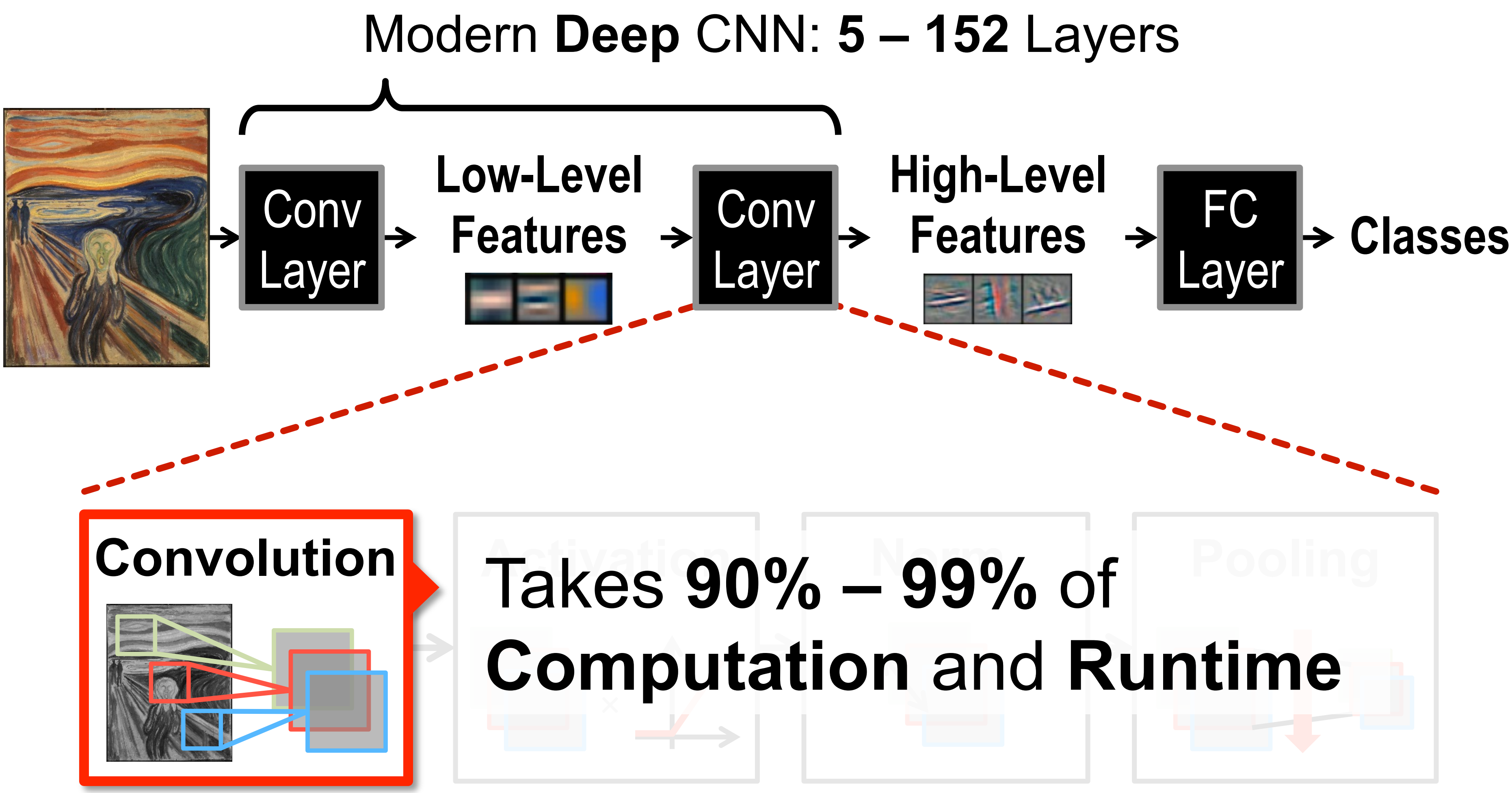
- **High Throughput** for Real-time Processing
- **Sub-watt Power/Energy** Consumption

- DPM: Deformable Part Model

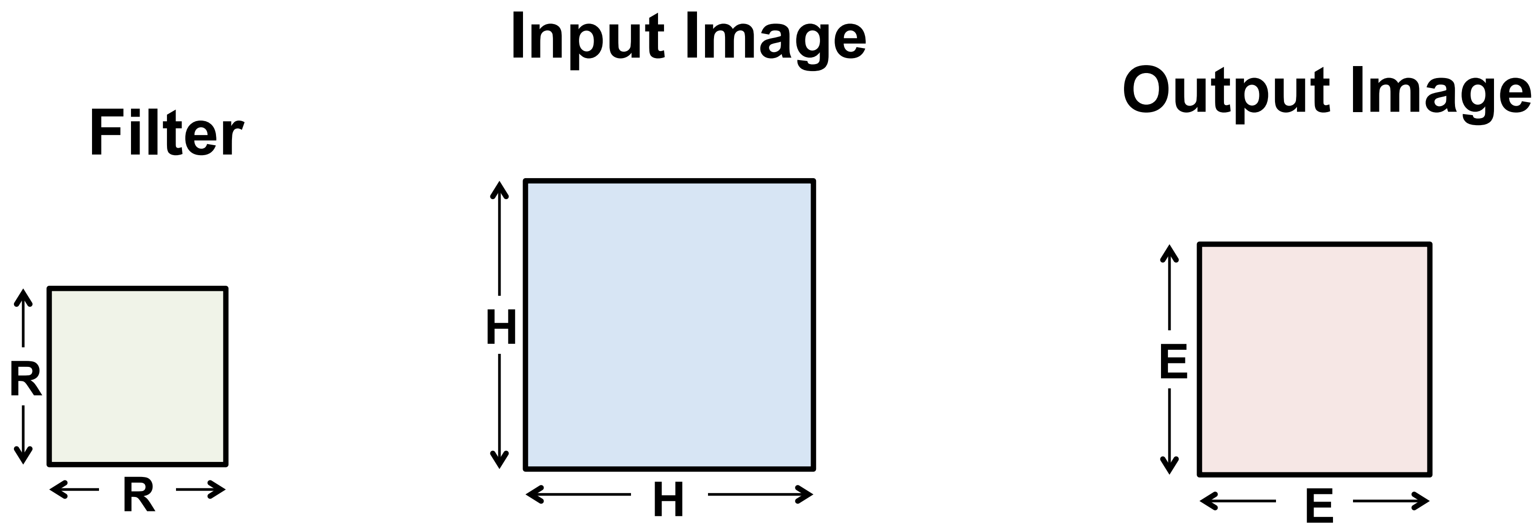
DCNN Explained



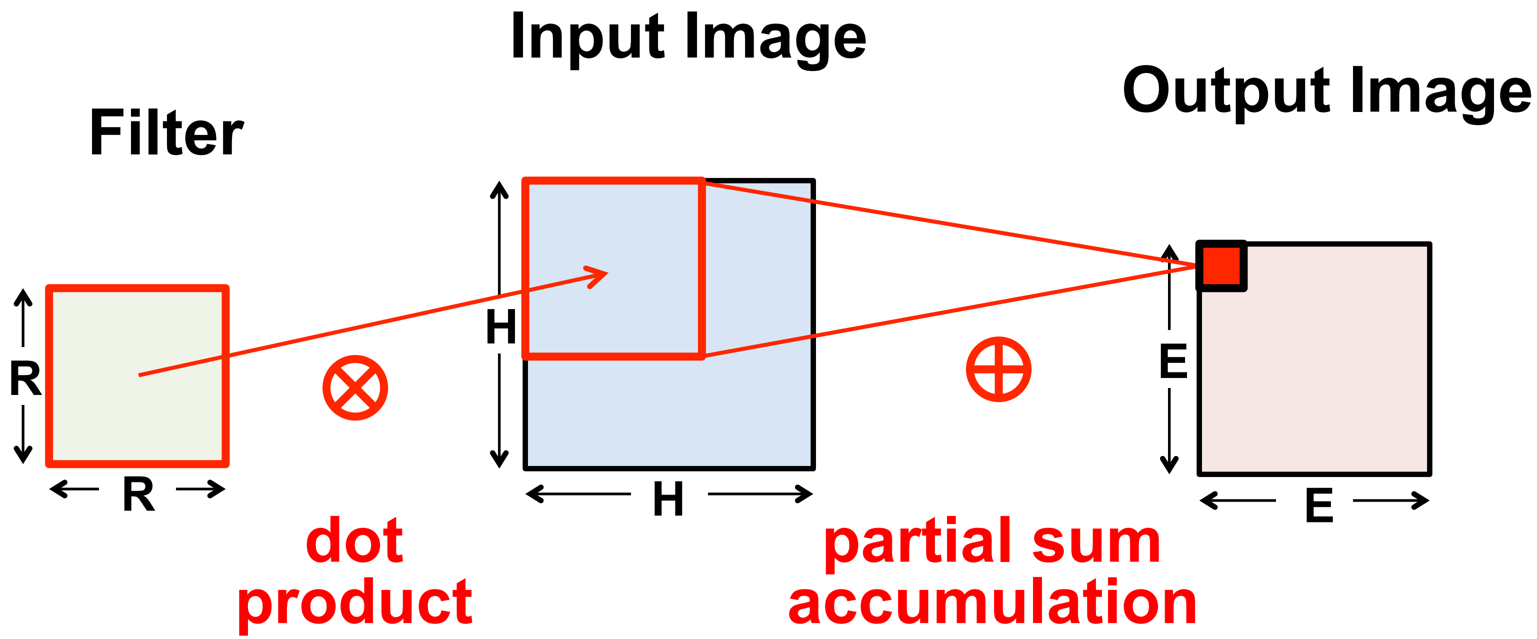
Convolution is the Most Important



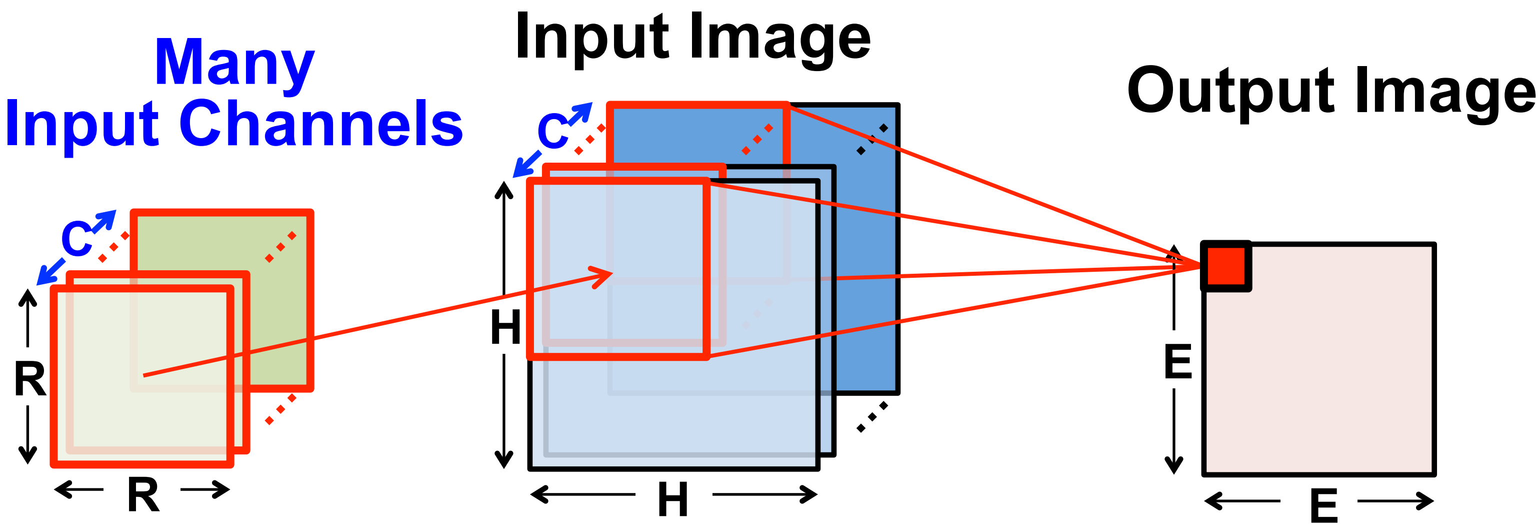
Convolution in CNN



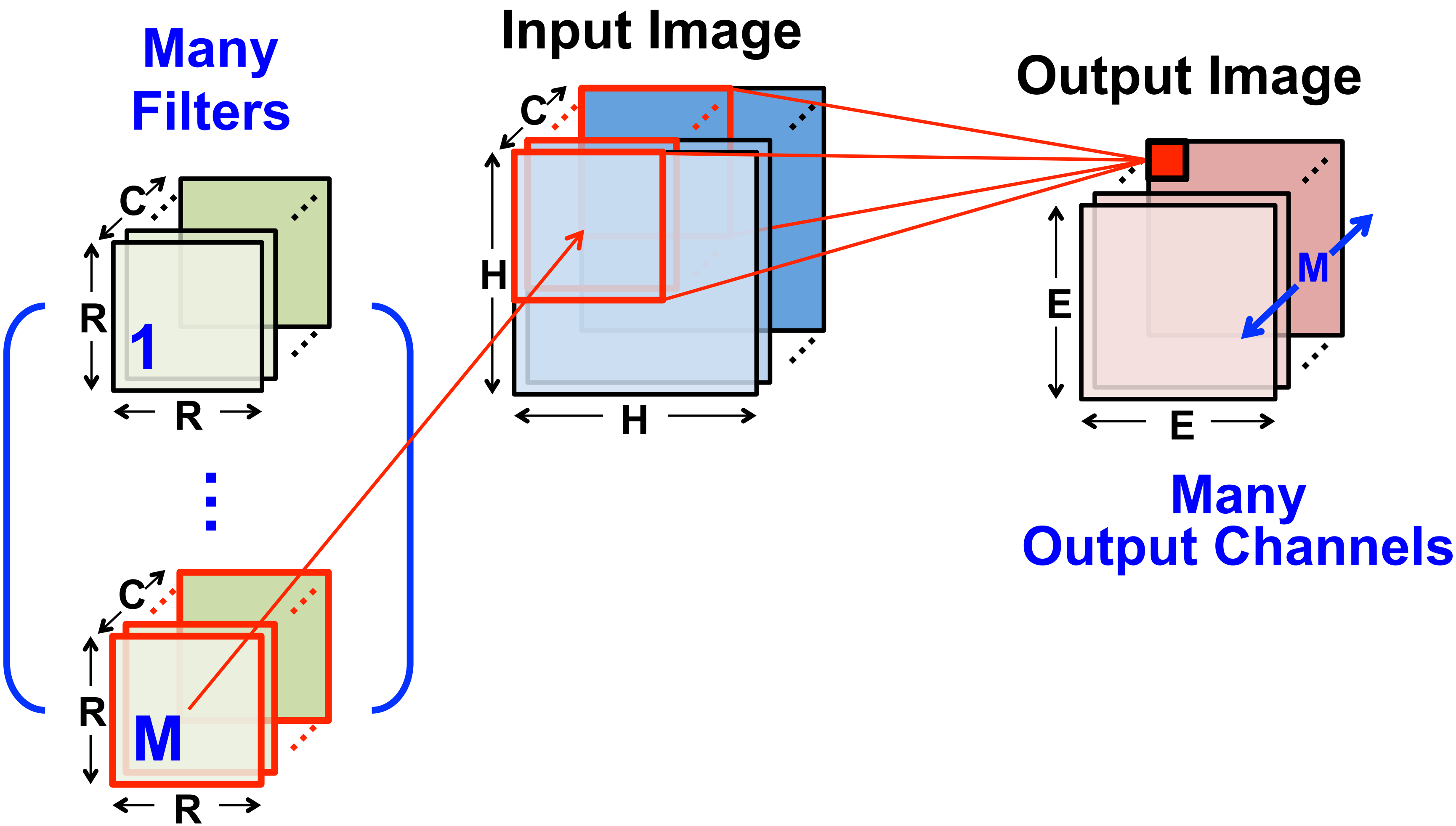
Convolution in CNN



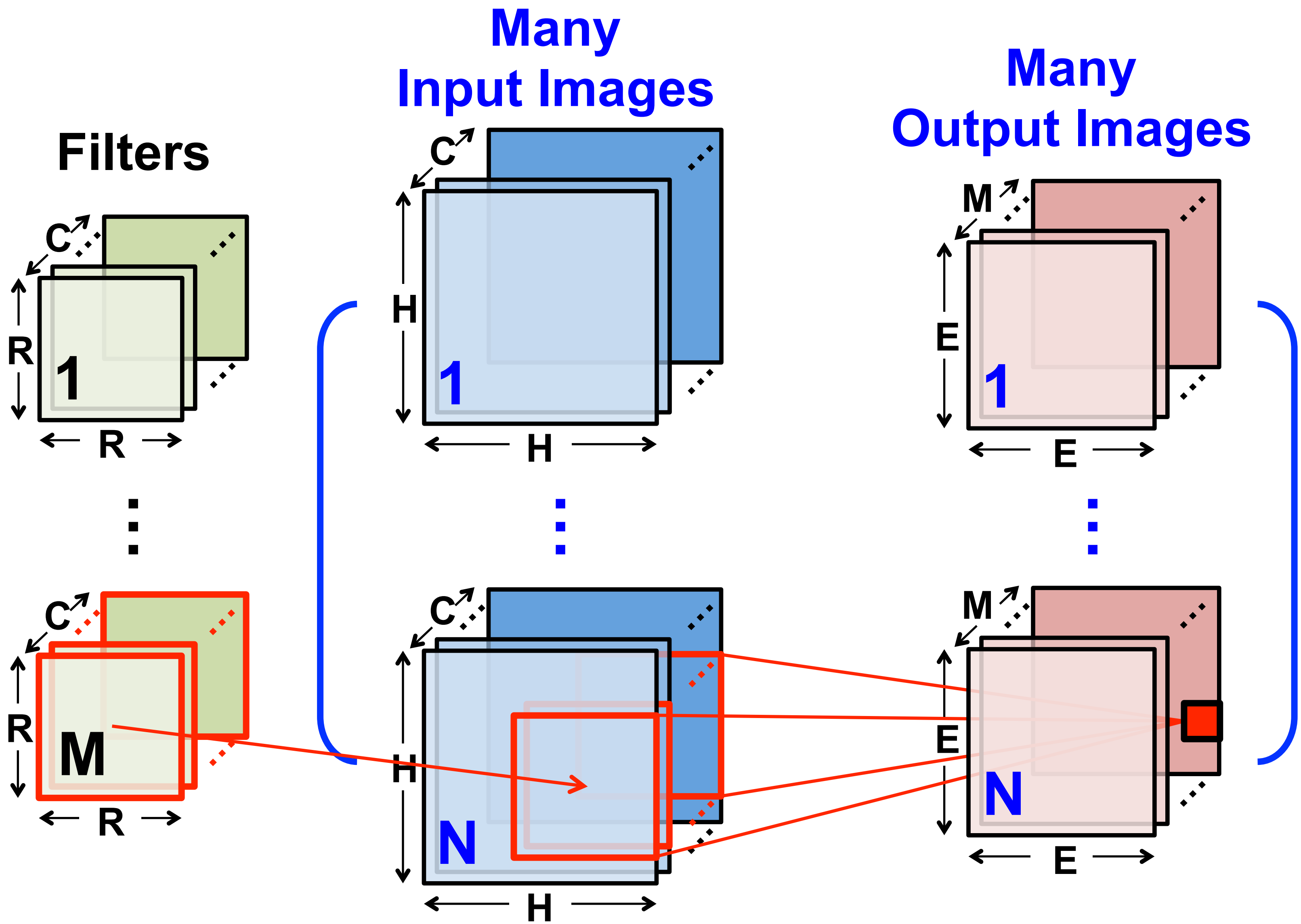
Convolution in CNN



Convolution in CNN



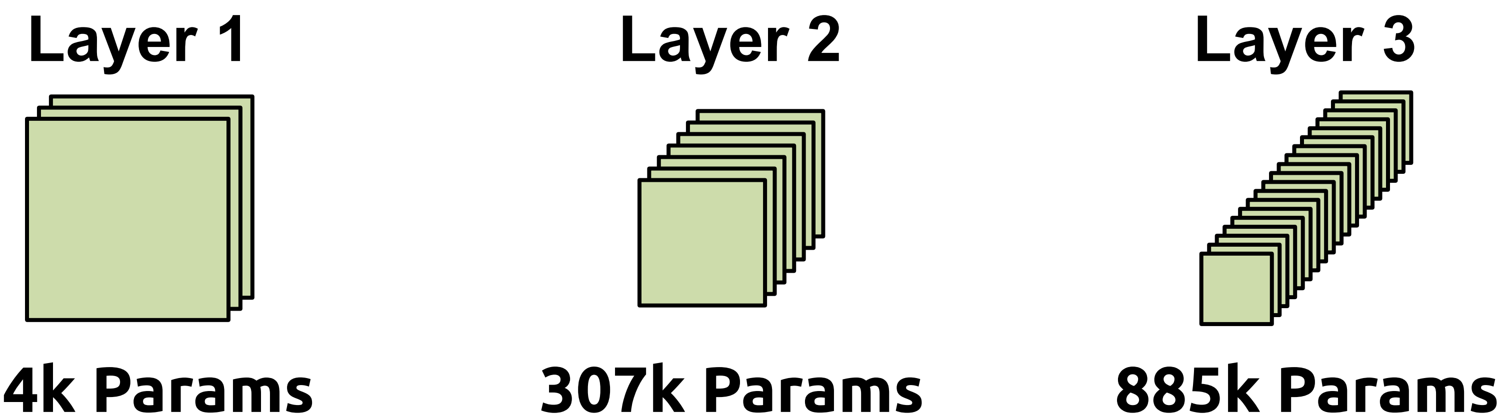
Convolution in CNN



Large Sizes with Varying Shapes

AlexNet[1] Convolutional Layer Configurations

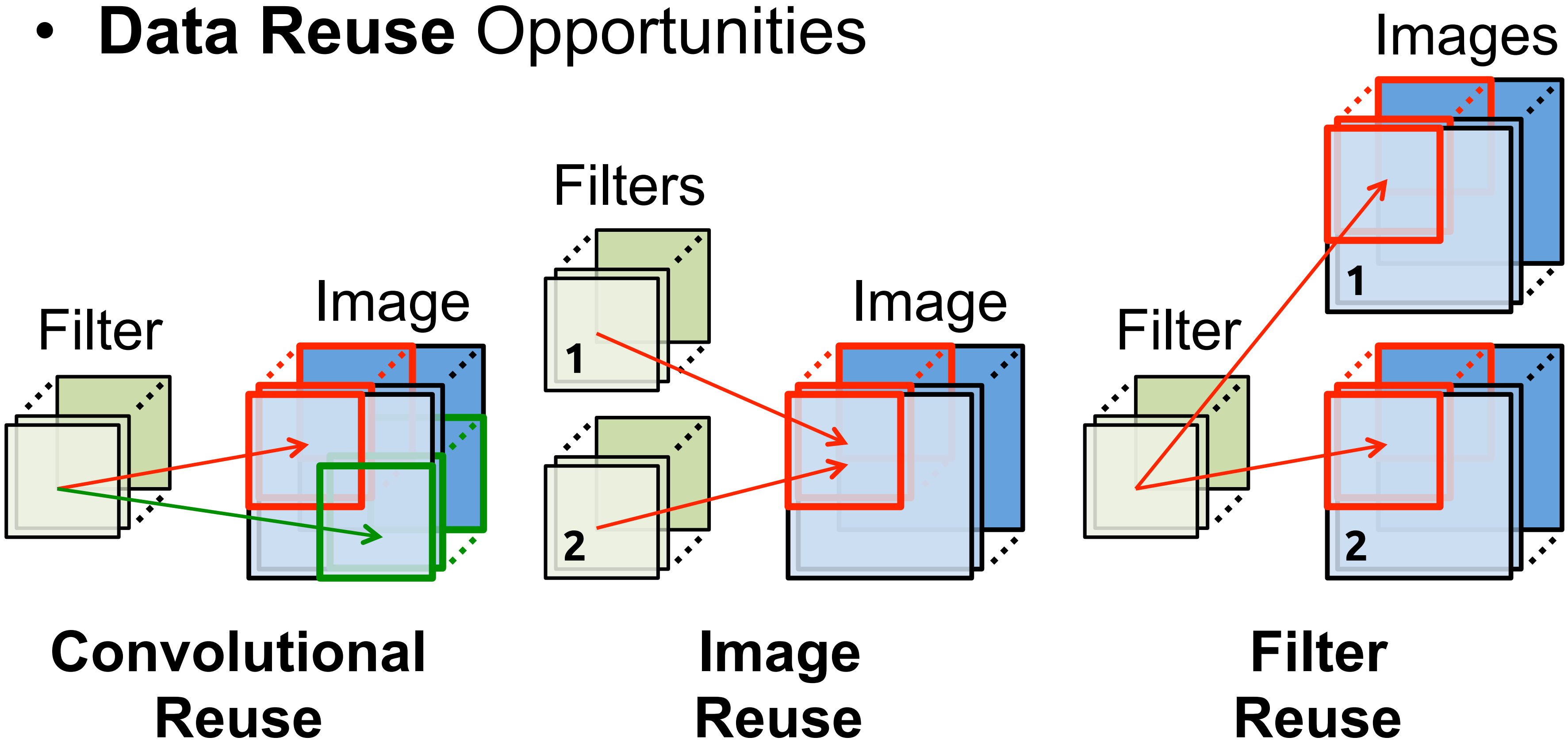
Layer	Filter Size (R)	# Filters (M)	# Channels (C)	Stride
1	11x11	96	3	4
2	5x5	256	48	1
3	3x3	384	256	1
4	3x3	384	192	1
5	3x3	256	192	1



1. [Krizhevsky, NIPS 2012]

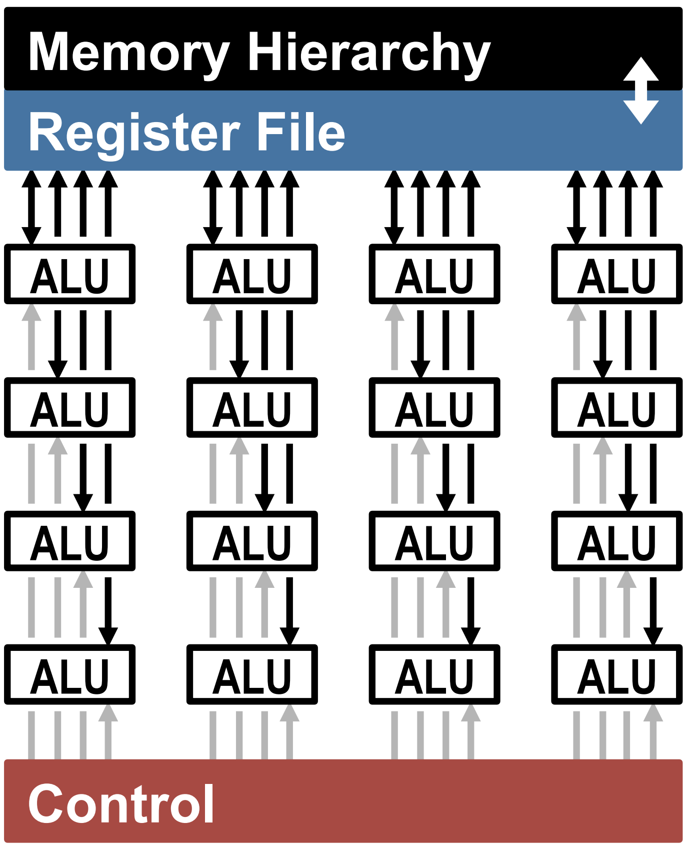
Properties We Can Leverage

- Operations exhibit **High Parallelism**
- **Data Reuse** Opportunities

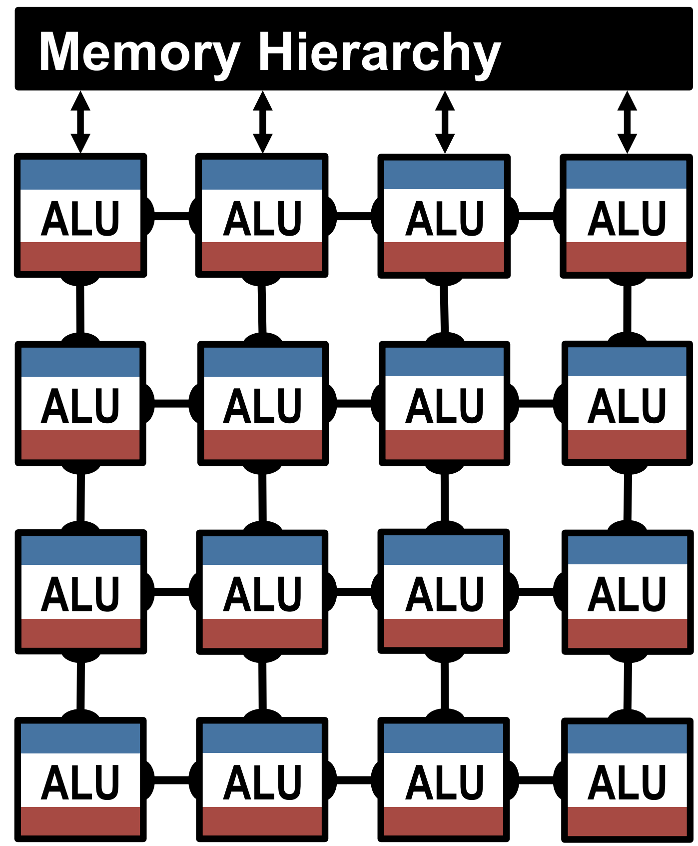


Highly Parallel Compute Paradigms

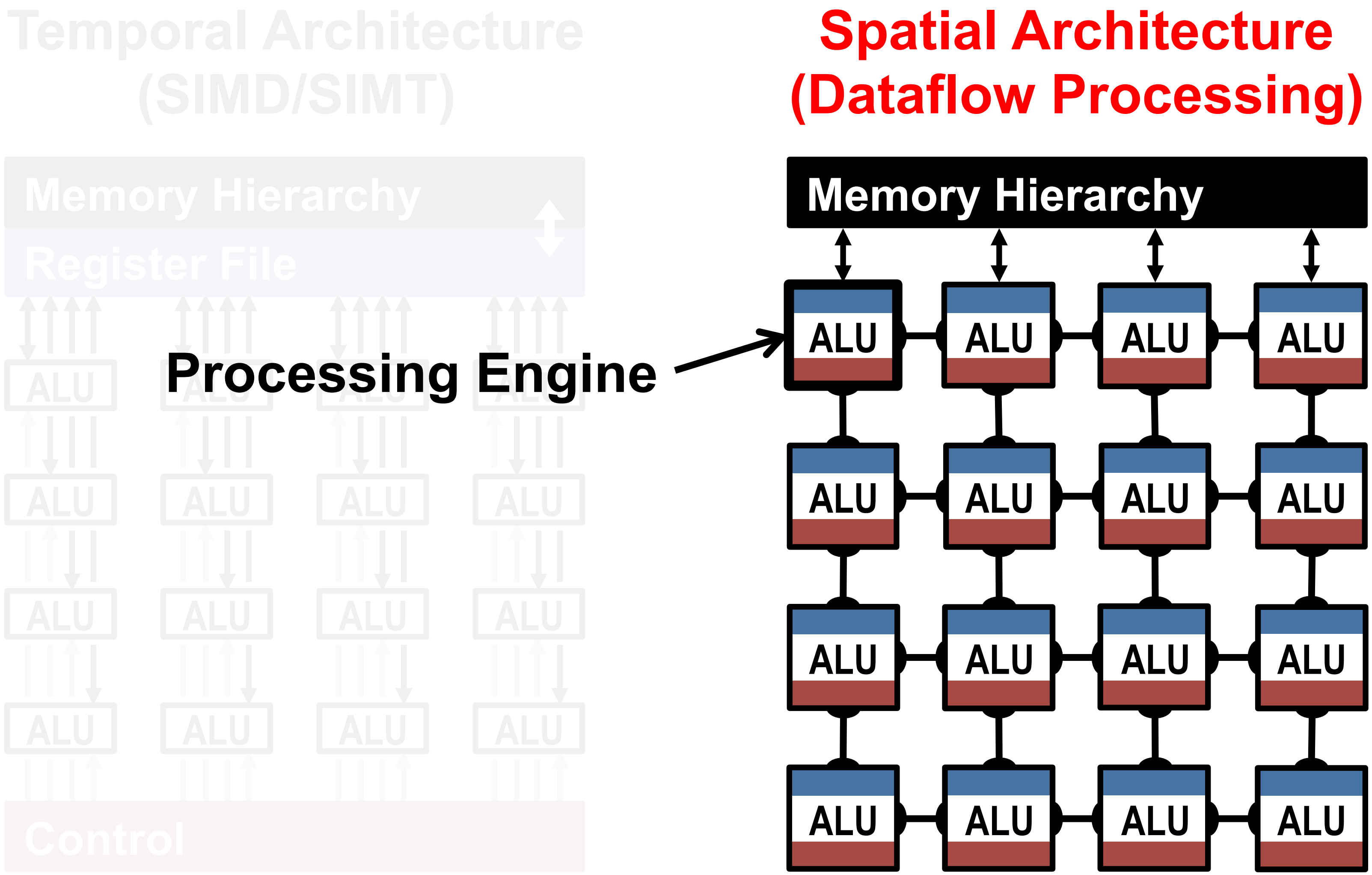
Temporal Architecture (SIMD/SIMT)



Spatial Architecture (Dataflow Processing)

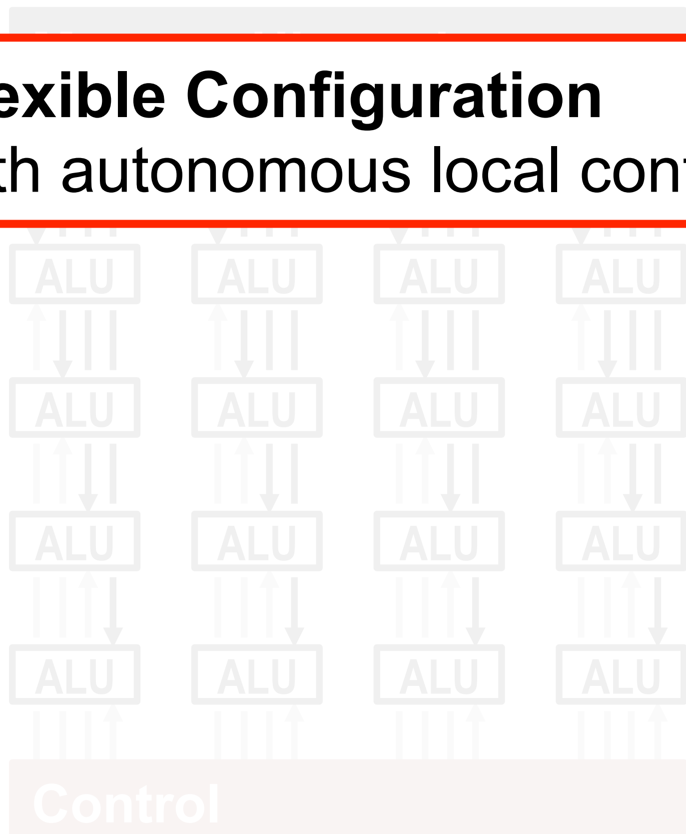


Highly Parallel Compute Paradigms



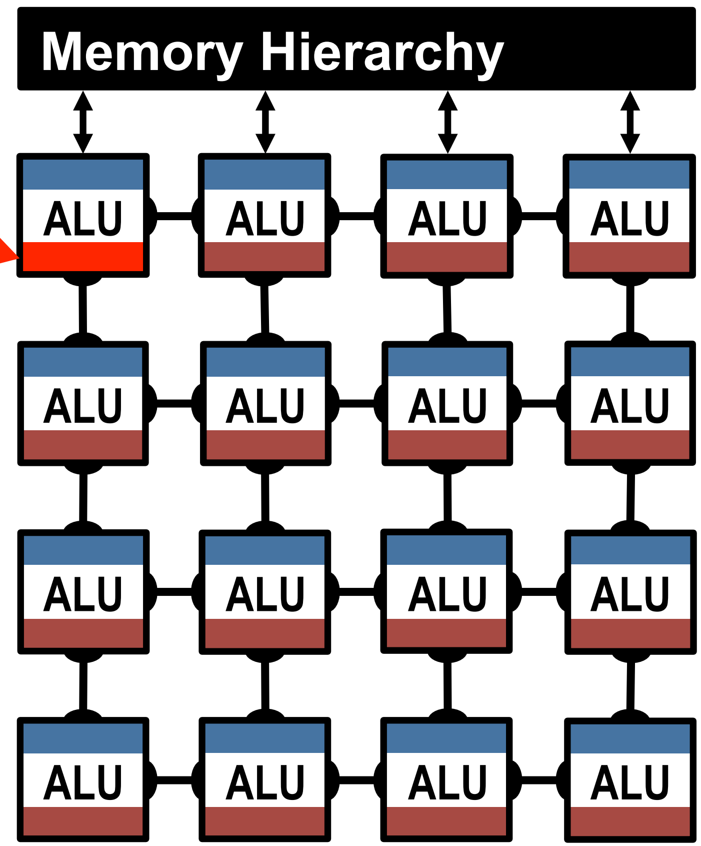
Highly Parallel Compute Paradigms

Temporal Architecture
(SIMD/SIMT)



Spatial Architecture
(Dataflow Processing)

Flexible Configuration
with autonomous local control

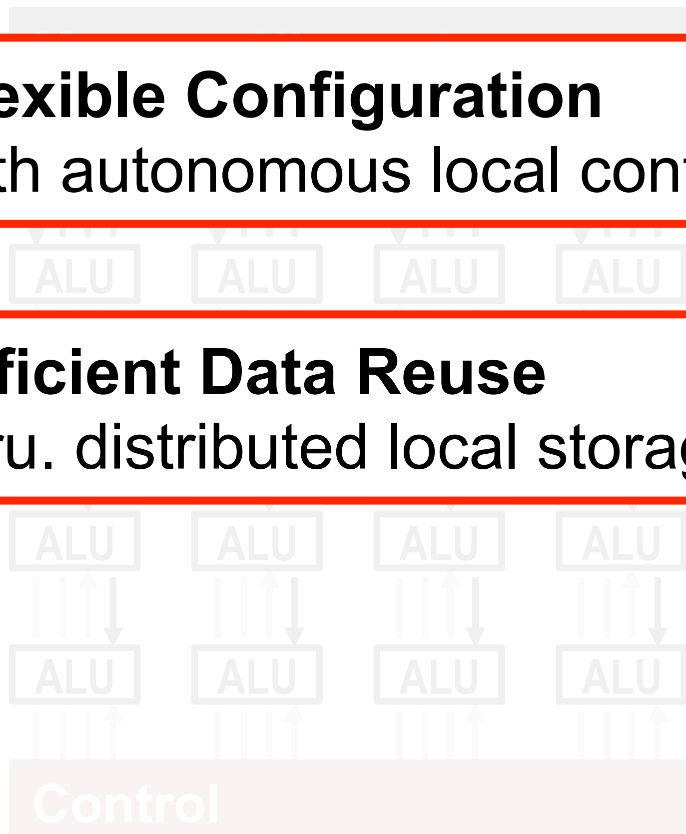


Highly Parallel Compute Paradigms

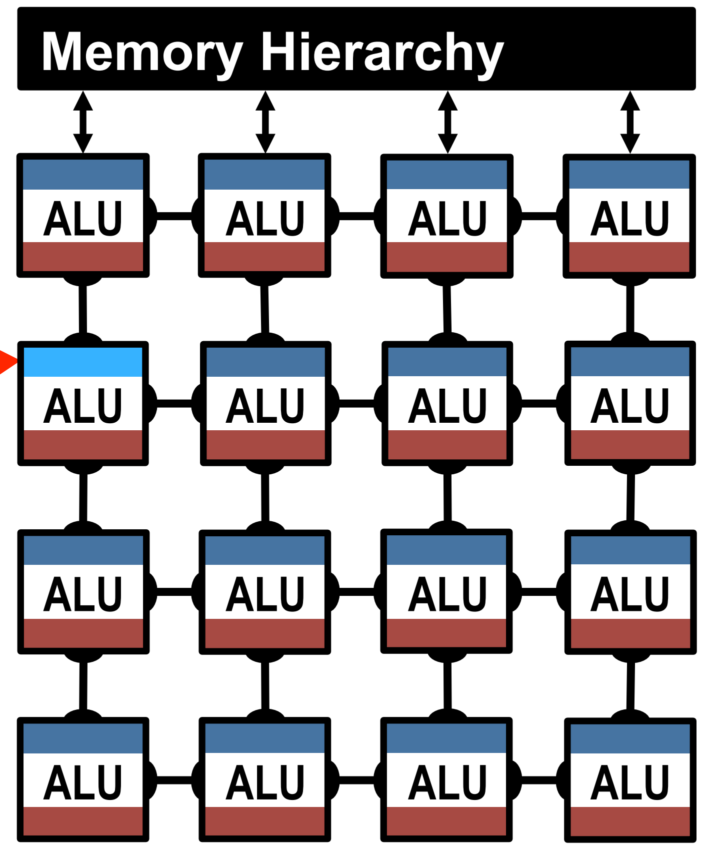
Temporal Architecture (SIMD/SIMT)

Flexible Configuration
with autonomous local control

Efficient Data Reuse
thru. distributed local storage



Spatial Architecture (Dataflow Processing)



Highly Parallel Compute Paradigms

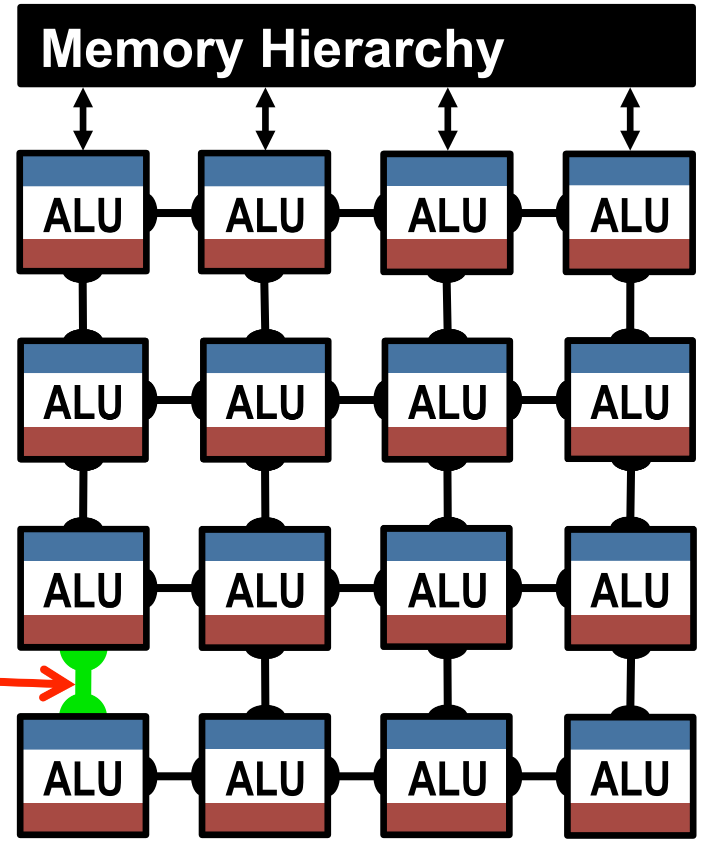
Temporal Architecture
(SIMD/SIMT)

Spatial Architecture (Dataflow Processing)

Flexible Configuration
with autonomous local control

Efficient Data Reuse
thru. distributed local storage

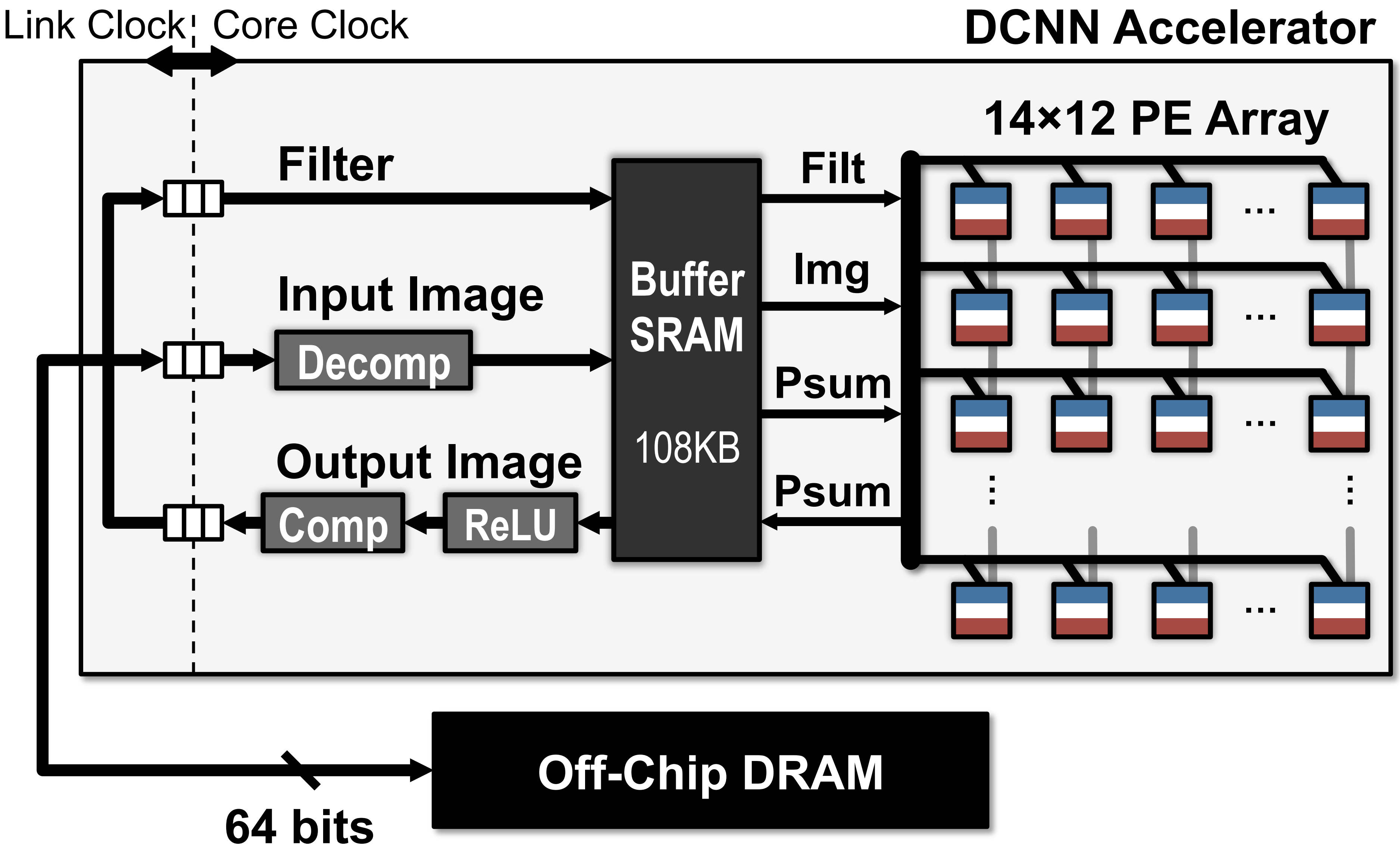
Natural Dataflow Mapping
in-place data consumption



Hardware Architecture

- Reduce Data Movement
- Exploit Data Statistics

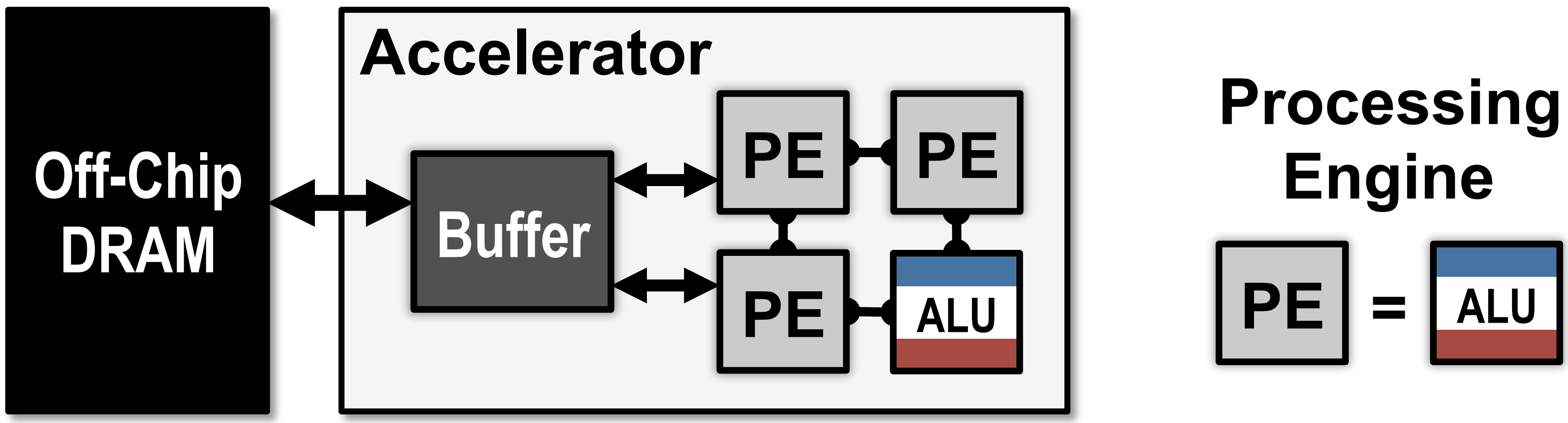
The DCNN Accelerator Architecture

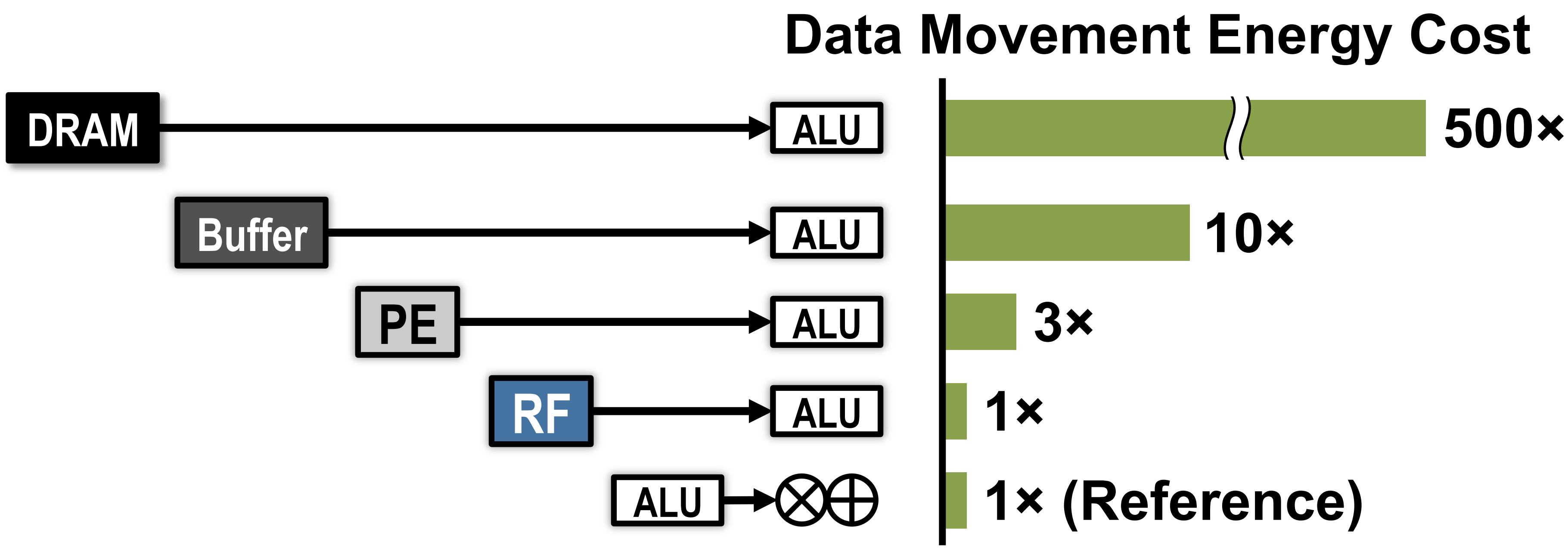


Hardware Architecture

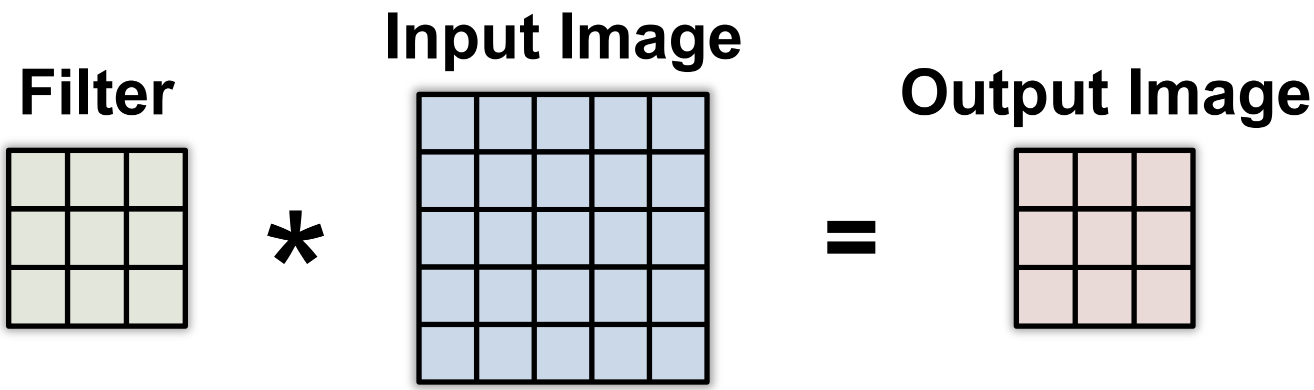
- **Reduce Data Movement**
- Exploit Data Statistics

Moving Data is Expensive

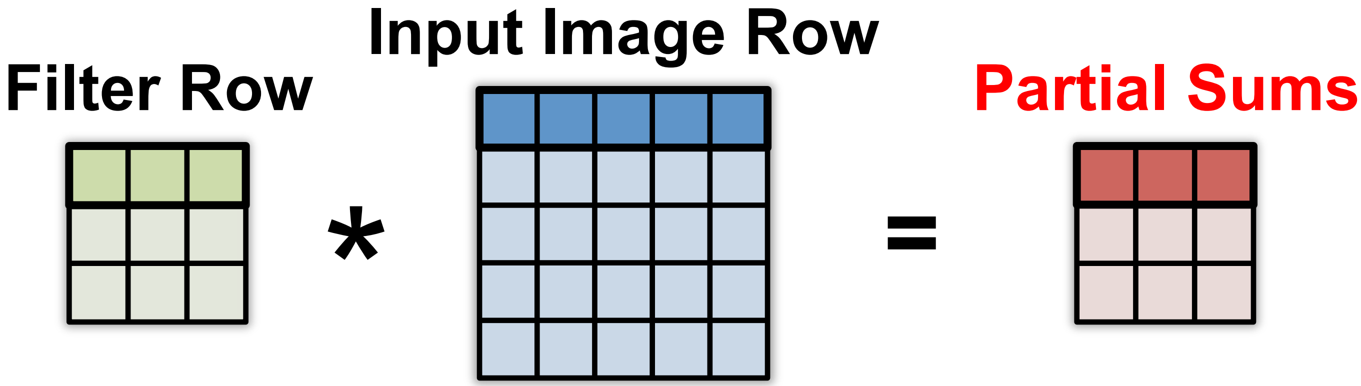




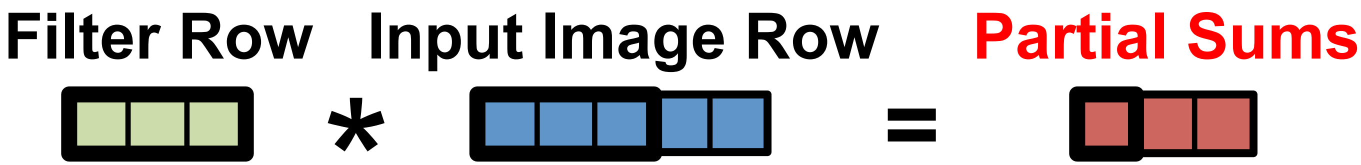
Maximize Data Reuse within PE

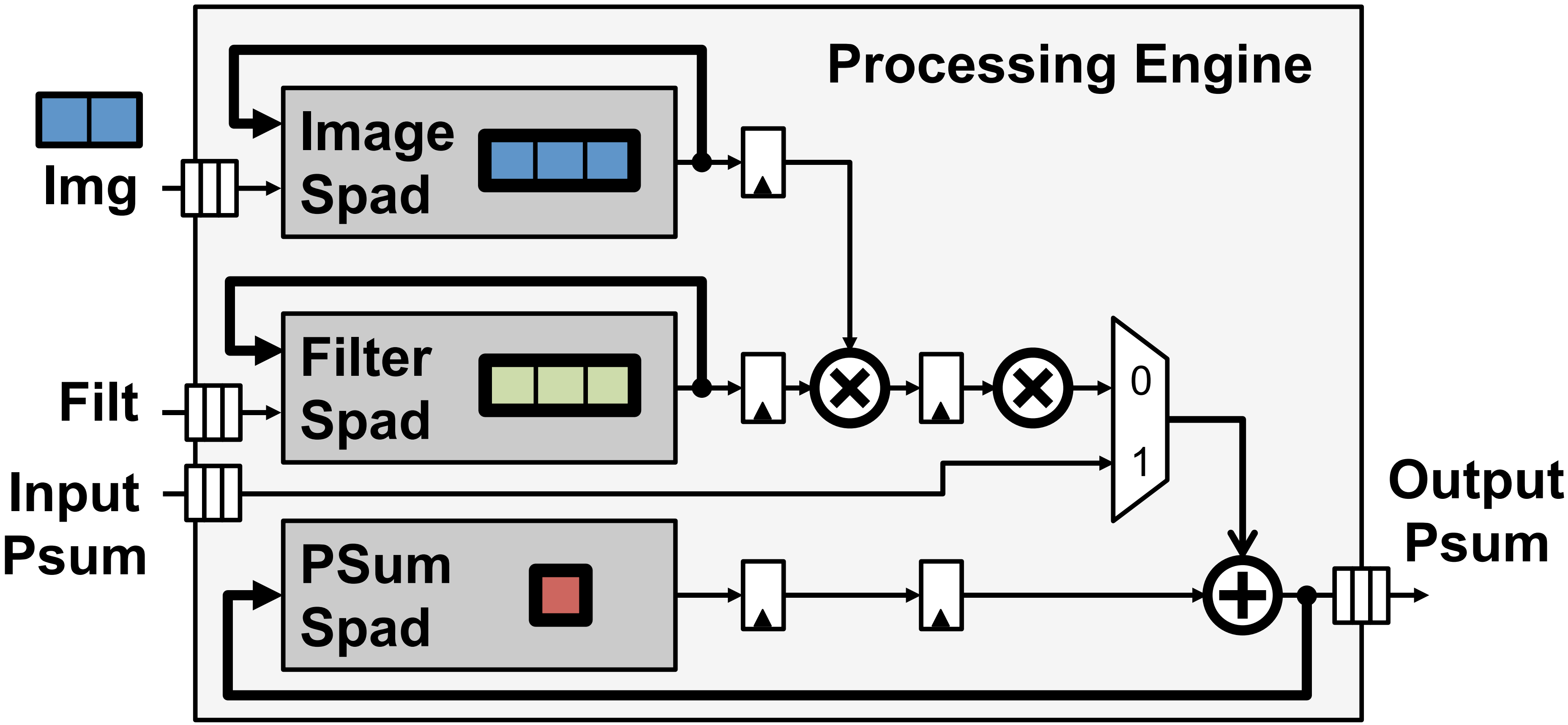


Maximize Data Reuse within PE

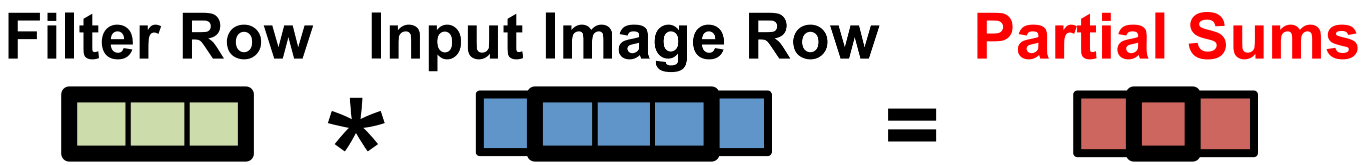


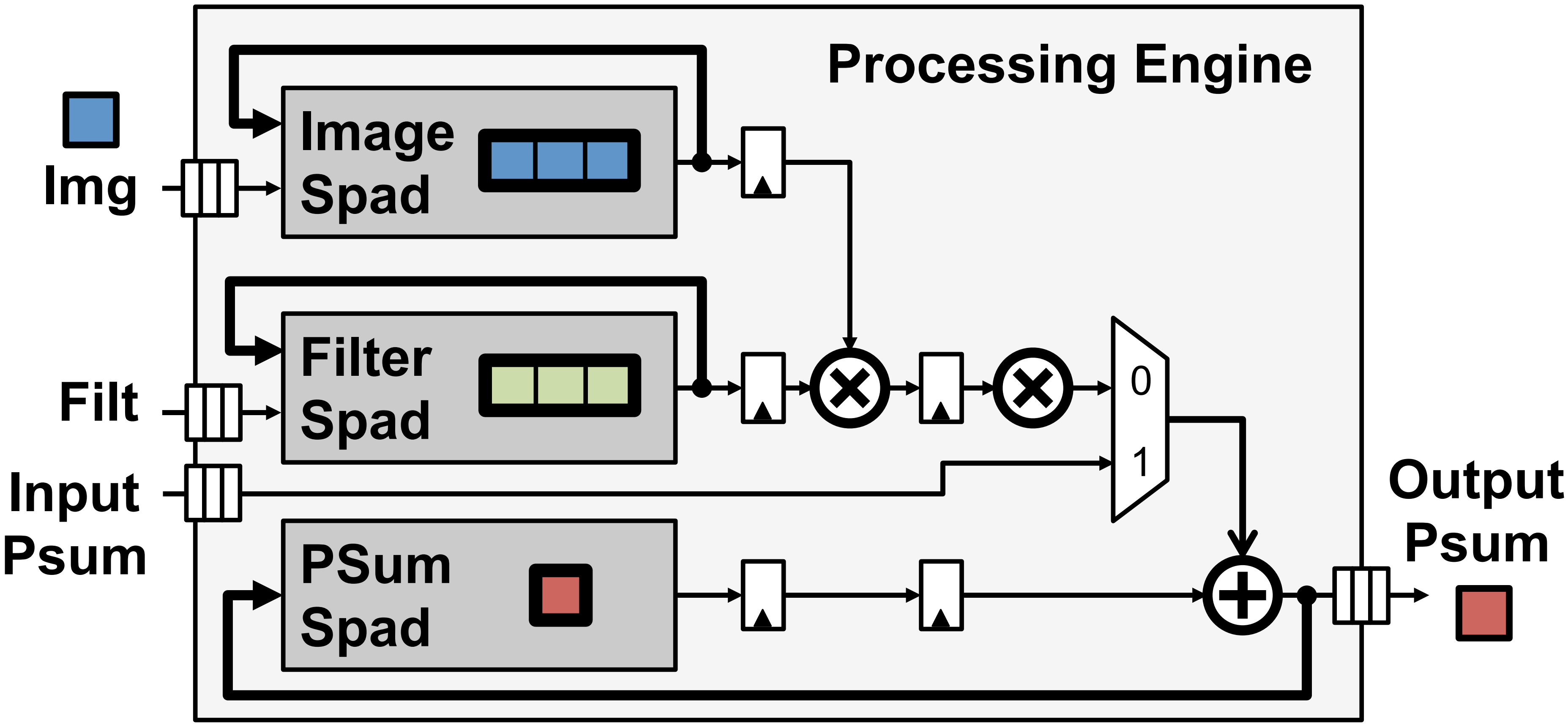
Maximize Data Reuse within PE



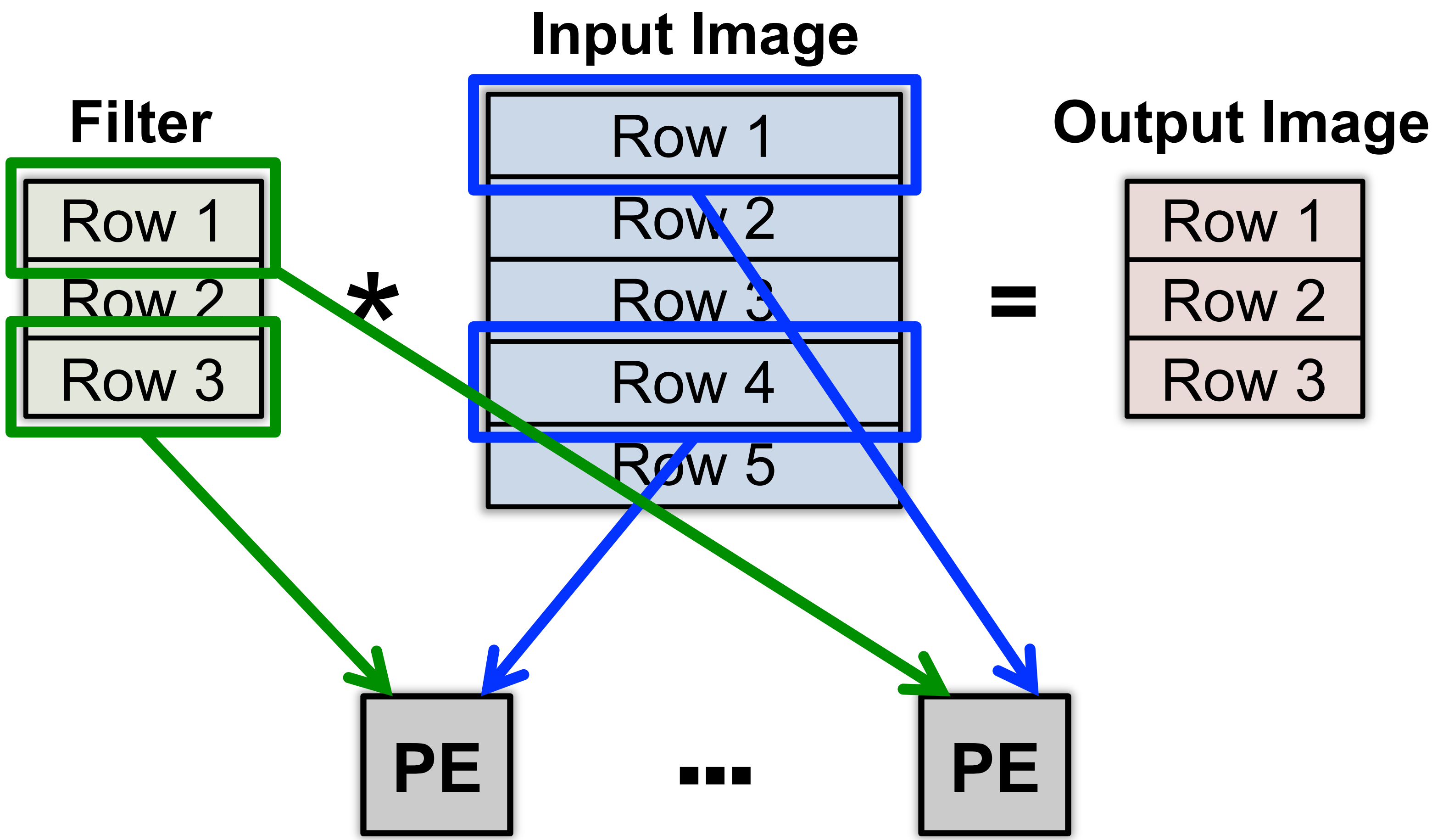


Maximize Data Reuse within PE



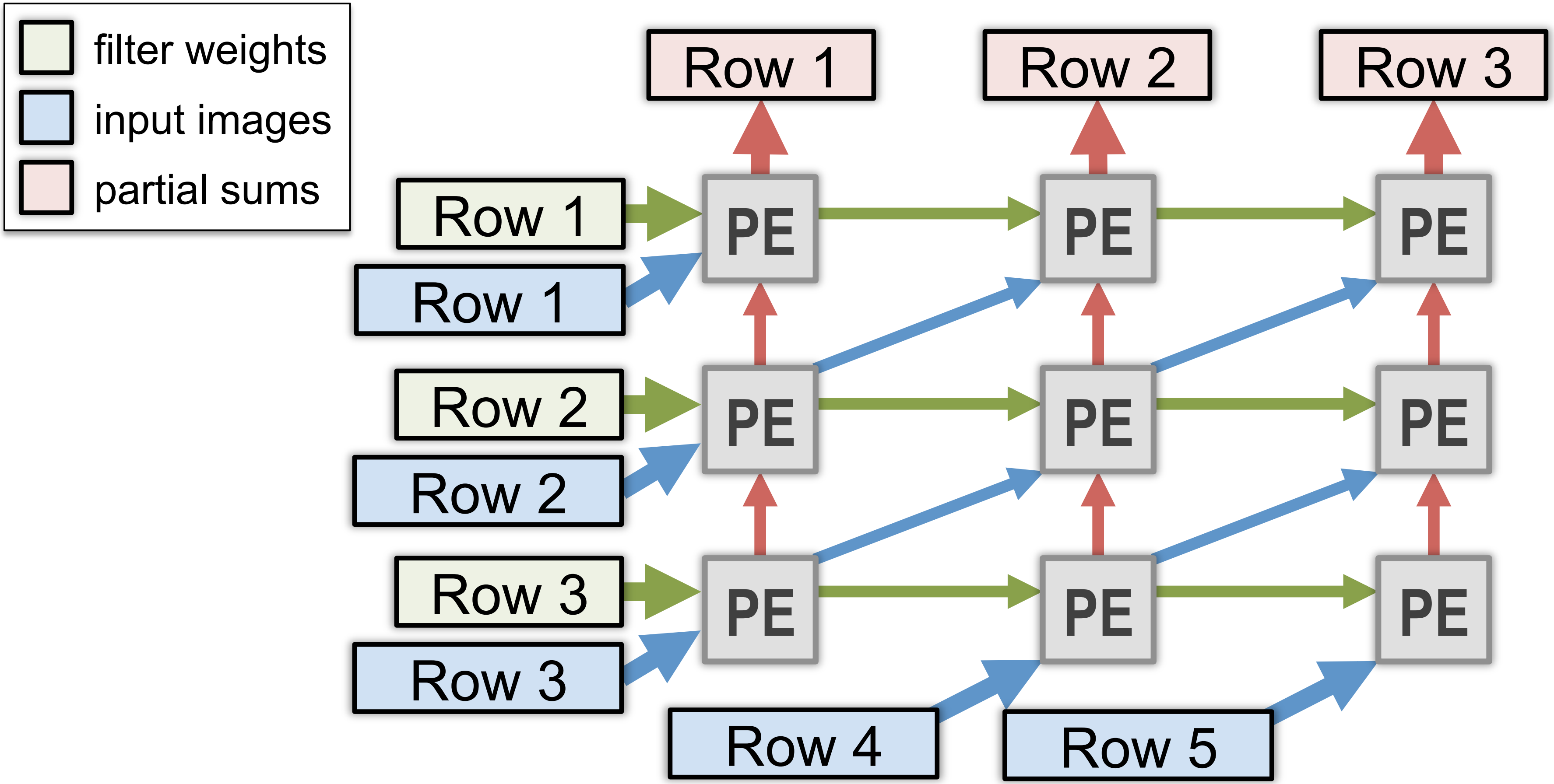


Maximize Data Reuse within PE Array



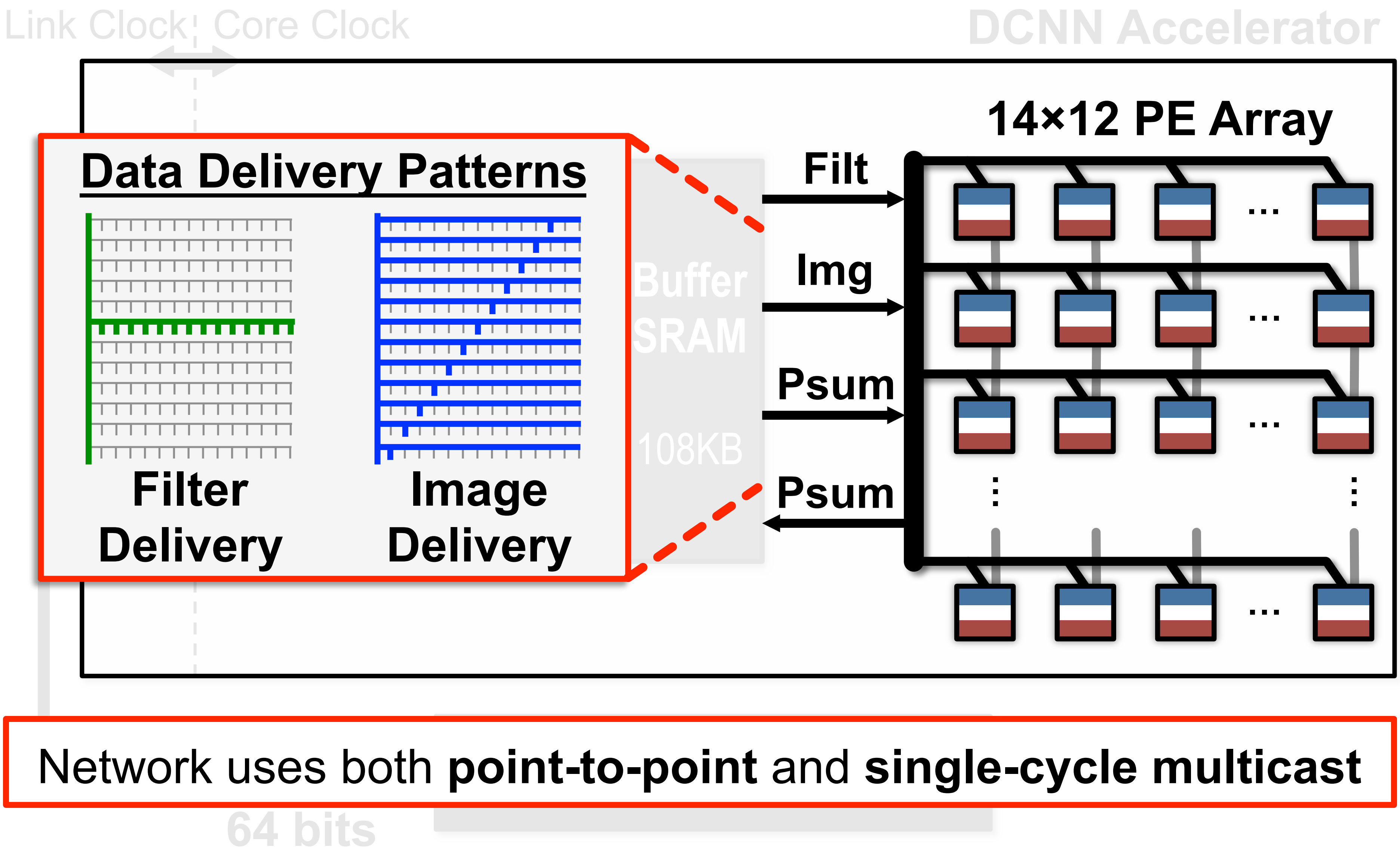
Map a pair of **Filter Row** and **Image Row** to **each PE**

Convolutional Reuse within PE Array



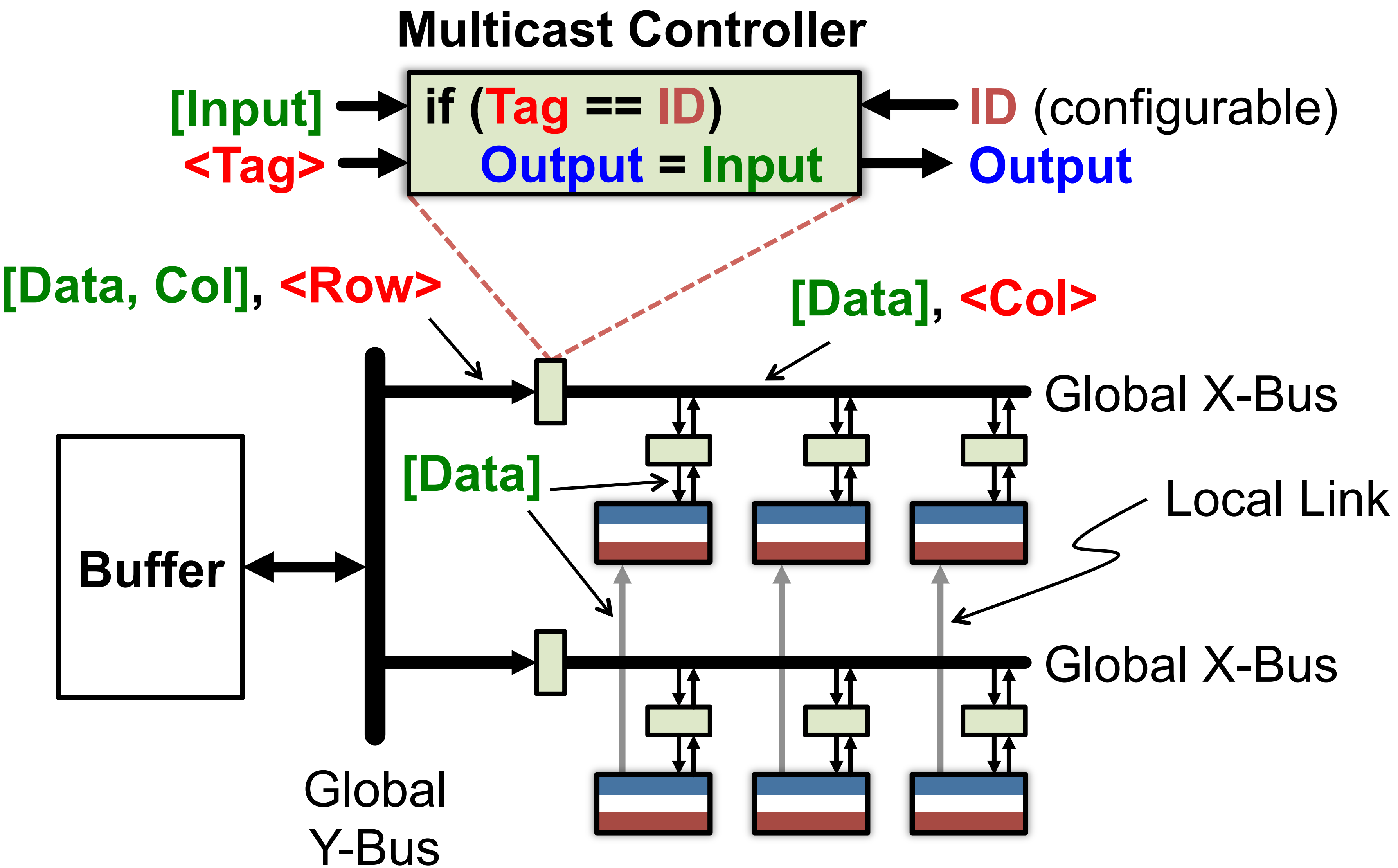
Mapping rows from **multiple channels** and/or **multiple filter/images** to each PE results in even more **reuse**

Data Delivery with On-Chip Network

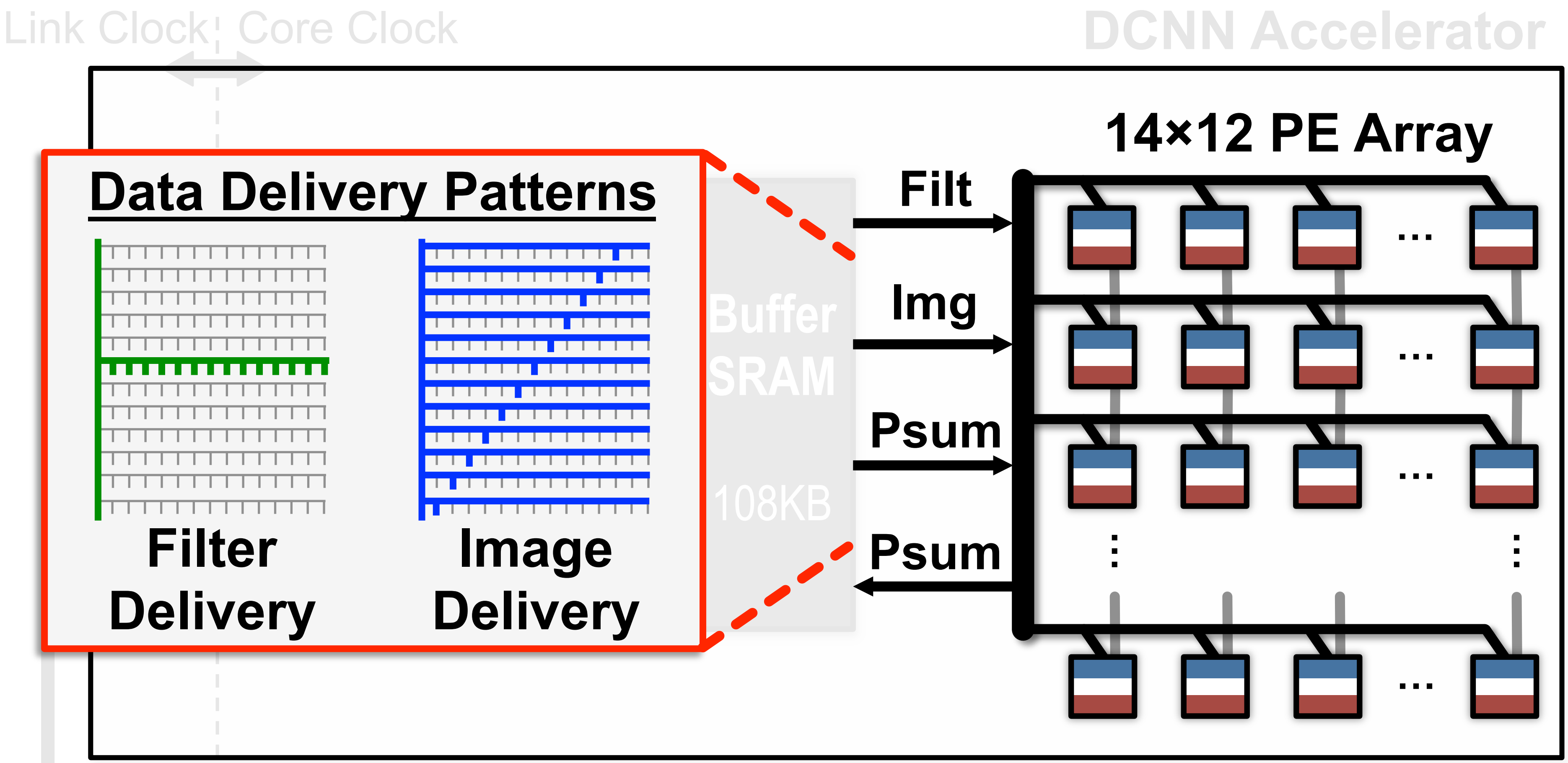


Network uses both **point-to-point** and **single-cycle multicast**

Multicast Network Design



Data Delivery with On-Chip Network

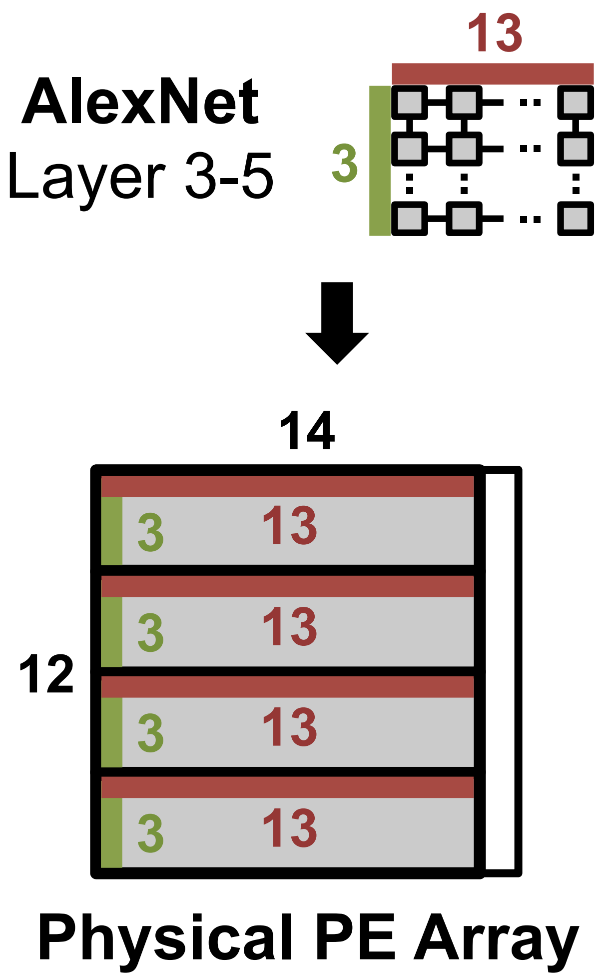


Compared to Broadcast, **Multicast** saves >**80%** of NoC energy

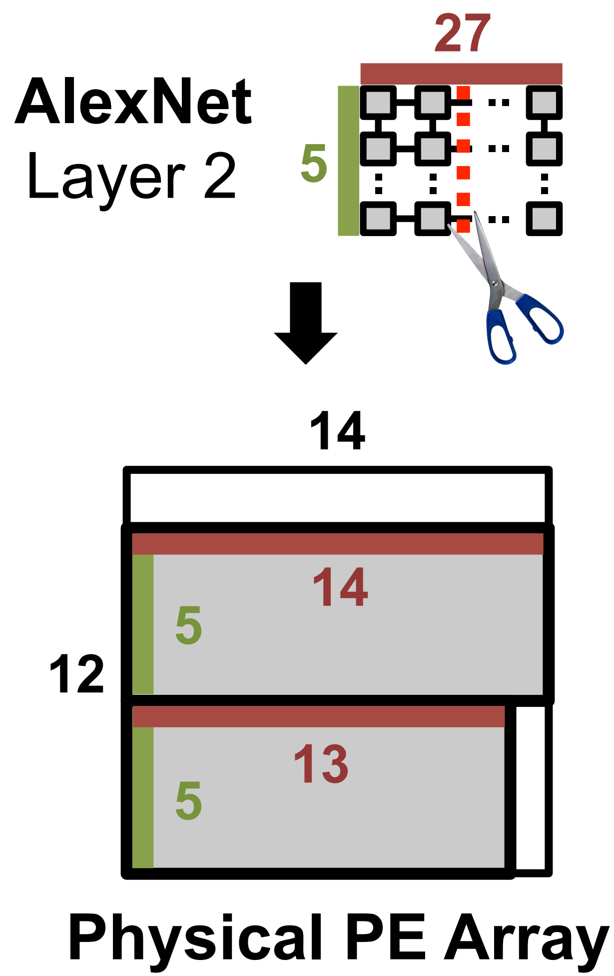
64 bits

Logical to Physical Mappings

Replication



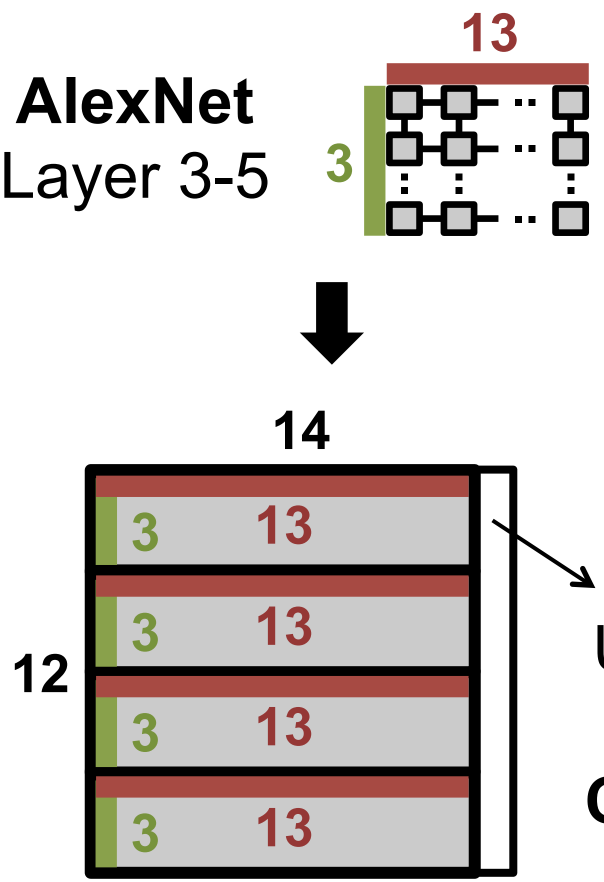
Folding



Logical to Physical Mappings

Replication

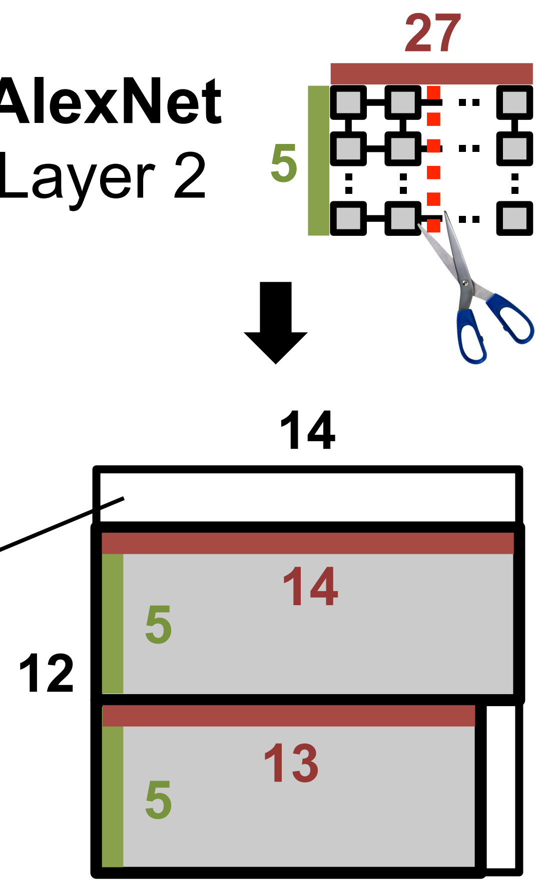
AlexNet
Layer 3-5



Physical PE Array

Folding

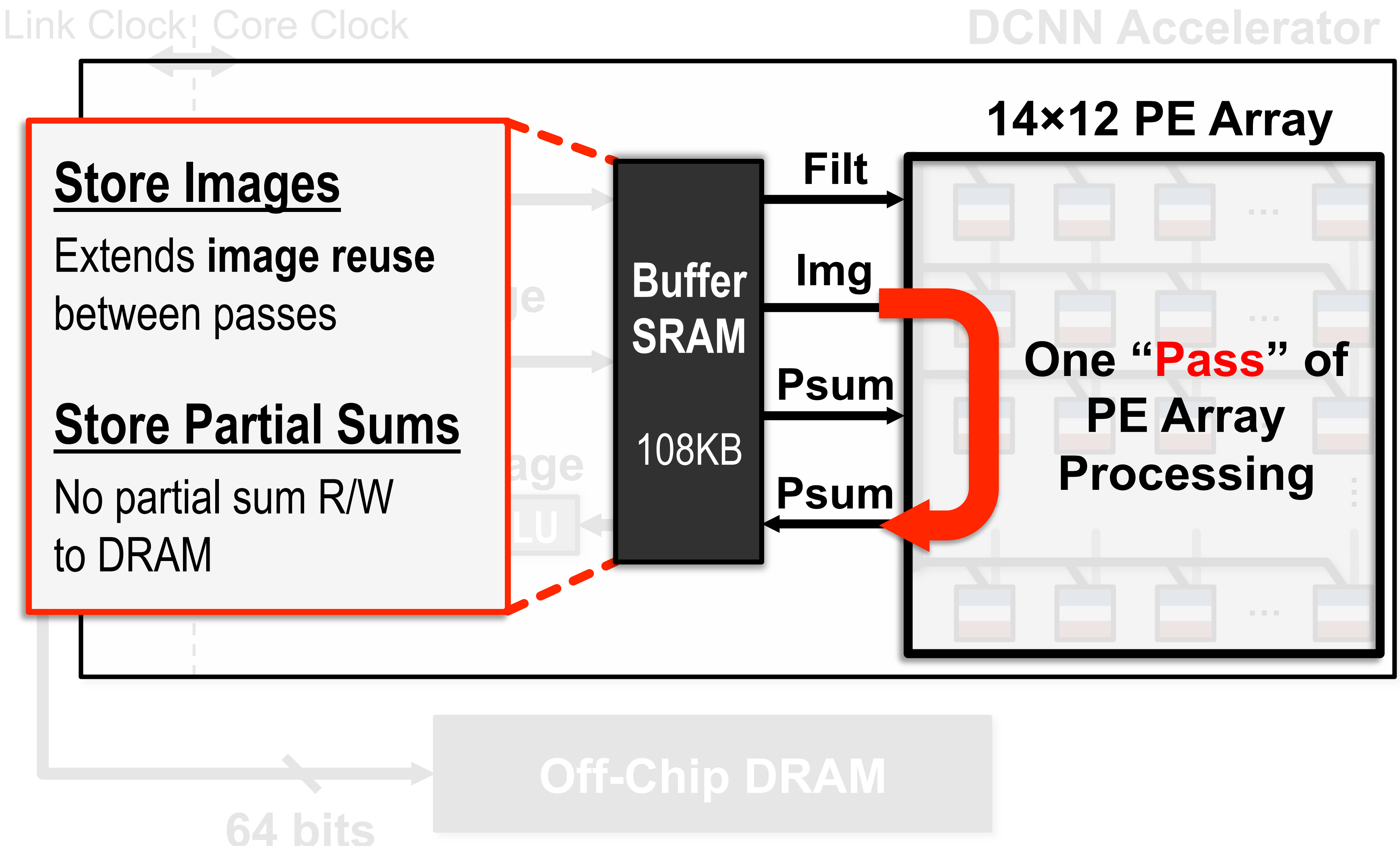
AlexNet
Layer 2



Physical PE Array

Unused PEs
are
Clock Gated

Maximize Data Reuse with Buffer

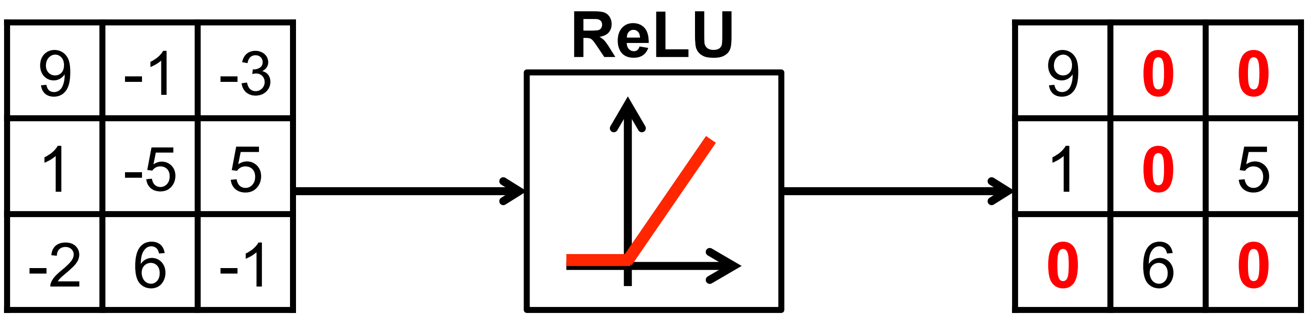


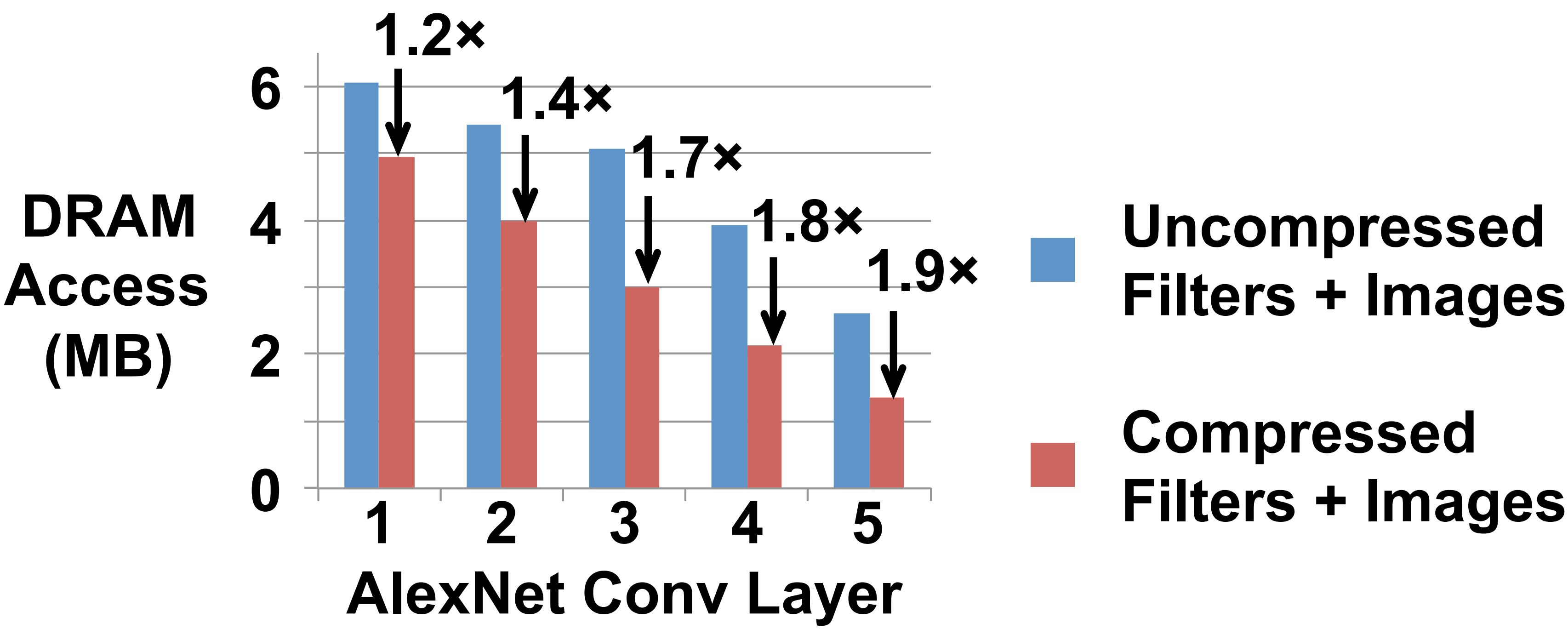
Hardware Architecture

- Reduce Data Movement
- **Exploit Data Statistics**

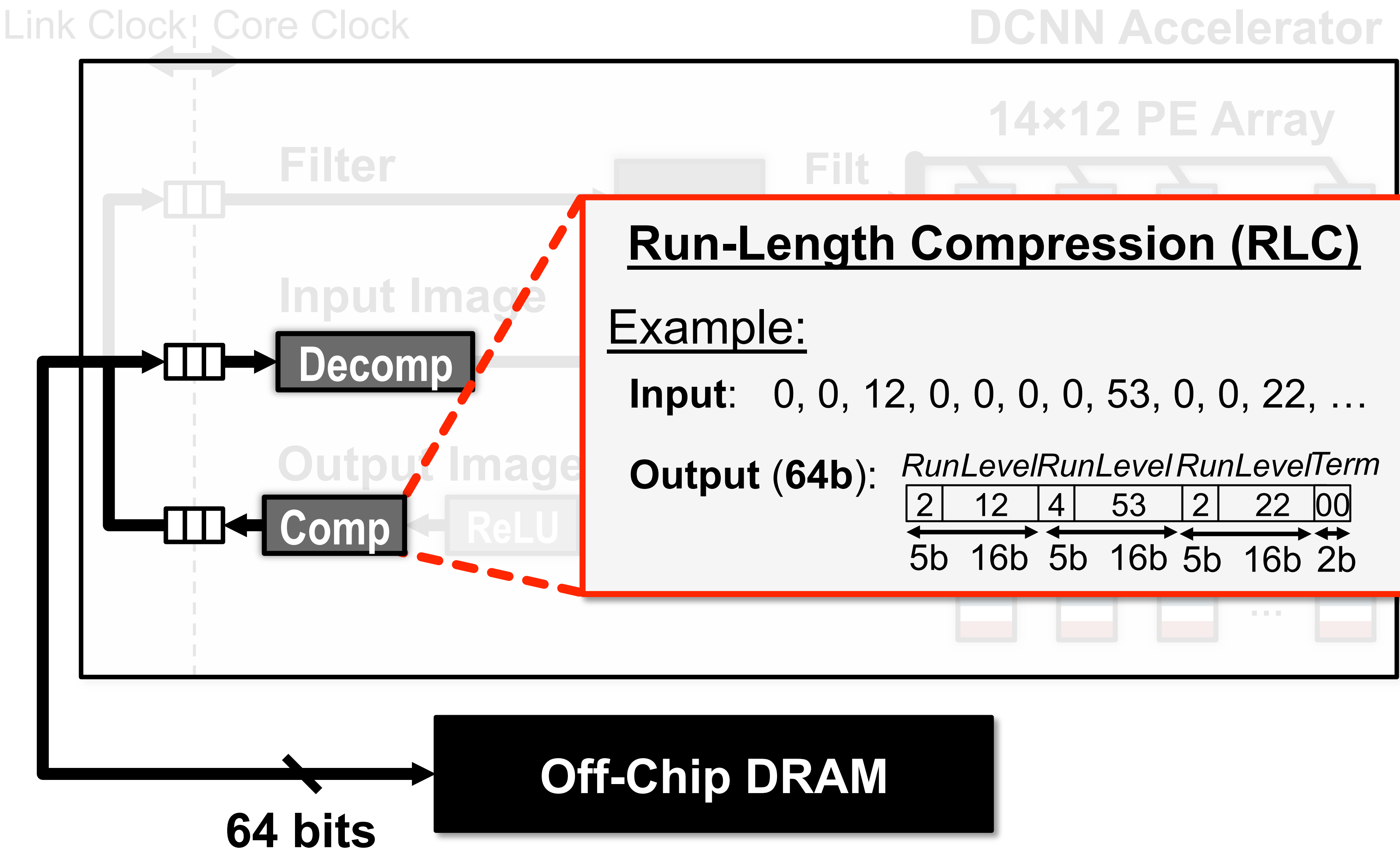
Data Compression

Apply Activation (**ReLU**) on Filtered Image Data

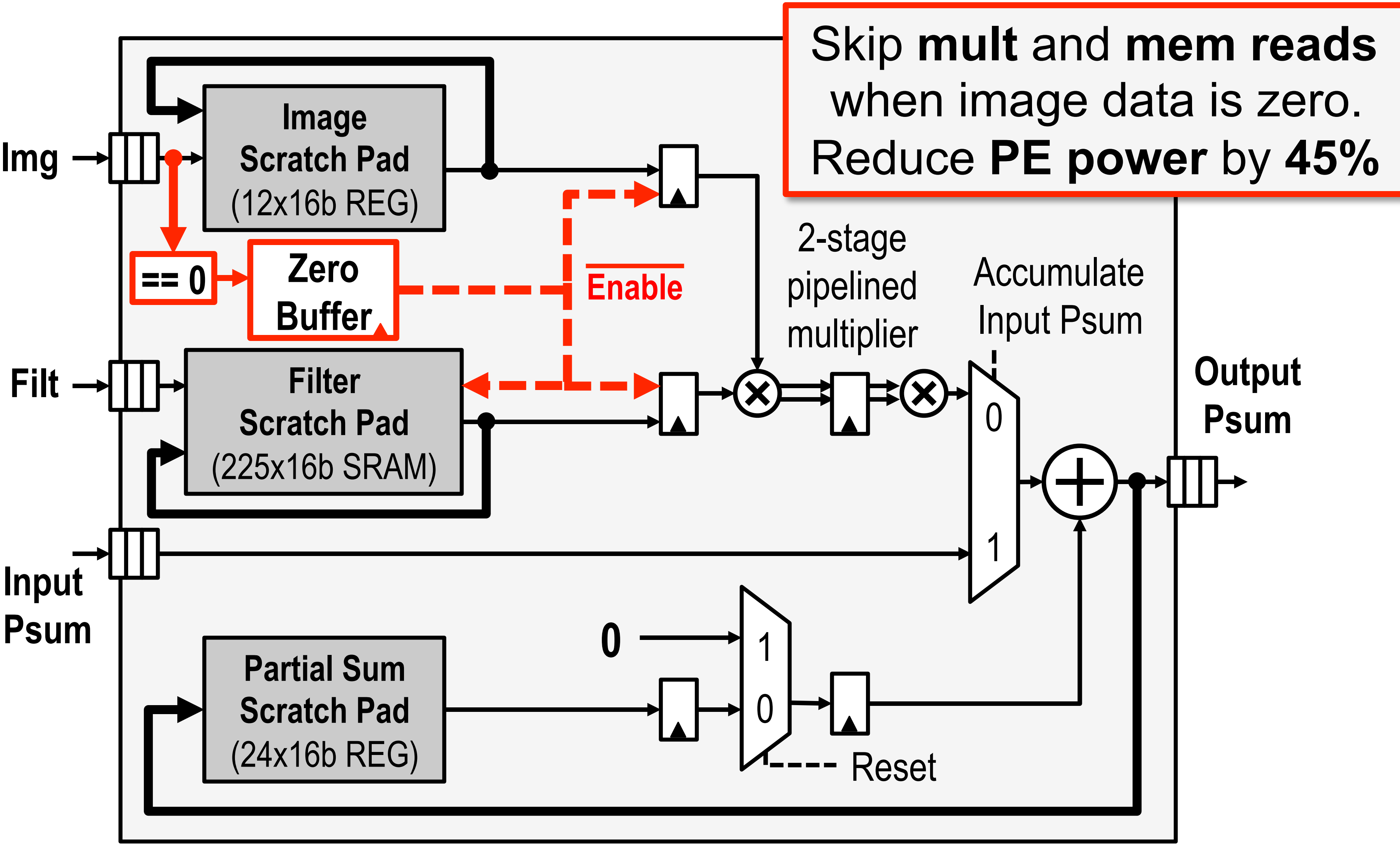




Data Compression



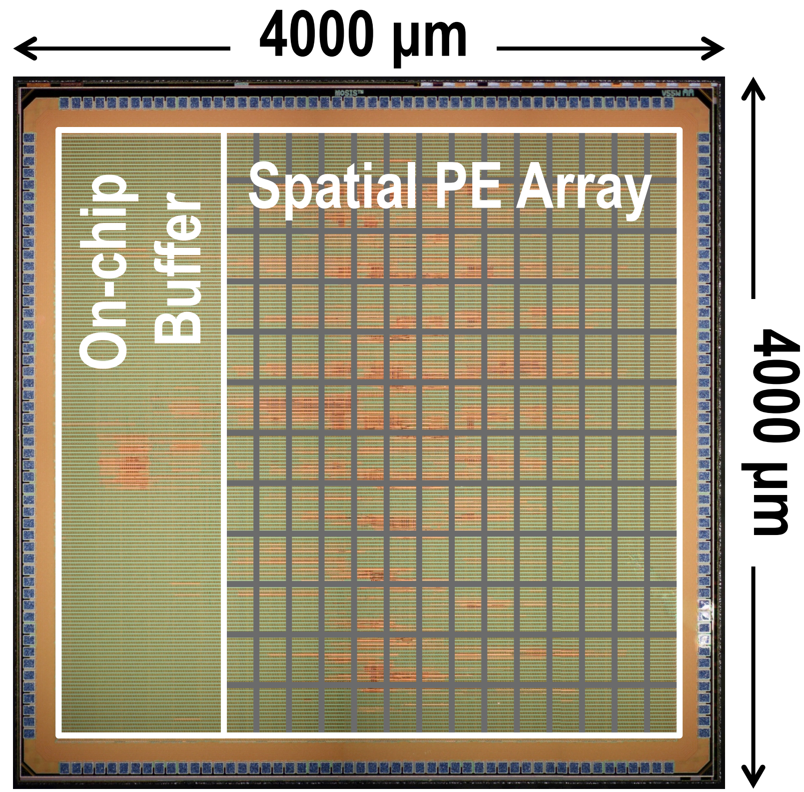
Data Gating / Zero Skipping



Results

Chip Spec & Measurement Results

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]



*** Natively Supported**

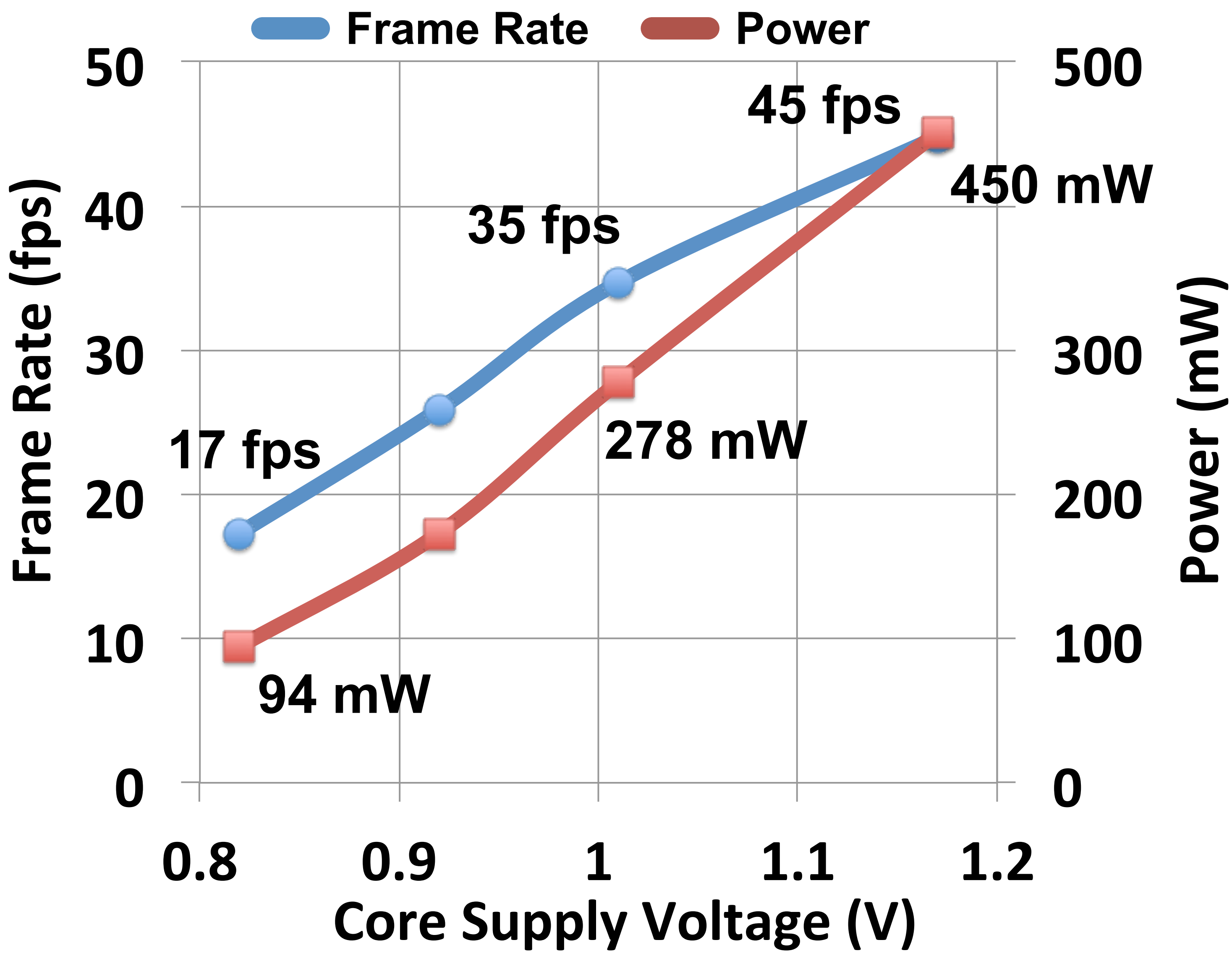
Benchmark – AlexNet Performance

Image Batch Size of **4** (i.e. 4 frames of 227x227)

Core Frequency = 200MHz / Link Frequency = 60 MHz

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**

AlexNet Throughput vs. Power



Comparison with GPU

	This Work	NVIDIA TK1 (Jetson Kit)
Technology	65nm	28nm
Clock Rate	200MHz	852MHz
# Multipliers	168	192
On-Chip Storage	Buffer: 108KB Spad: 75.3KB	Shared Mem: 64KB Reg File: 256KB
Word Bit-Width	16b Fixed	32b Float
Throughput[1]	34.7 fps	68 fps
Measured Power	278 mW	Idle/Active[2]: 3.7W/10.2W
DRAM Bandwidth	127 MB/s	1120 MB/s [3]

1. AlexNet Convolutional Layers Only
2. Board Power
3. Modeled from [Tan, SC11]

Demo (DS2 Today!)

Video Link: https://vimeo.com/154012013

AlexNet: [Krizhevsky, NIPS 2012]

Summary

- **A 278mW reconfigurable accelerator for state-of-the-art deep CNNs**
  – A 168-PE spatial architecture that supports an efficient dataflow to minimize data movement
  – A configurable multicast NoC that saves energy compared to a broadcast design
- **Exploits data statistics for higher efficiency**
  – Compression to reduce memory bandwidth
  – Zero-skipping logic to reduce PE power
- **Integrated with the Caffe DL framework and demonstrated an image classification system**

Acknowledgement: funding by DARPA YFA, MIT CICS and a gift from Intel

INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS

ISSCC®

UNIVERSITY OF PENNSYLVANIA