DNN Accelerator Architectures (cont.)

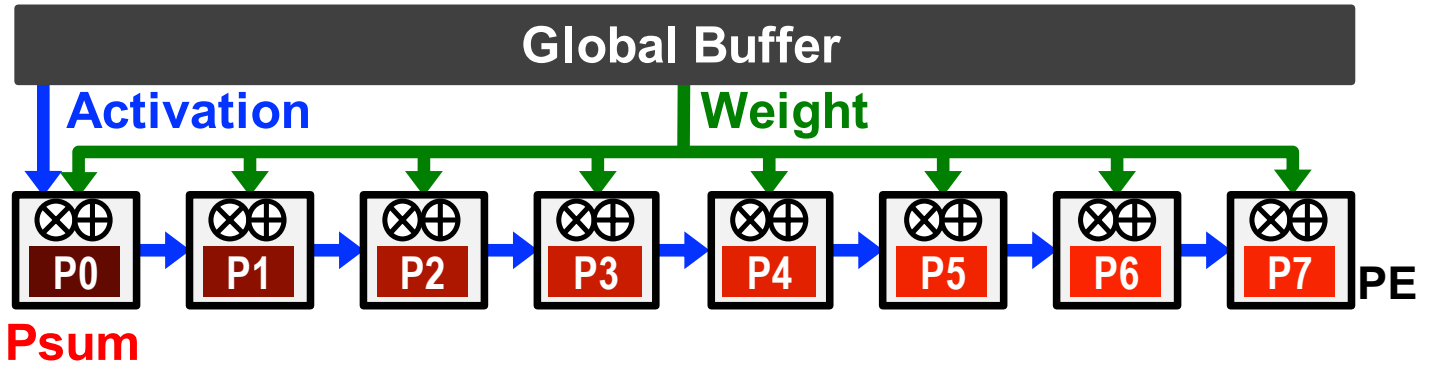
ISCA Tutorial (2019)

Website: http://eyeriss.mit.edu/tutorial.html

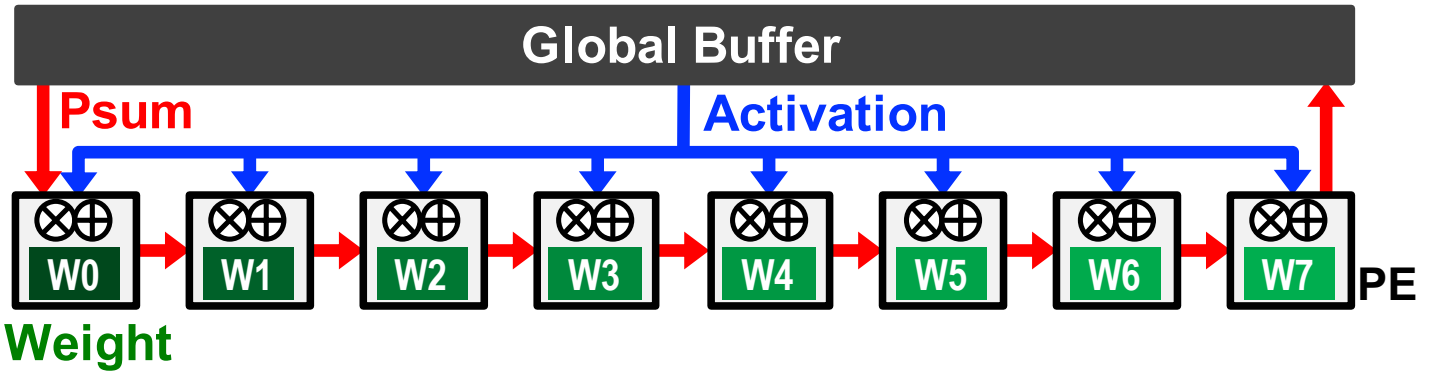
Joel Emer, Vivienne Sze, Yu-Hsin Chen

Recap on Dataflow Taxonomy

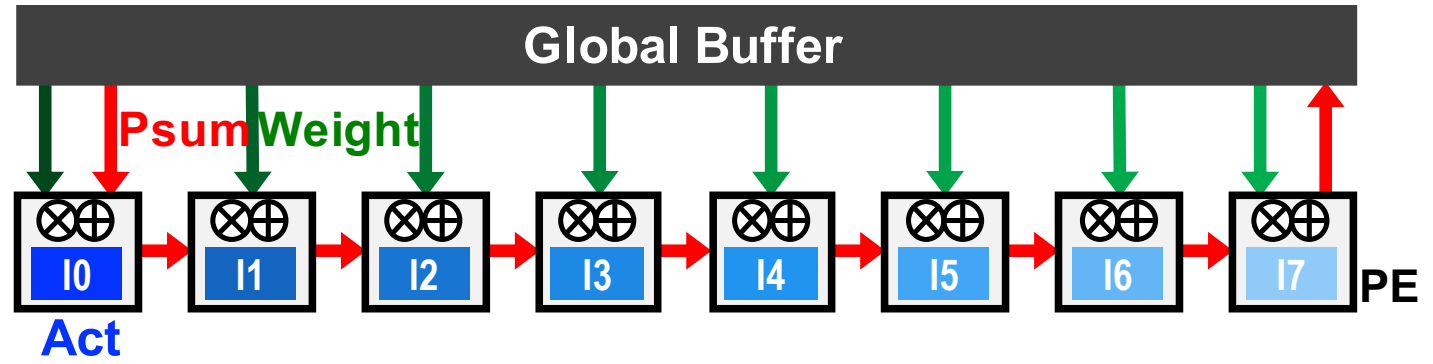
Output Stationary



Weight Stationary



Input Stationary

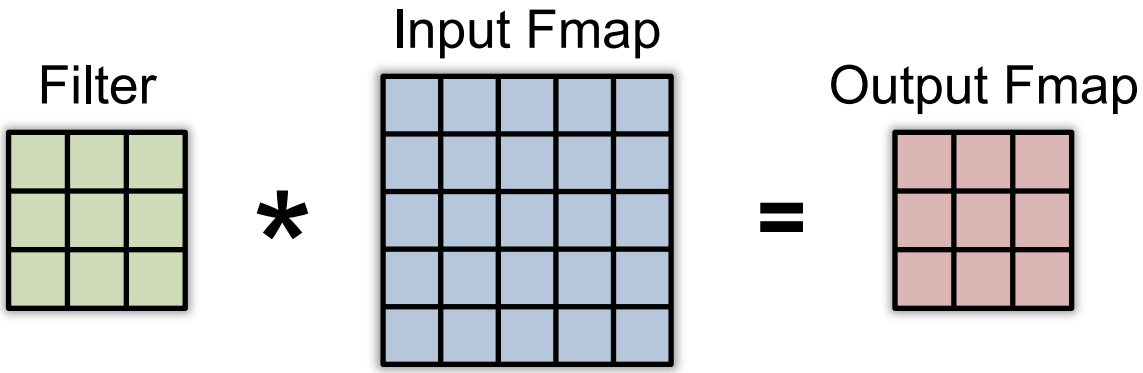


Energy-Efficient Dataflow: Row Stationary (RS)

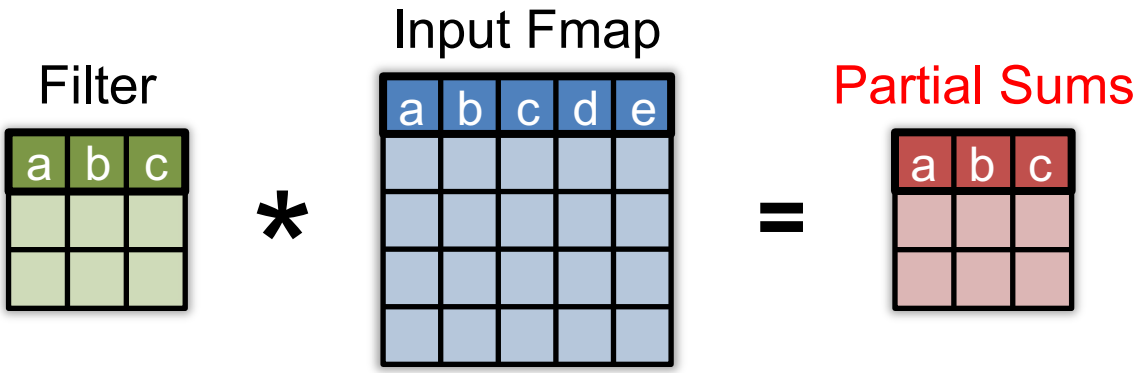
- **Maximize** data reuse at **RF**
- Optimize for **overall** energy efficiency instead for *only* a certain data type

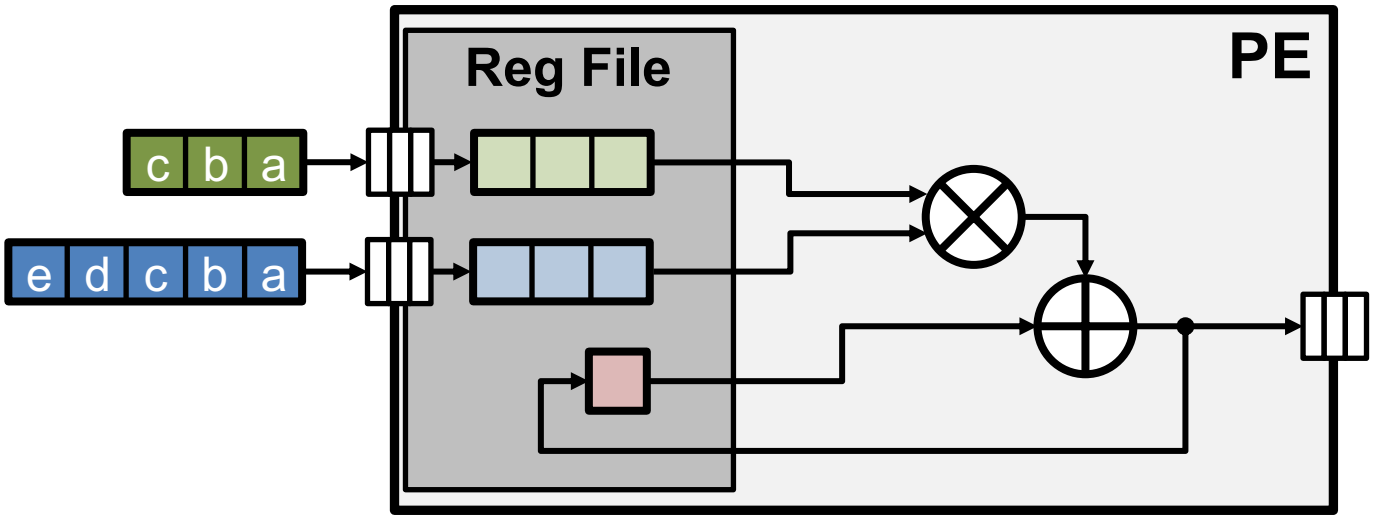
[Chen et al., ISCA 2016]

Row Stationary: Energy-efficient Dataflow

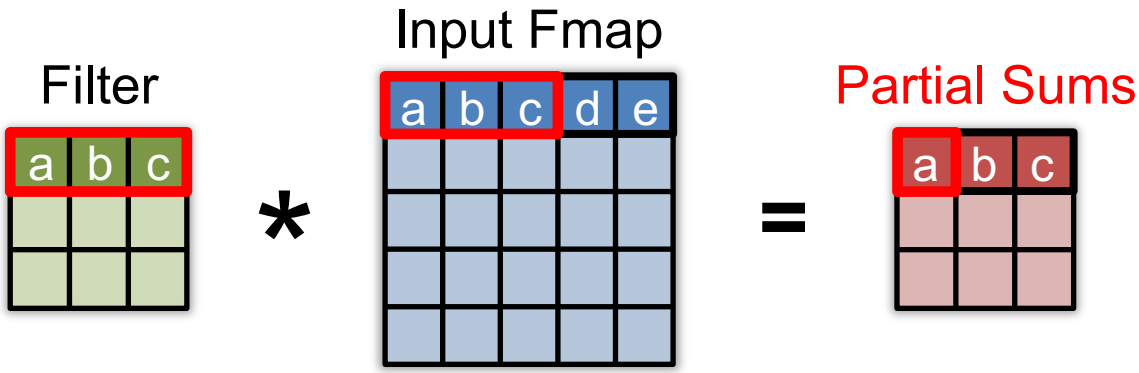


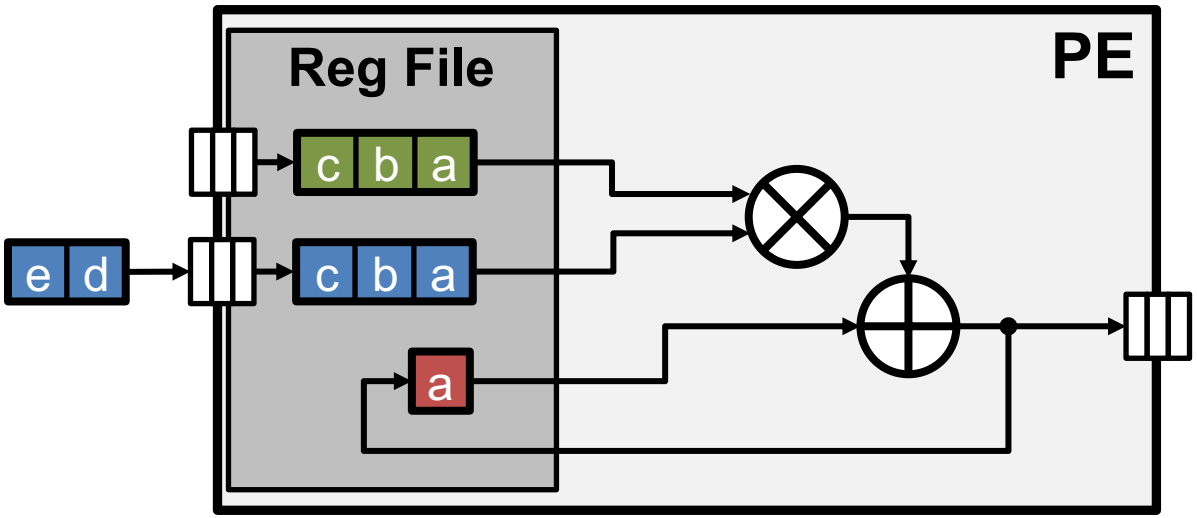
1D Row Convolution in PE



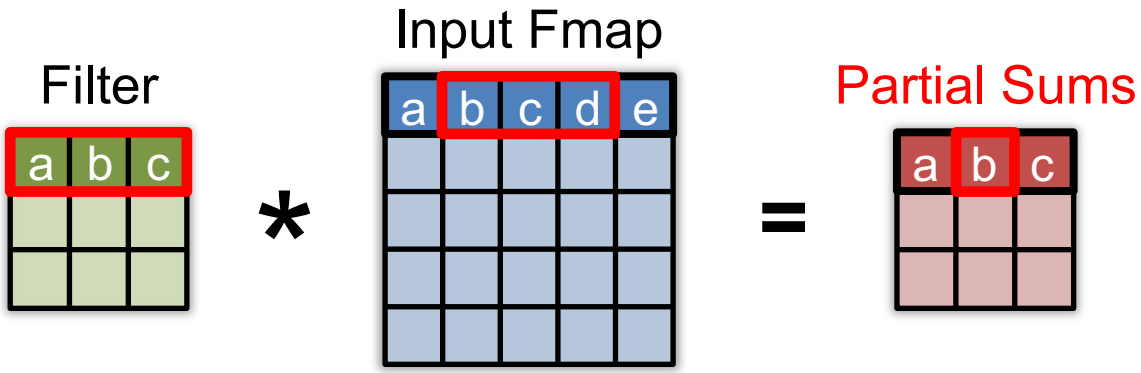


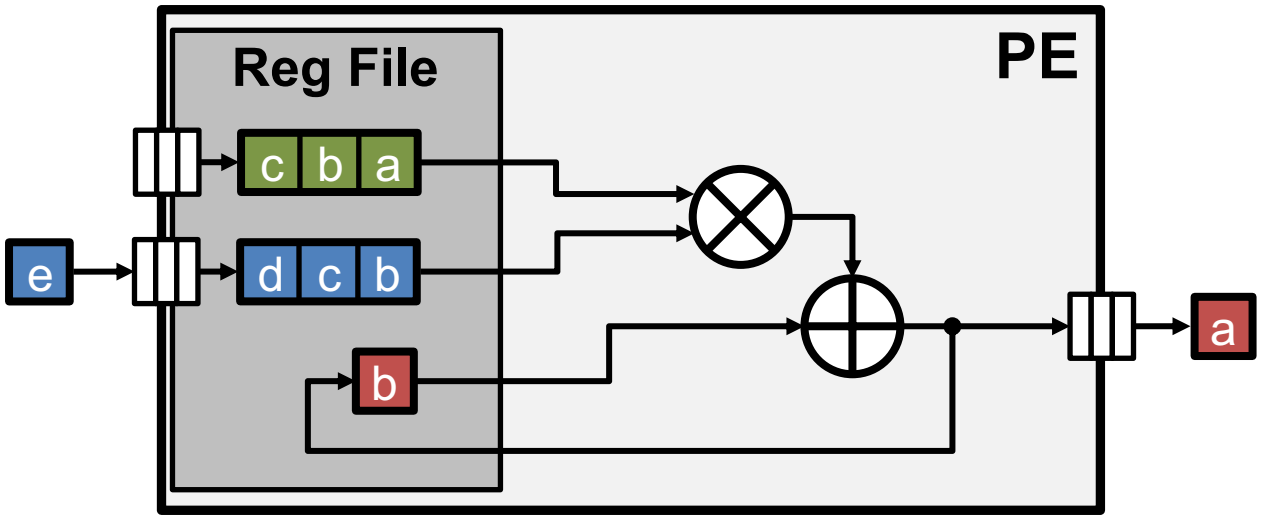
1D Row Convolution in PE



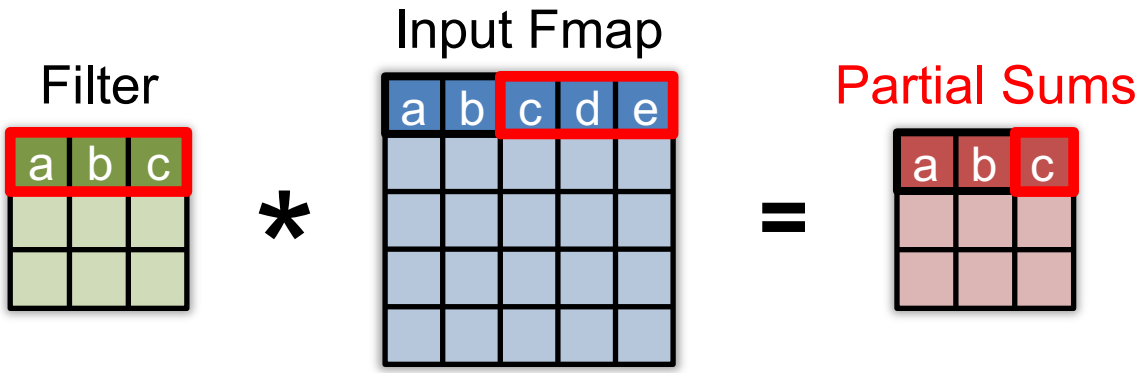


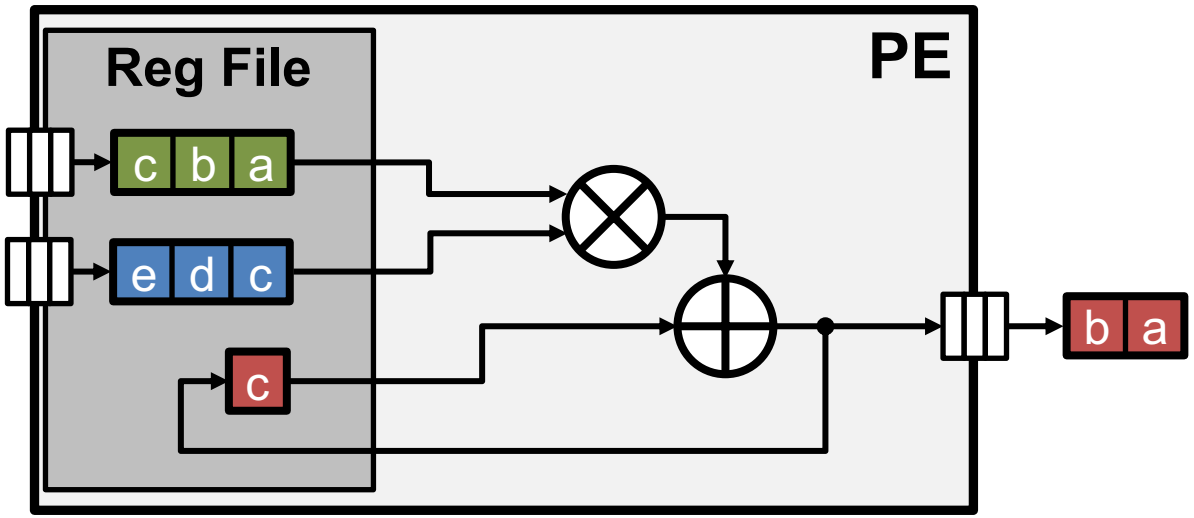
1D Row Convolution in PE





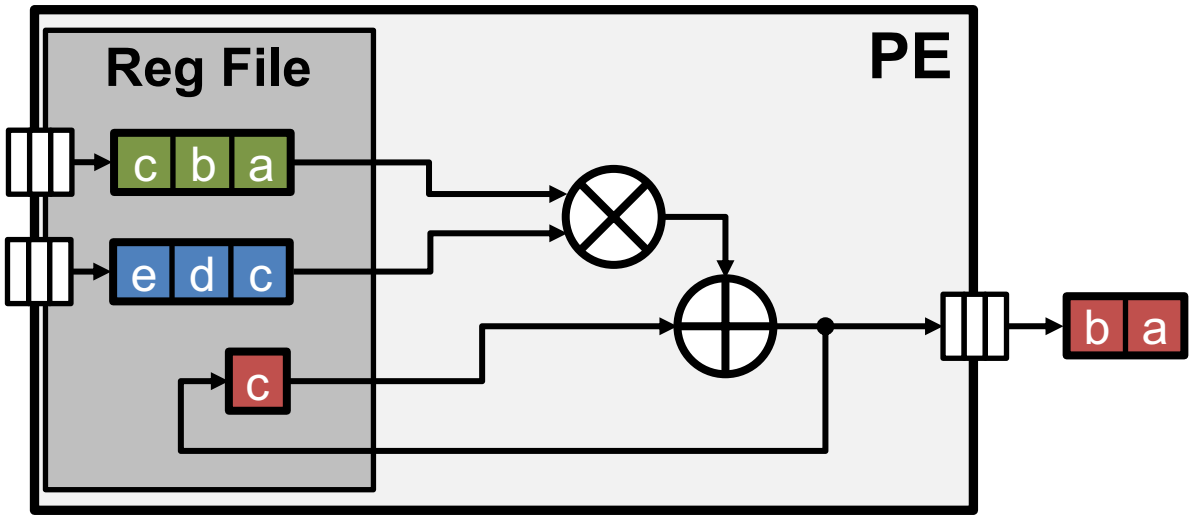
1D Row Convolution in PE



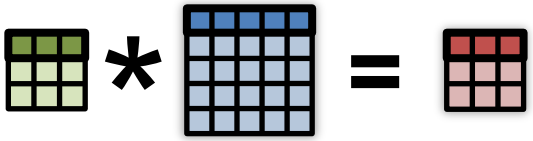


1D Row Convolution in PE

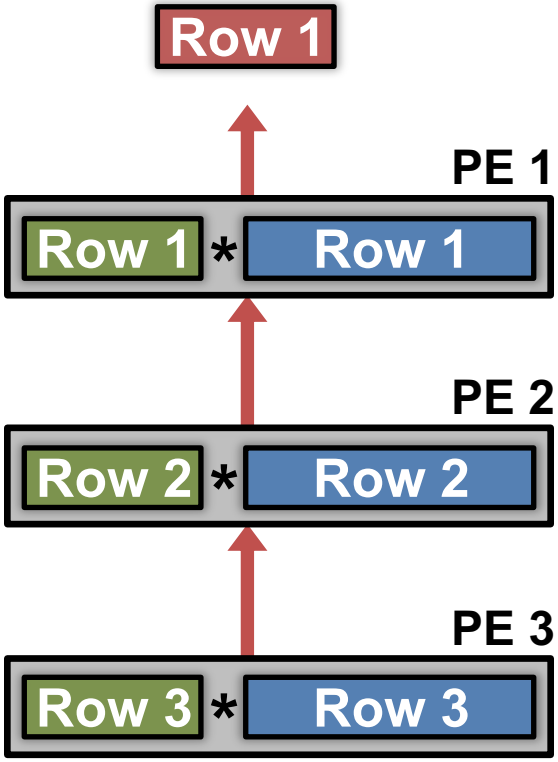
- Maximize row **convolutional reuse** in RF
  - Keep a **filter** row and **fmap** sliding window in RF
- Maximize row **psum accumulation** in RF

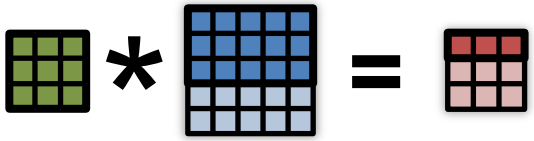


2D Convolution in PE Array

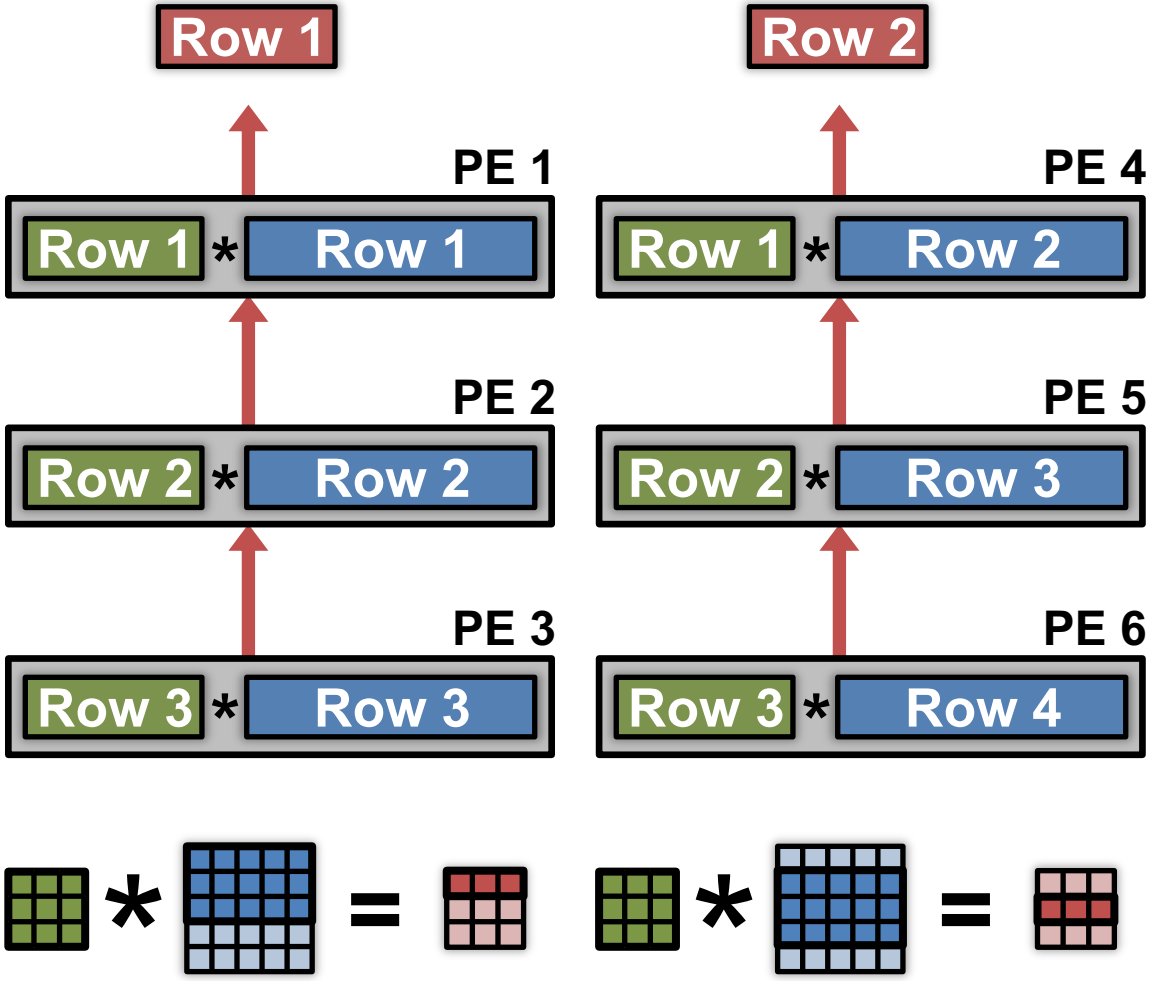


2D Convolution in PE Array

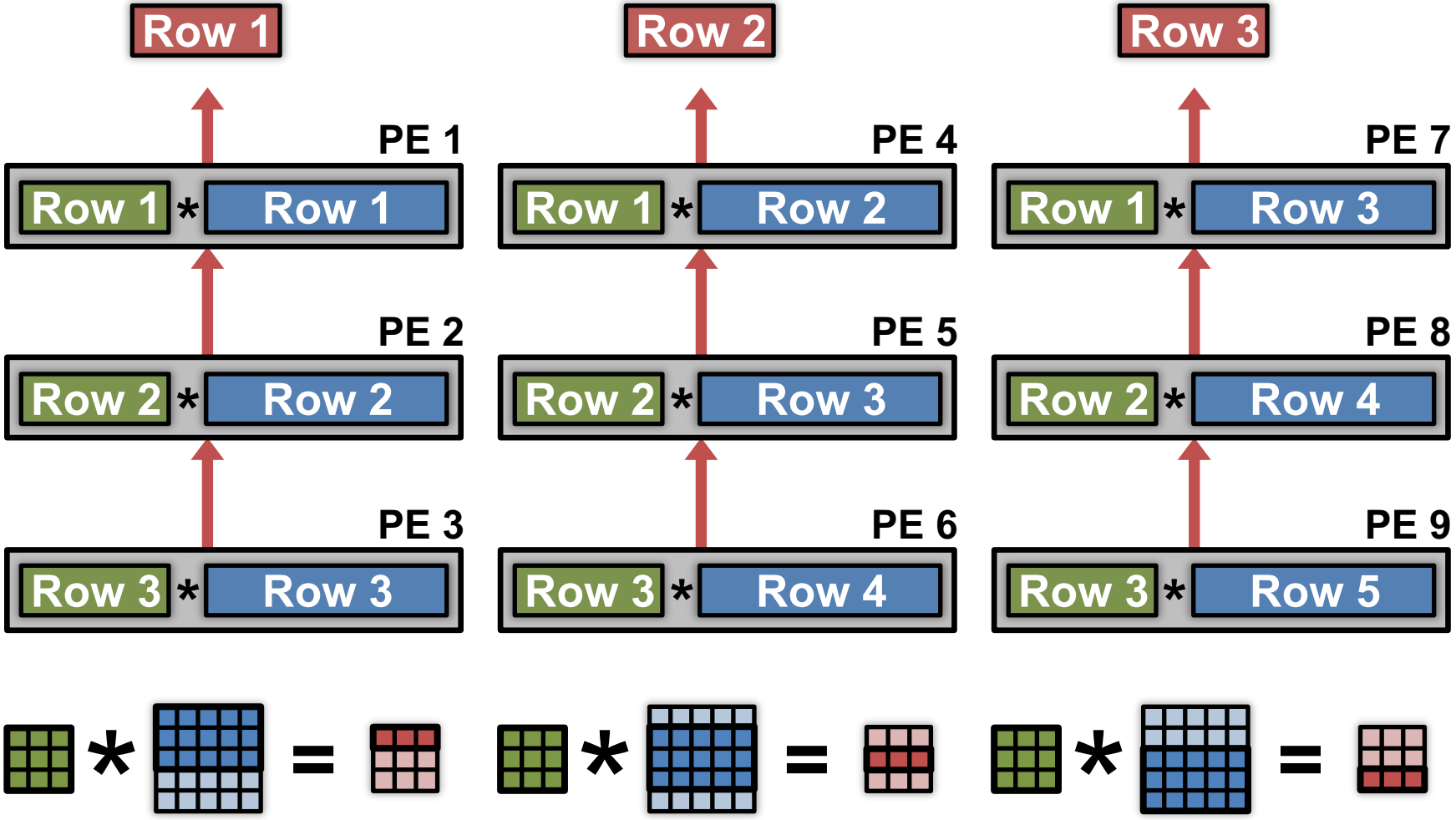




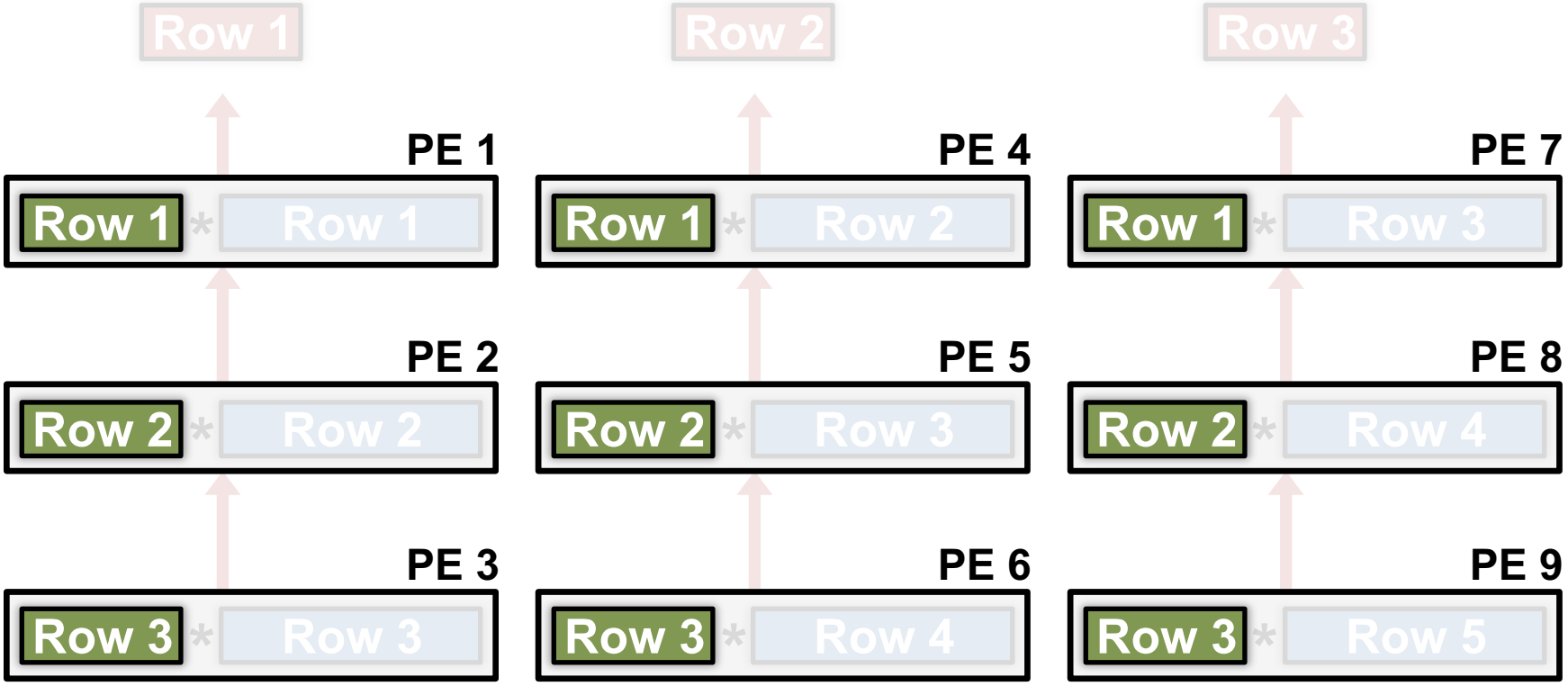
2D Convolution in PE Array



2D Convolution in PE Array

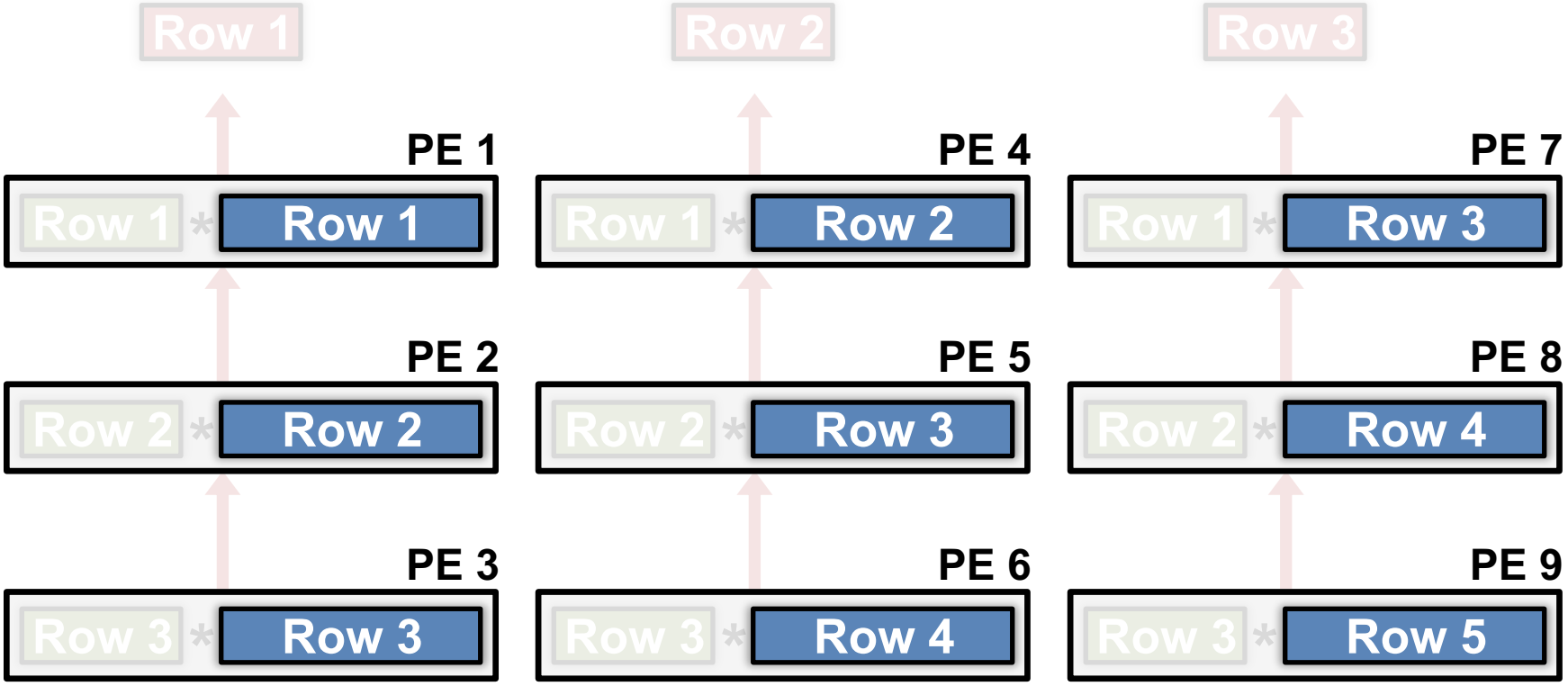


Convolutional Reuse Maximized



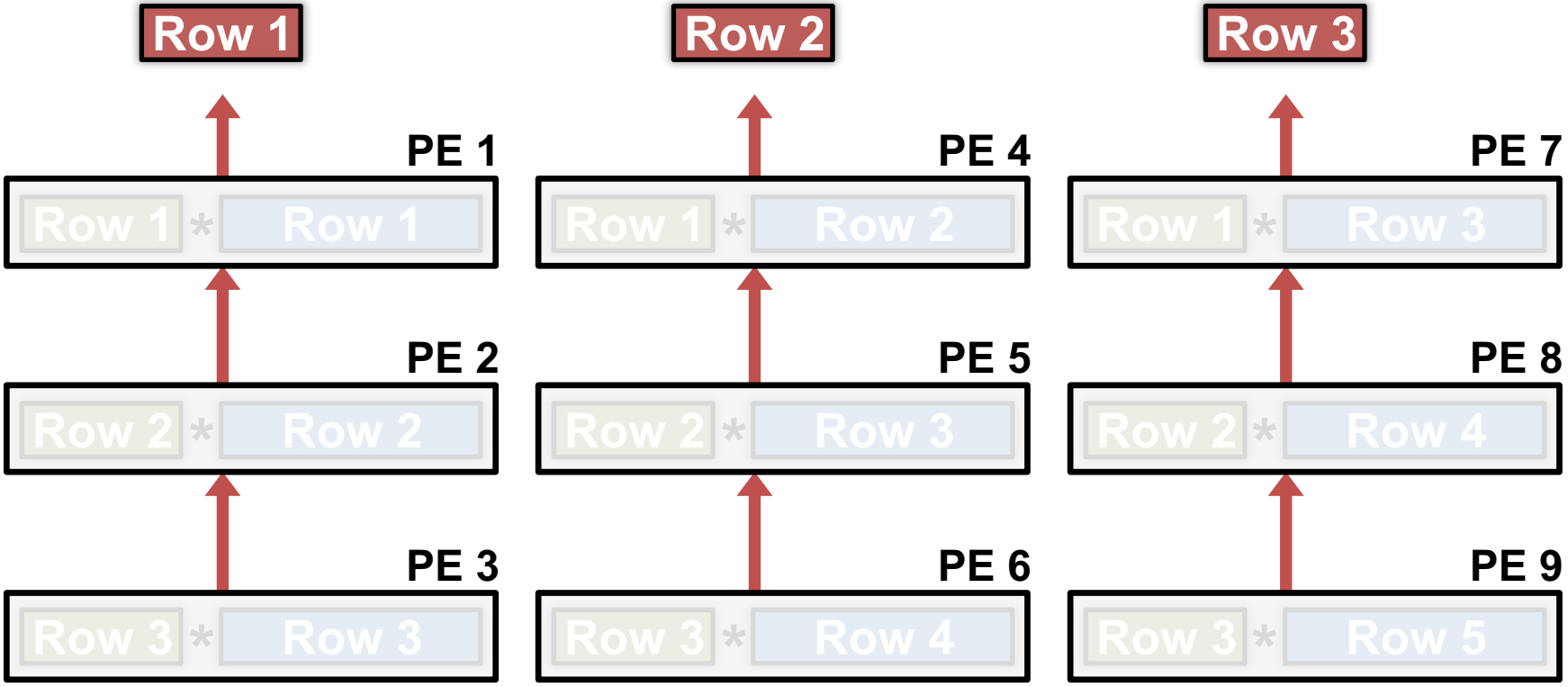
Filter rows are reused across PEs **horizontally**

Convolutional Reuse Maximized



Fmap rows are reused across PEs **diagonally**

Maximize 2D Accumulation in PE Array



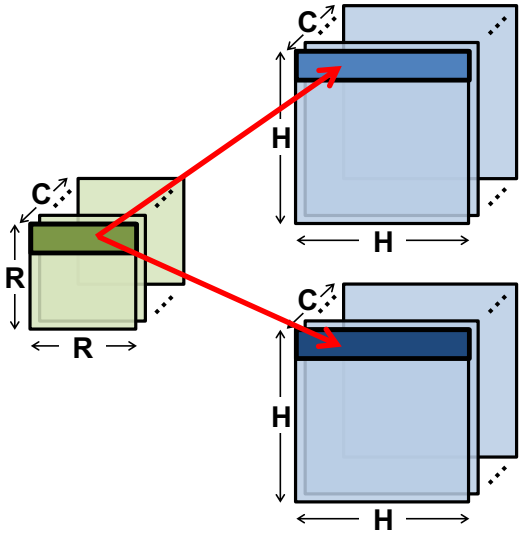
Partial sums accumulate across PEs **vertically**

Dimensions Beyond 2D Convolution

1. Multiple Fmaps
2. Multiple Filters
3. Multiple Channels

Filter Reuse in PE

1 Multiple Fmaps

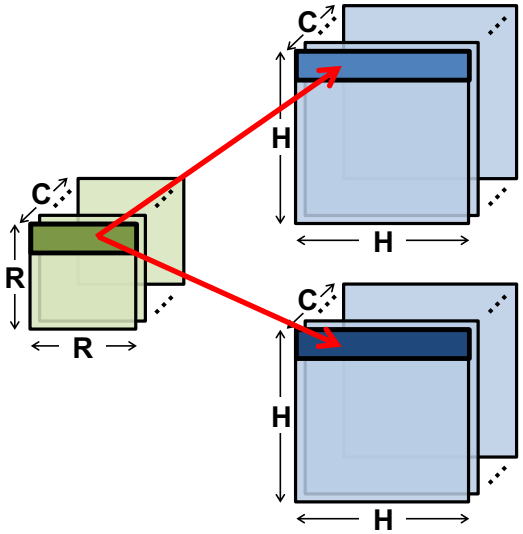


	Filter 1		Fmap 1		Psum 1
Channel 1	Row 1	*	Row 1	=	Row 1

	Filter 1		Fmap 2		Psum 2
Channel 1	Row 1	*	Row 1	=	Row 1

Filter Reuse in PE

1 Multiple Fmaps



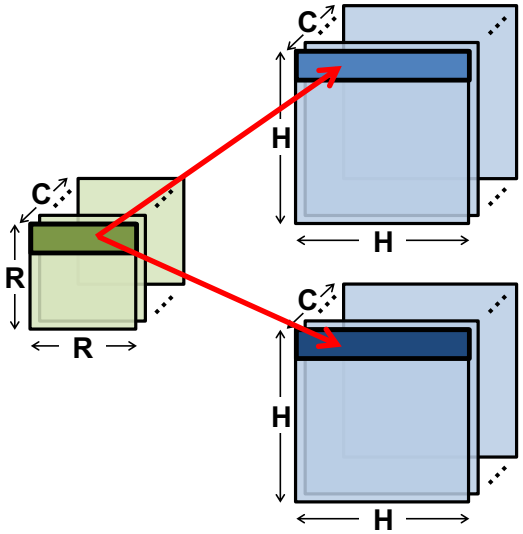
	Filter 1		Fmap 1		Psum 1
Channel 1	Row 1	*	Row 1	=	Row 1

	Filter 1		Fmap 2		Psum 2
Channel 1	Row 1	*	Row 1	=	Row 1

share the same filter row

Filter Reuse in PE

1 Multiple Fmaps



	Filter 1		Fmap 1		Psum 1
Channel 1	Row 1	*	Row 1	=	Row 1

	Filter 1		Fmap 2		Psum 2
Channel 1	Row 1	*	Row 1	=	Row 1

share the same filter row

2 Multiple Filters

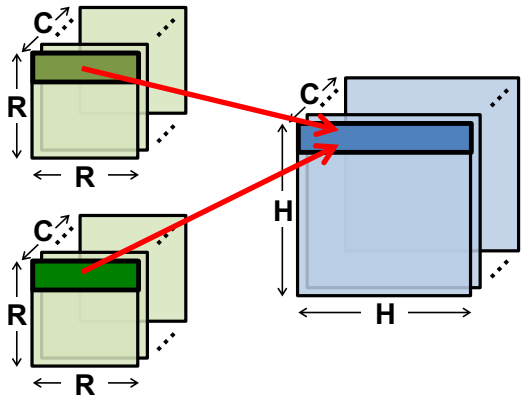
3 Multiple Channels

Processing in PE: **concatenate fmap rows**

	Filter 1		Fmap 1 & 2		Psum 1 & 2
Channel 1	Row 1	*	Row 1 Row 1	=	Row 1 Row 1

Fmap Reuse in PE

1 Multiple Fmaps **2 Multiple Filters** 3 Multiple Channels

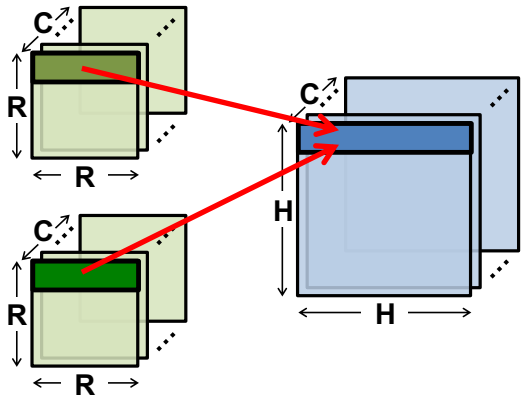


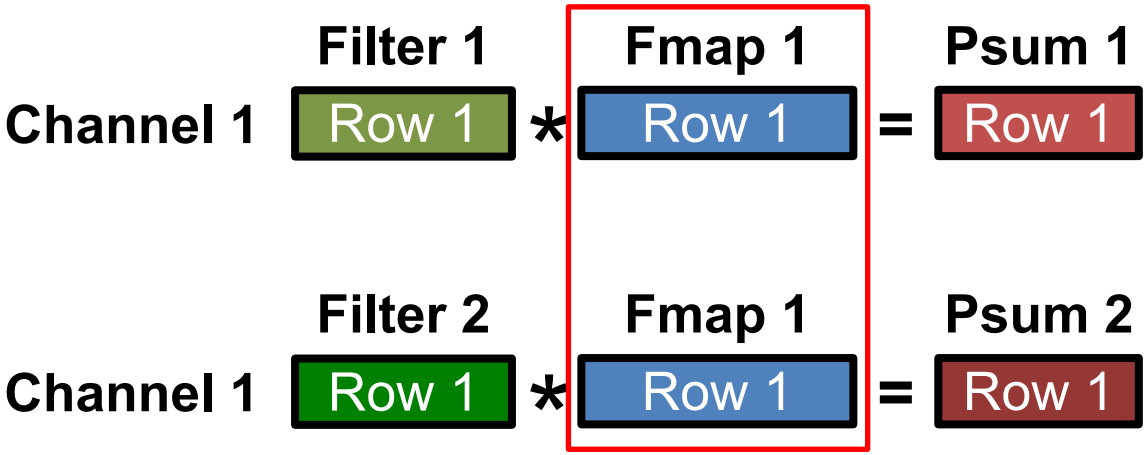
	Filter 1		Fmap 1		Psum 1
Channel 1	Row 1	*	Row 1	=	Row 1

	Filter 2		Fmap 1		Psum 2
Channel 1	Row 1	*	Row 1	=	Row 1

Fmap Reuse in PE

1 Multiple Fmaps **2 Multiple Filters** 3 Multiple Channels

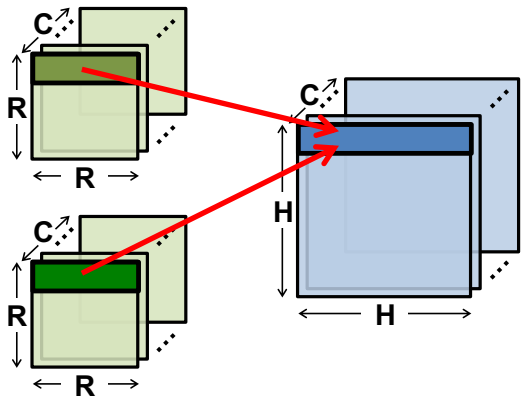


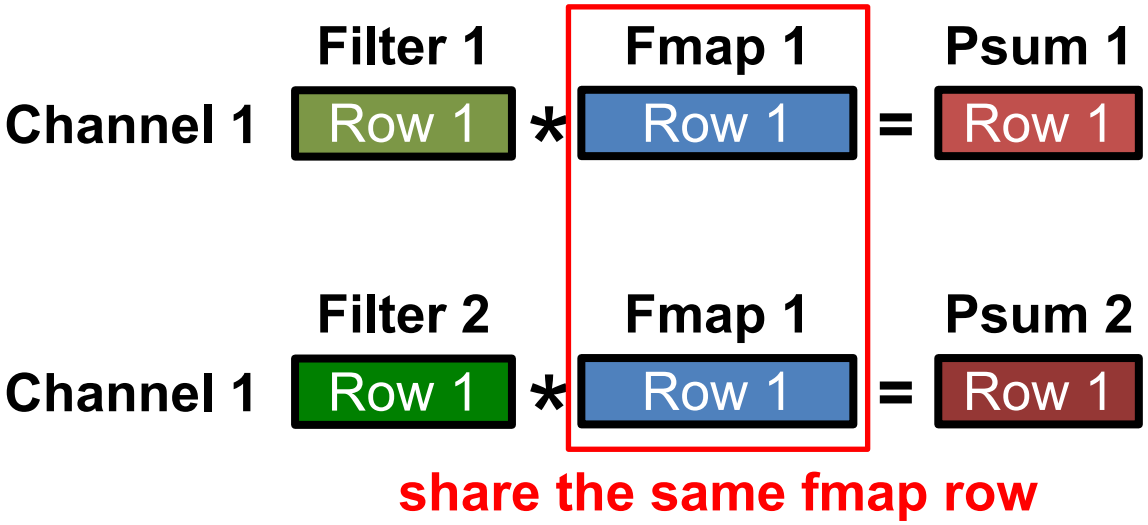


share the same fmap row

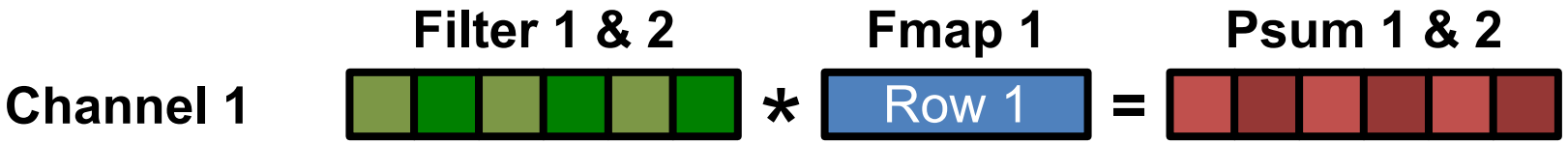
Fmap Reuse in PE

1 Multiple Fmaps **2 Multiple Filters** 3 Multiple Channels



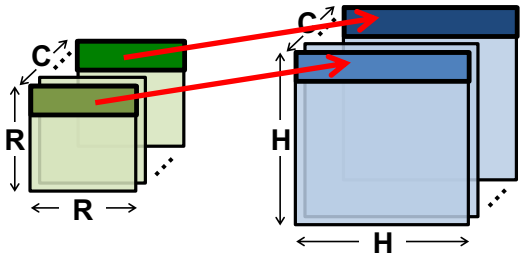


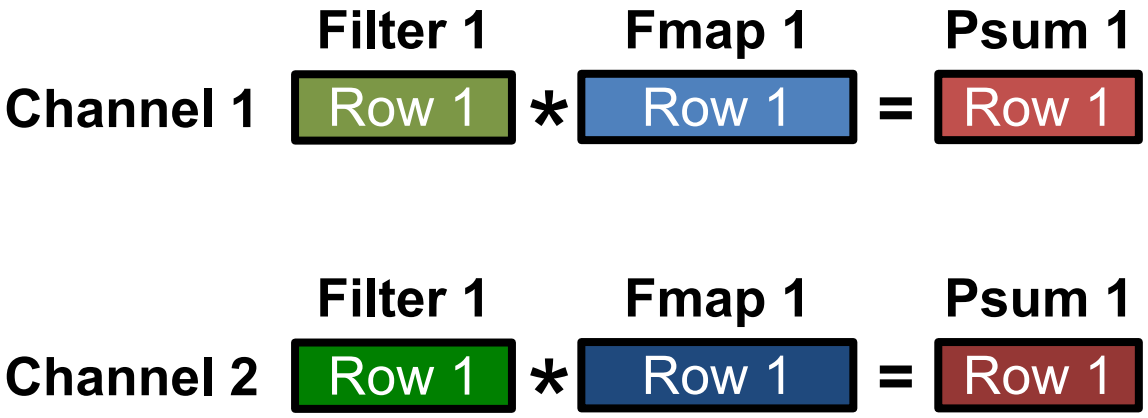
Processing in PE: **interleave filter rows**



Channel Accumulation in PE

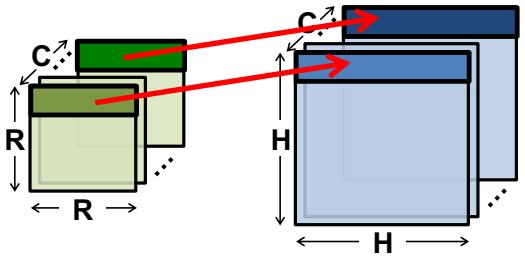
1 Multiple Fmaps 2 Multiple Filters **3 Multiple Channels**

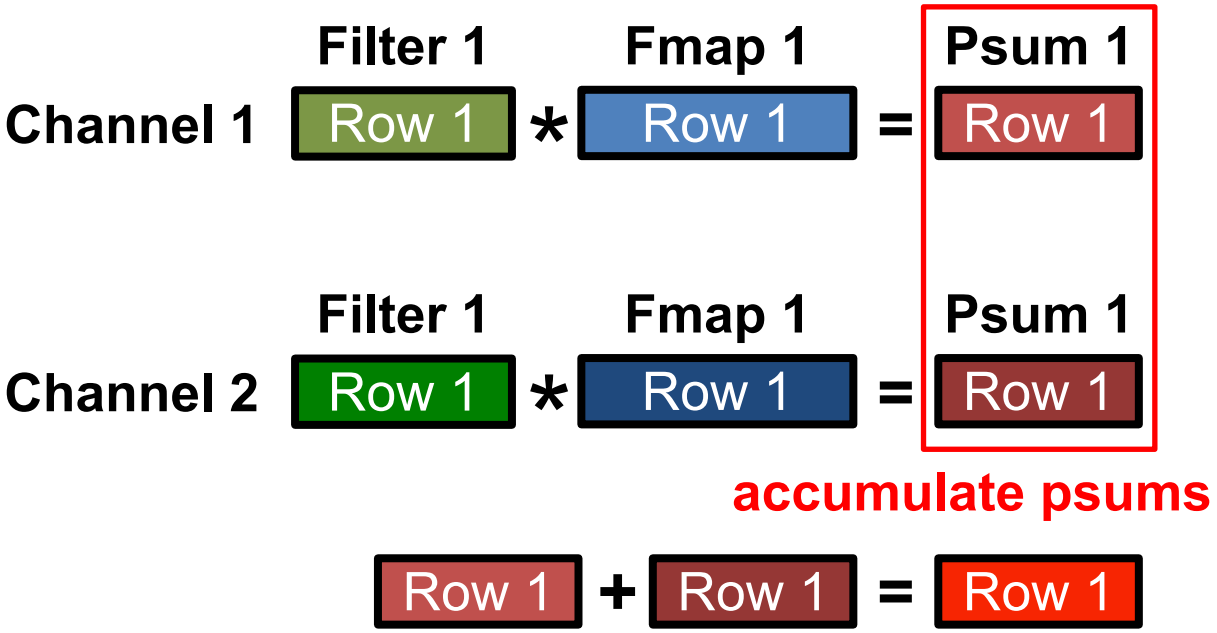




Channel Accumulation in PE

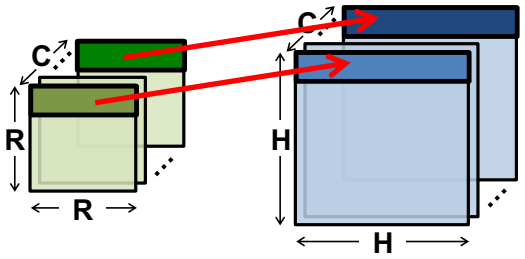
1 Multiple Fmaps 2 Multiple Filters **3 Multiple Channels**

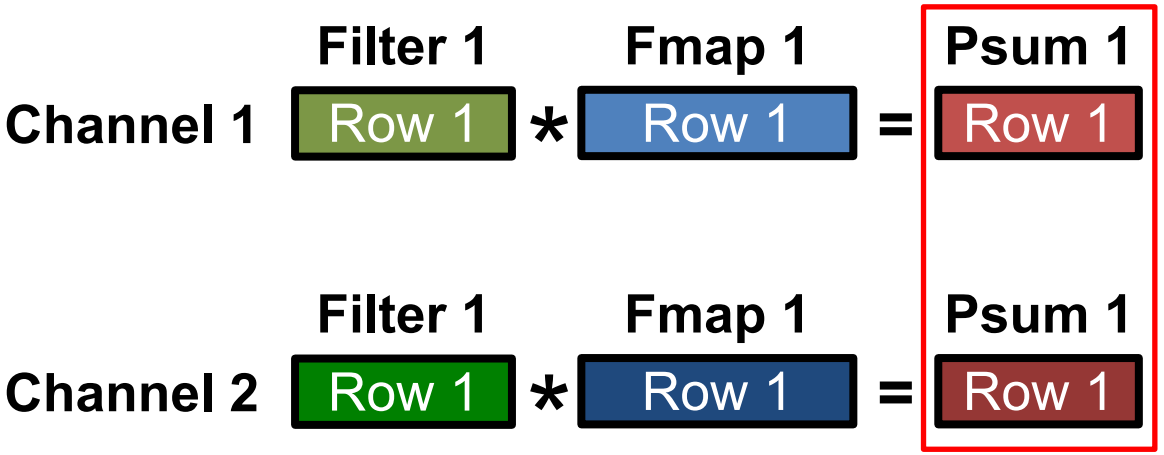




Channel Accumulation in PE

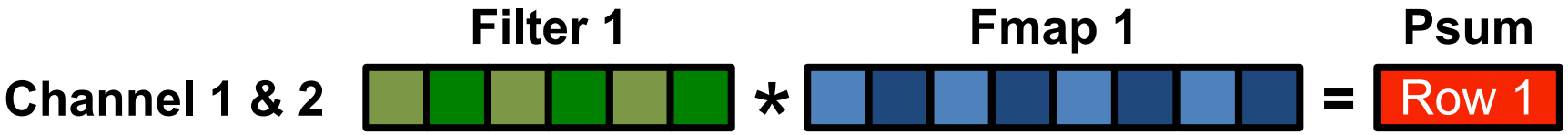
1 Multiple Fmaps 2 Multiple Filters **3 Multiple Channels**



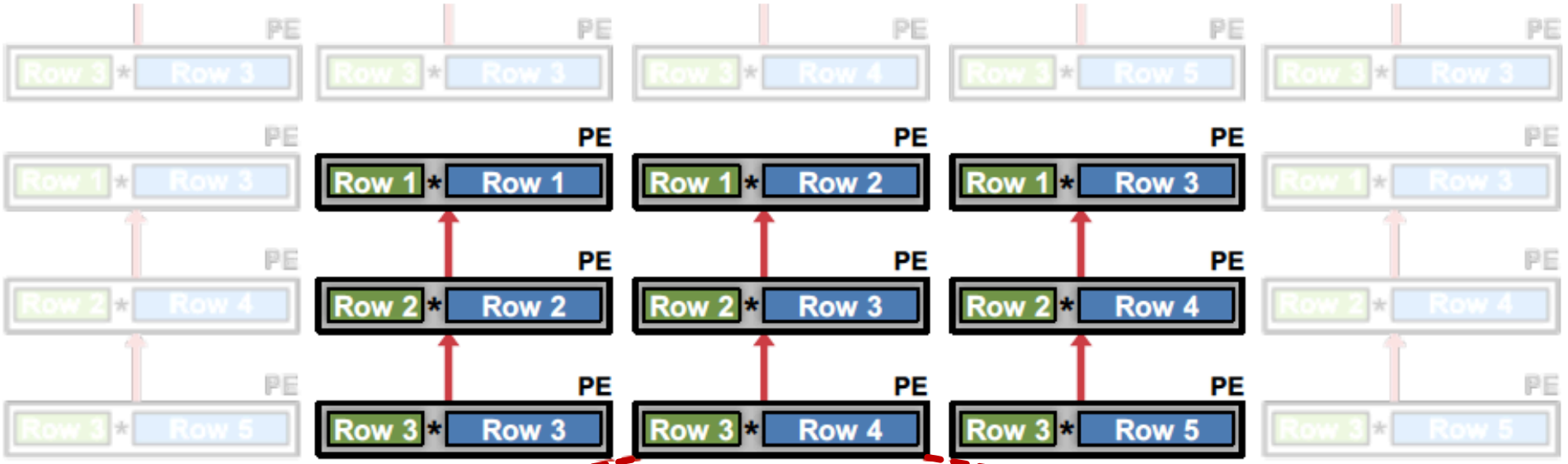


accumulate psums

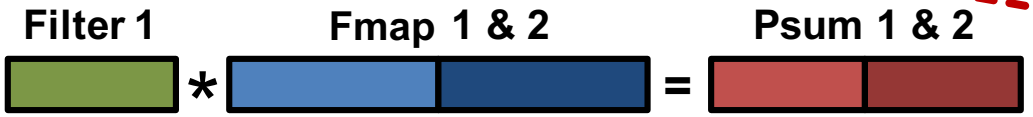
Processing in PE: **interleave channels**



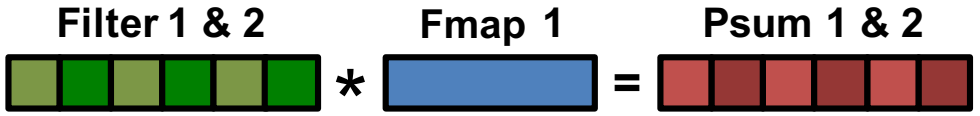
DNN Processing – The Full Picture



Multiple **fmaps**:



Multiple **filters**:

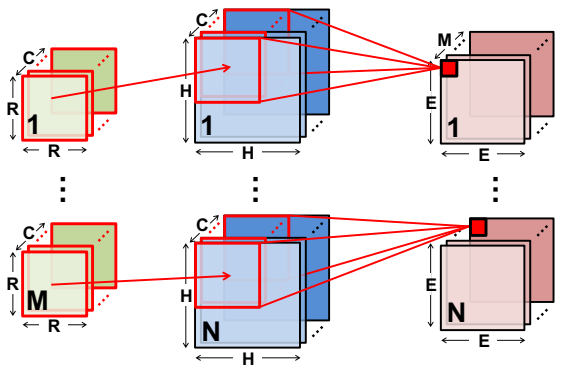


Multiple **channels**:

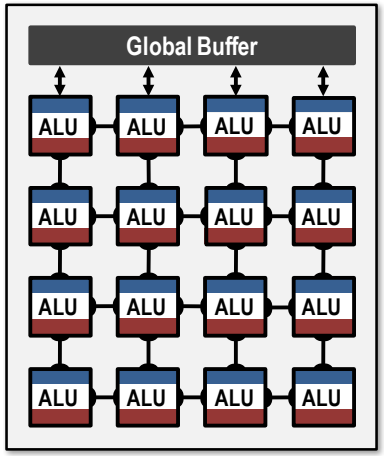
Map rows from **multiple fmaps, filters** and **channels** to same PE to exploit other forms of reuse and local accumulation

Optimal Mapping in Row Stationary

DNN Configurations

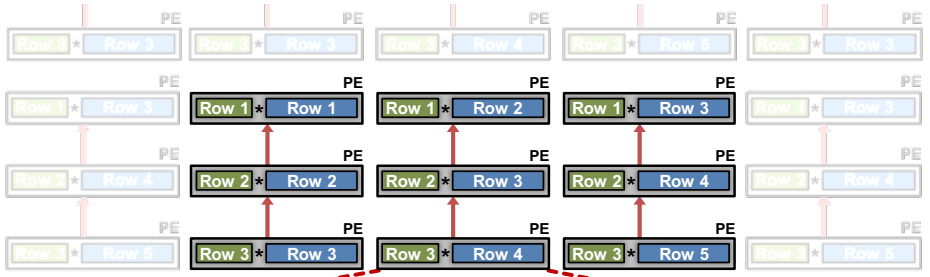


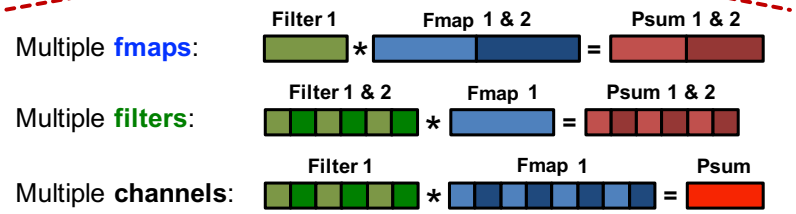
Hardware Resources



Optimization Compiler (Mapper)

Row Stationary Mapping

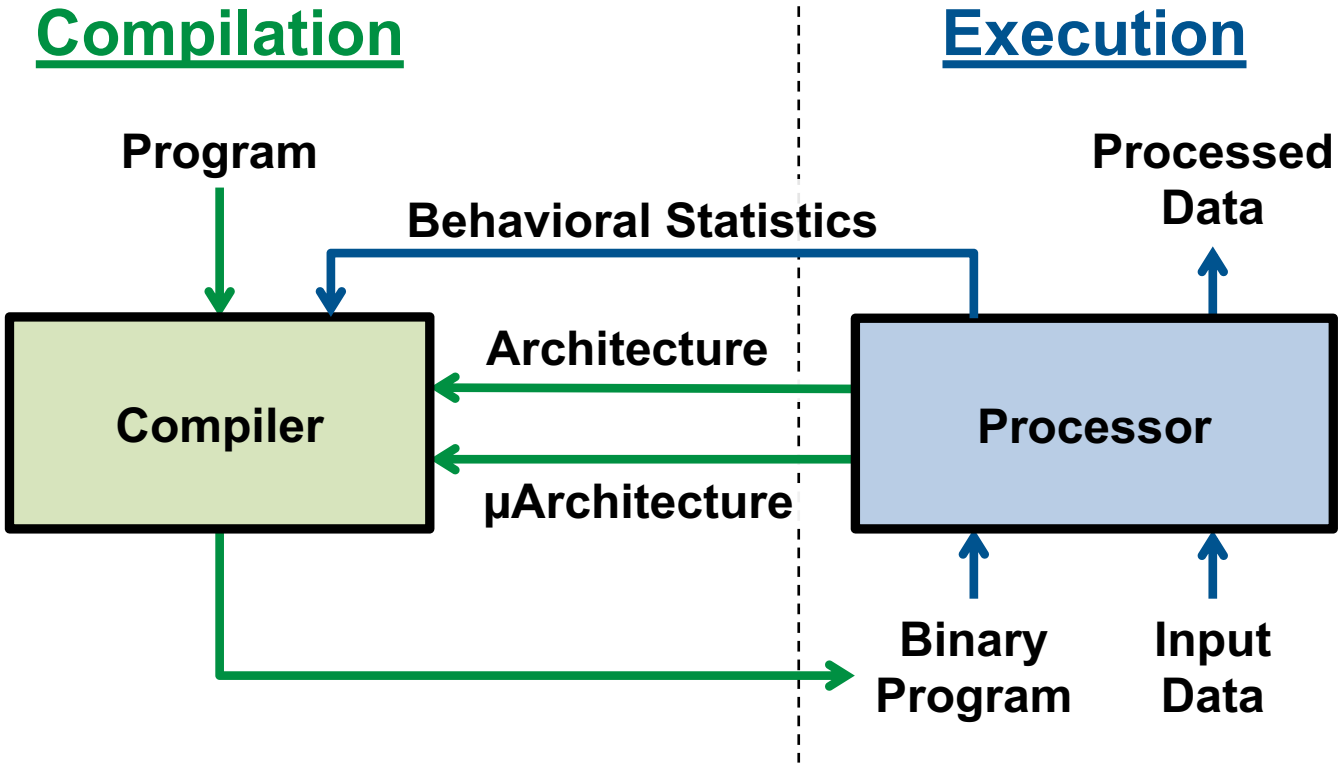




[Chen et al., ISCA 2016]

Computer Architecture Analogy

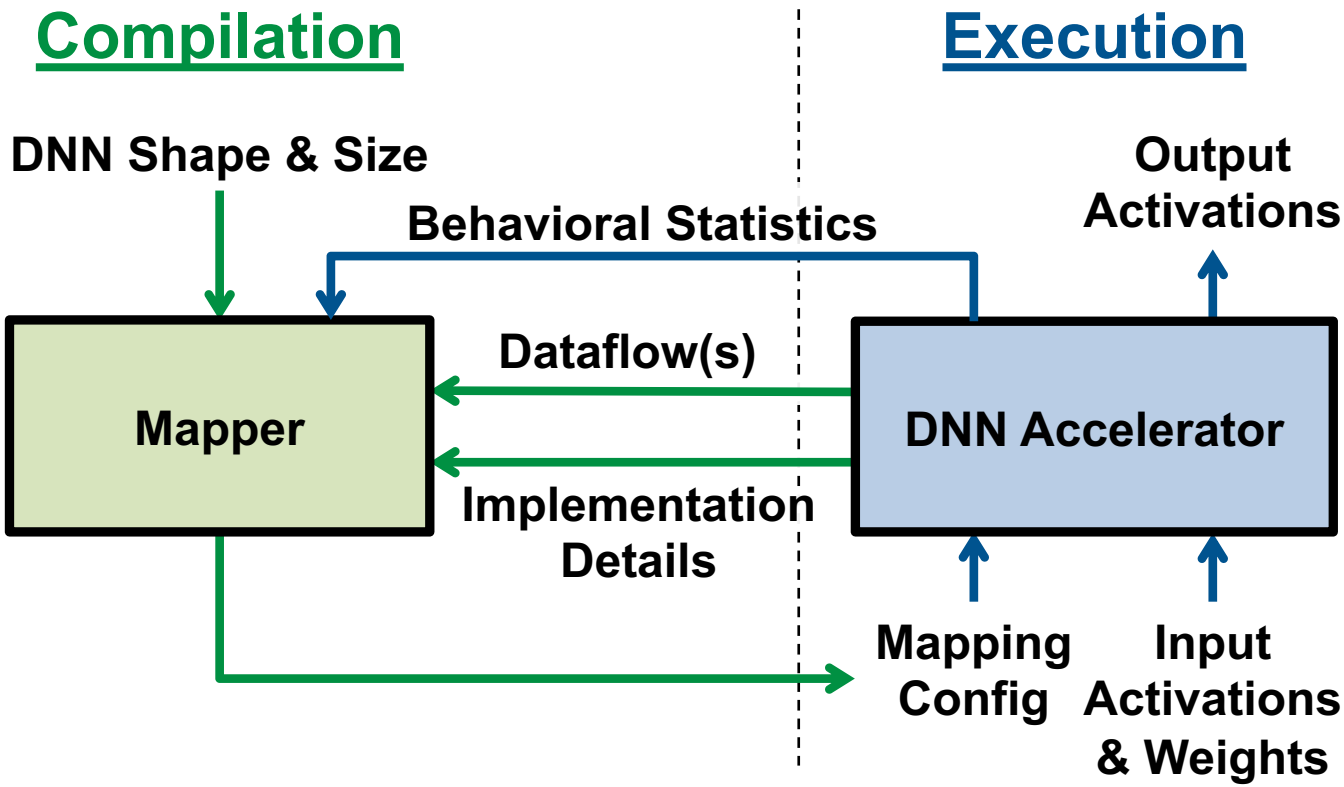
CPU Compute Model



[Chen et al., Micro Top-Picks 2017]

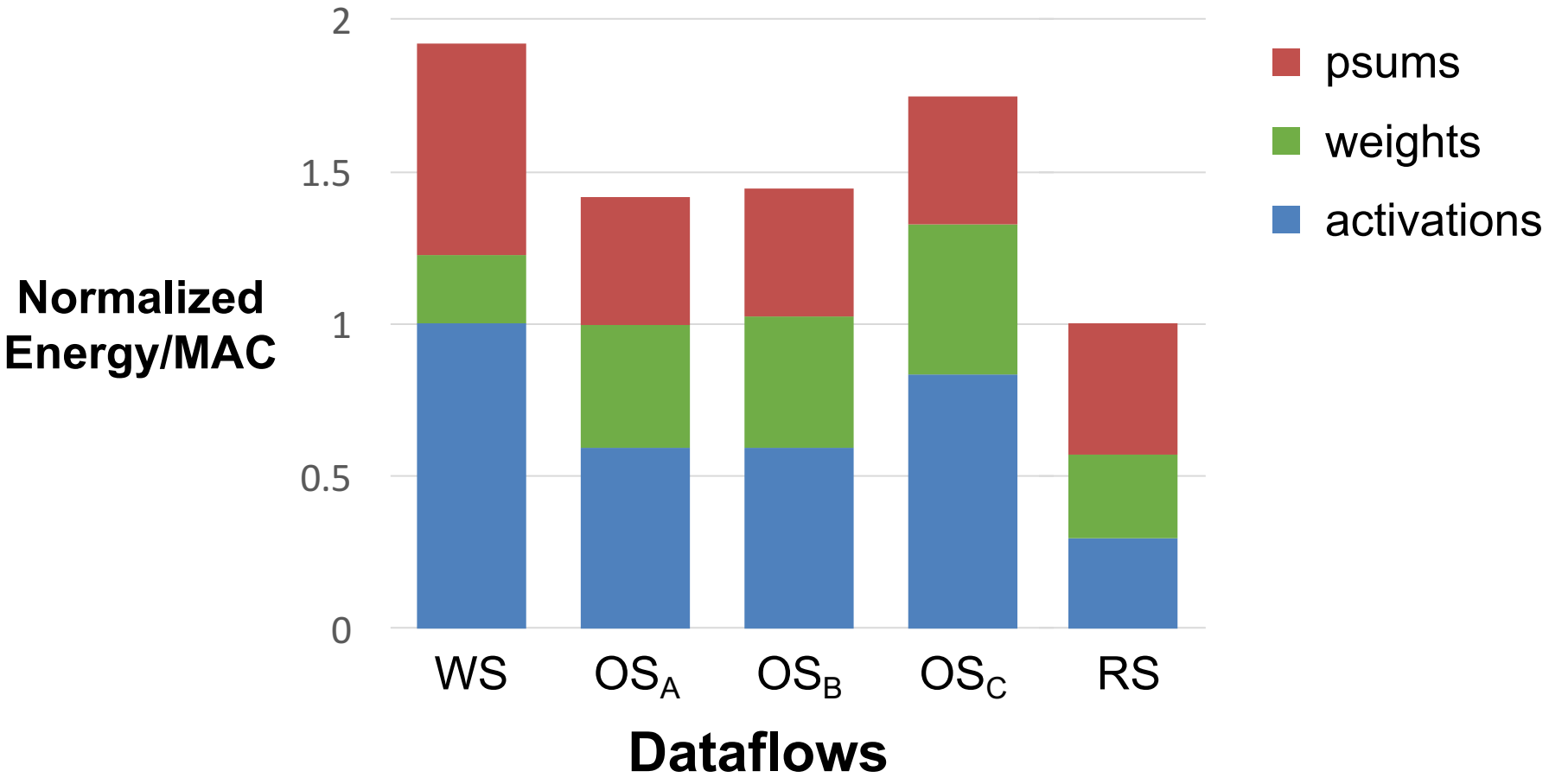
Computer Architecture Analogy

DNN Compute Model



[Chen et al., Micro Top-Picks 2017]

Dataflow Comparison



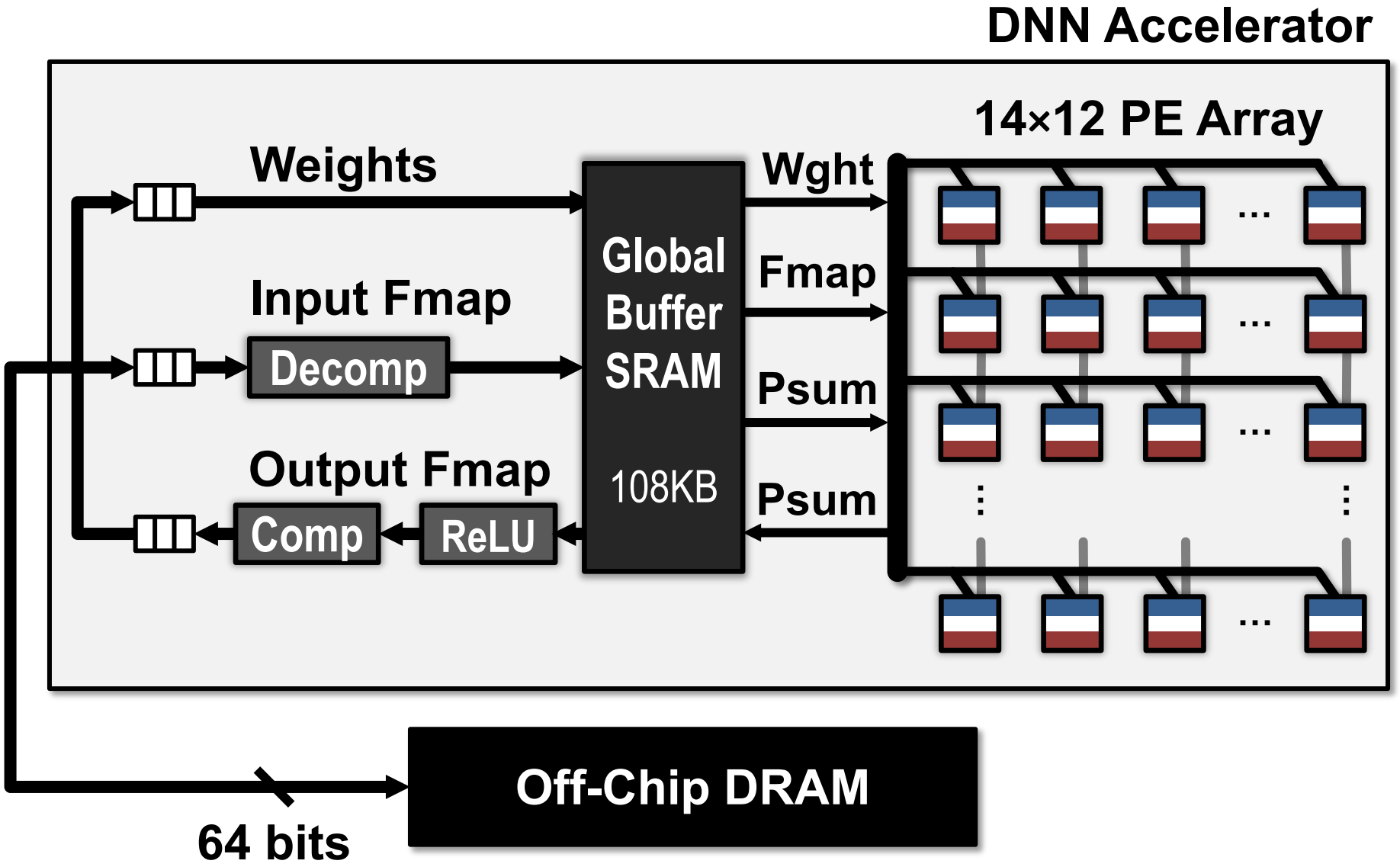
RS optimizes for the best **overall** energy efficiency

[Chen et al., ISCA 2016]

Hardware Architecture for RS Dataflow

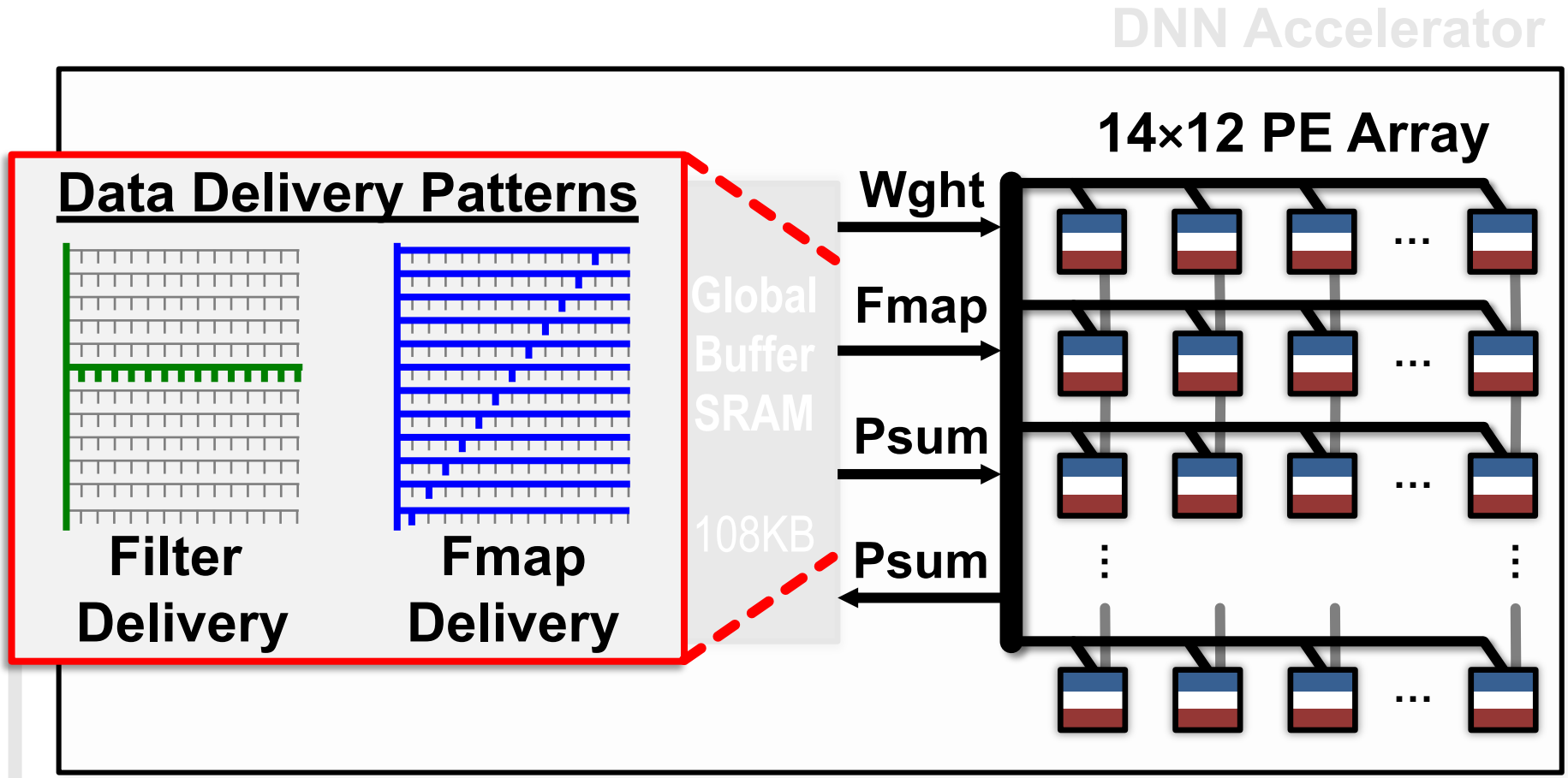
[Chen et al., ISSCC 2016]

Eyeriss DNN Accelerator



[Chen et al., ISSCC 2016]

Data Delivery with On-Chip Network

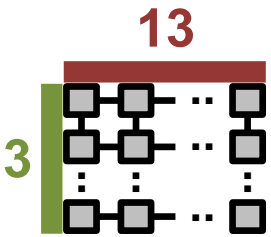


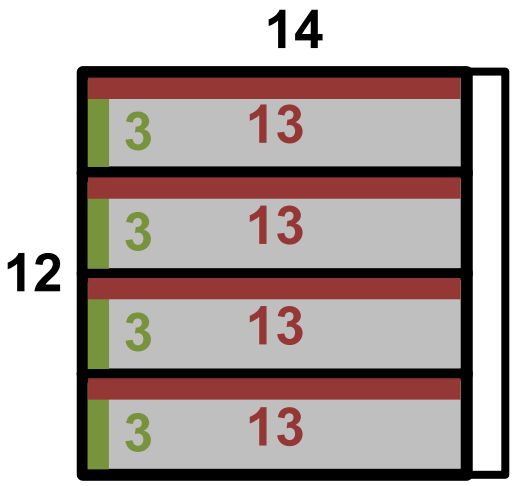
Multicast NoC to support any delivery patterns

Logical to Physical Mappings

Replication

AlexNet
Layer 3-5

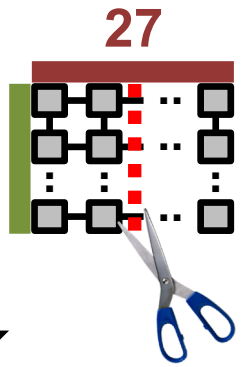


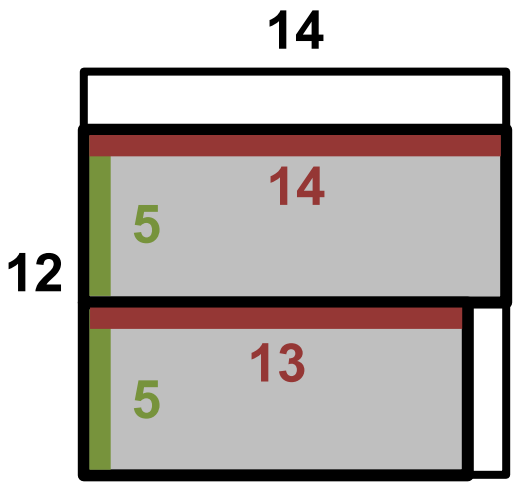


Physical PE Array

Folding

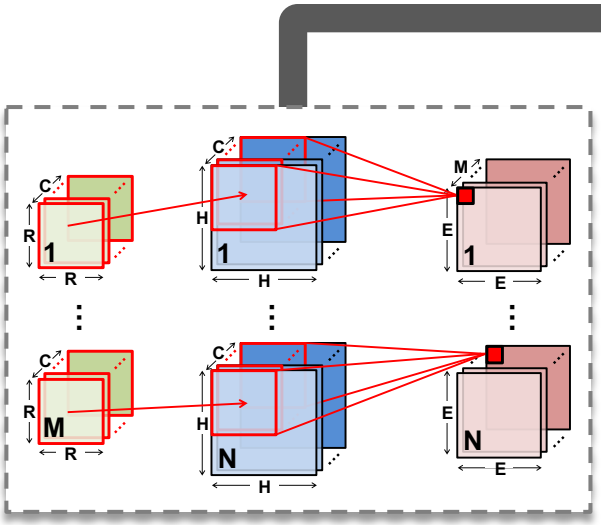
AlexNet
Layer 2



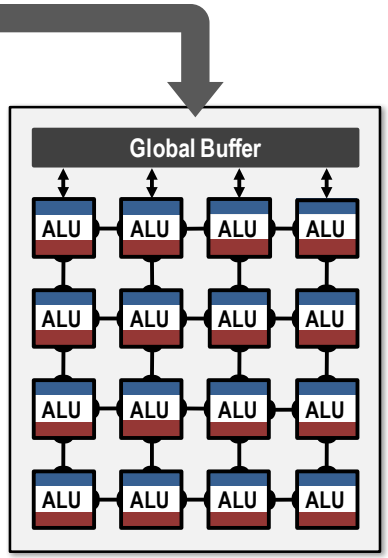


Physical PE Array

Hardware Performance Analysis

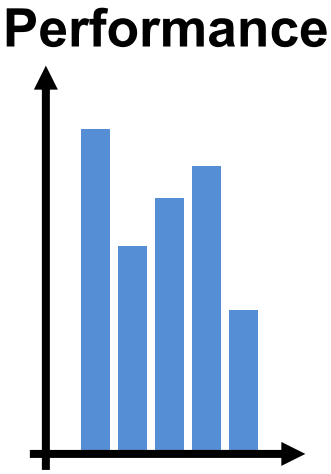


DNN Workloads



Accelerator

?



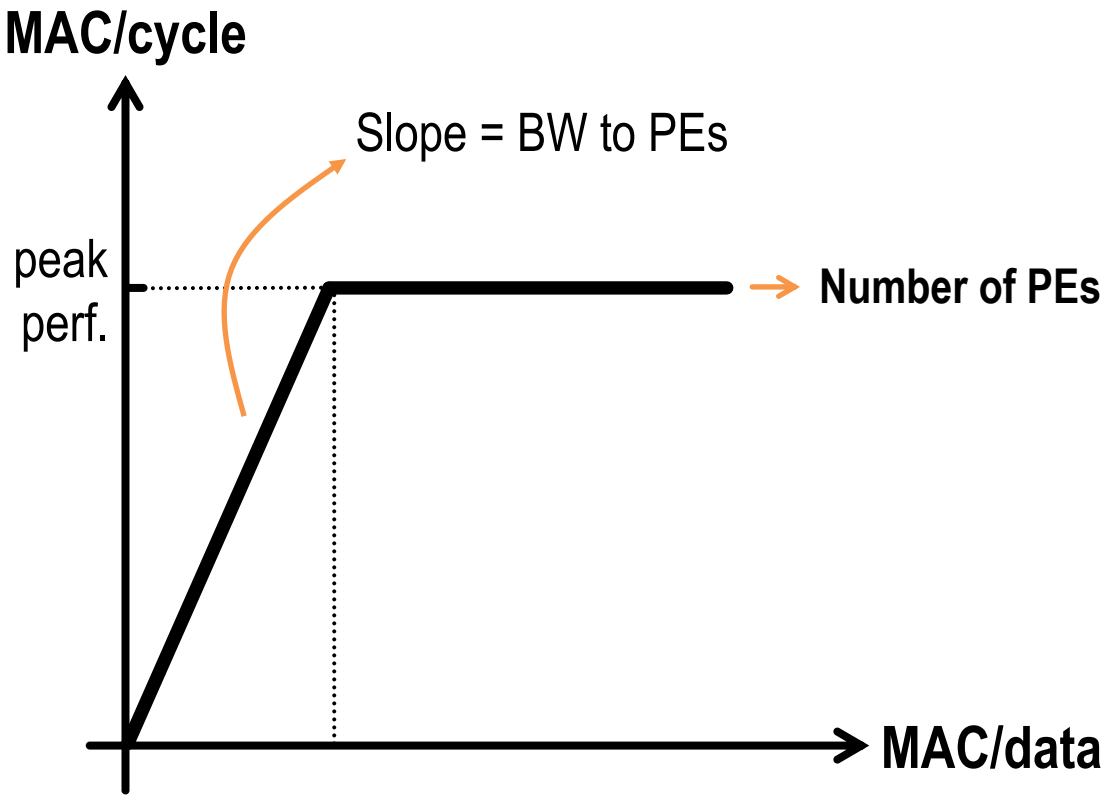
How to quickly estimate the performance of a DNN accelerator across various DNN workloads?

Eyexam:
Performance Eval Framework

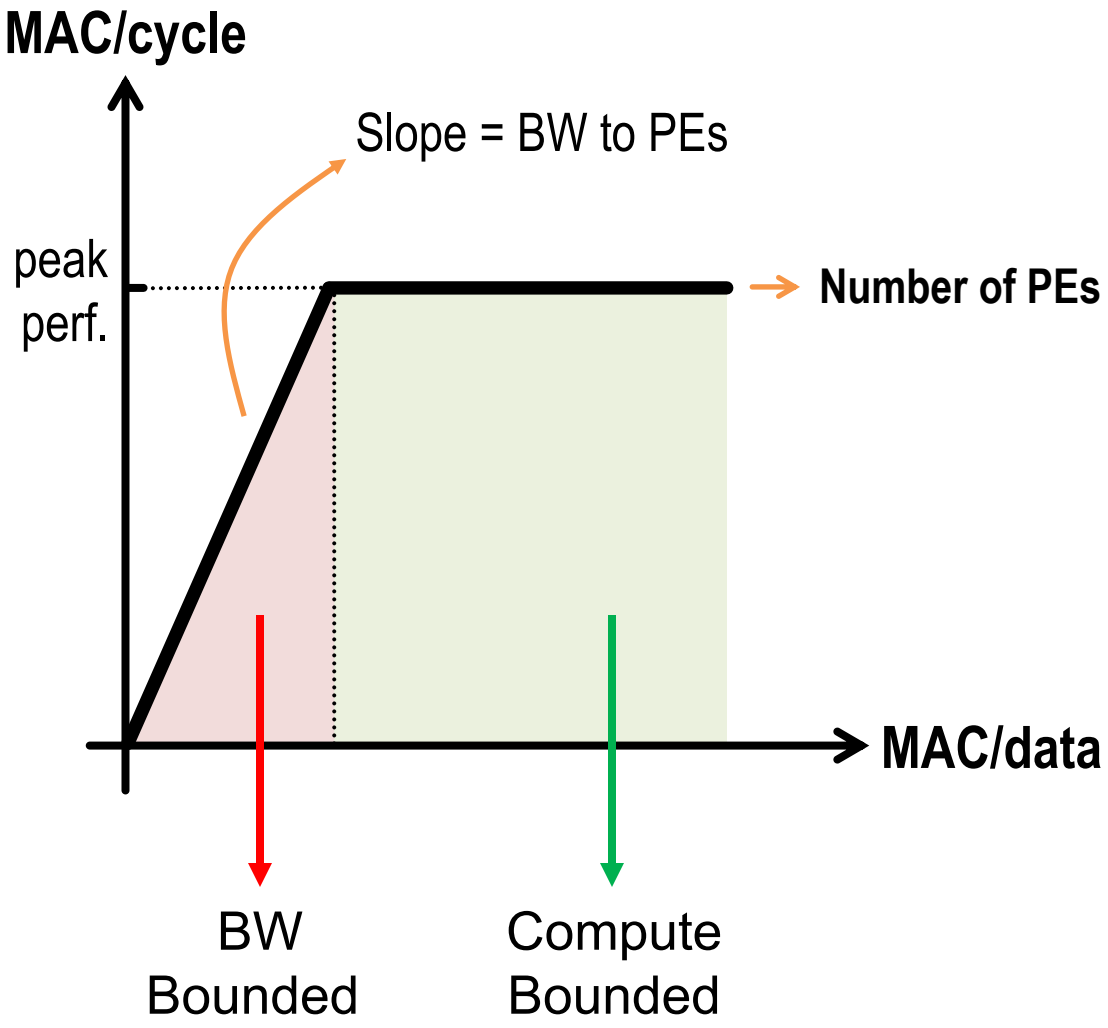
- A systematic way to quickly understand the performance limits of DNN accelerators in a step-by-step process

[Chen et al., JETCAS 2019]

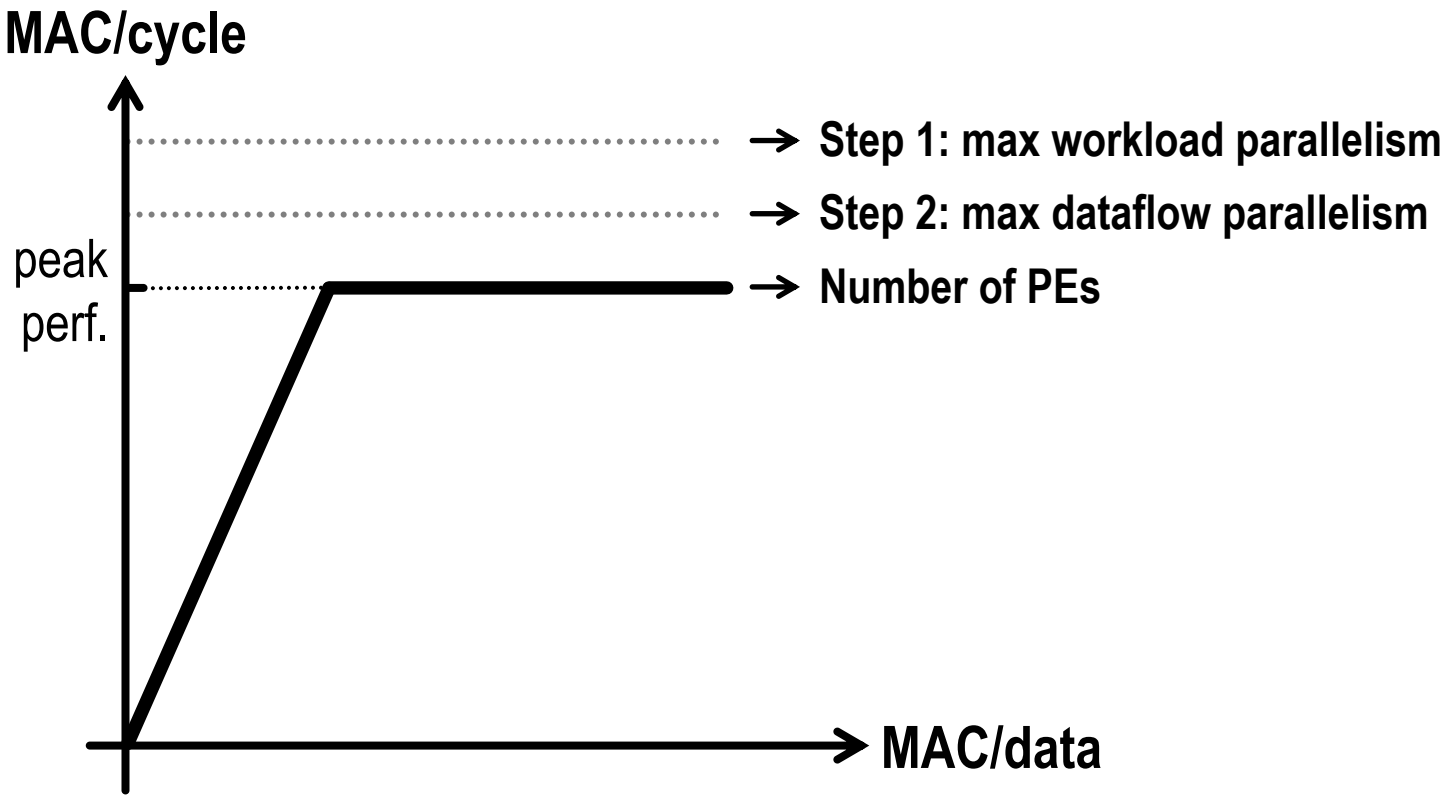
Eyexam: Performance Eval Framework



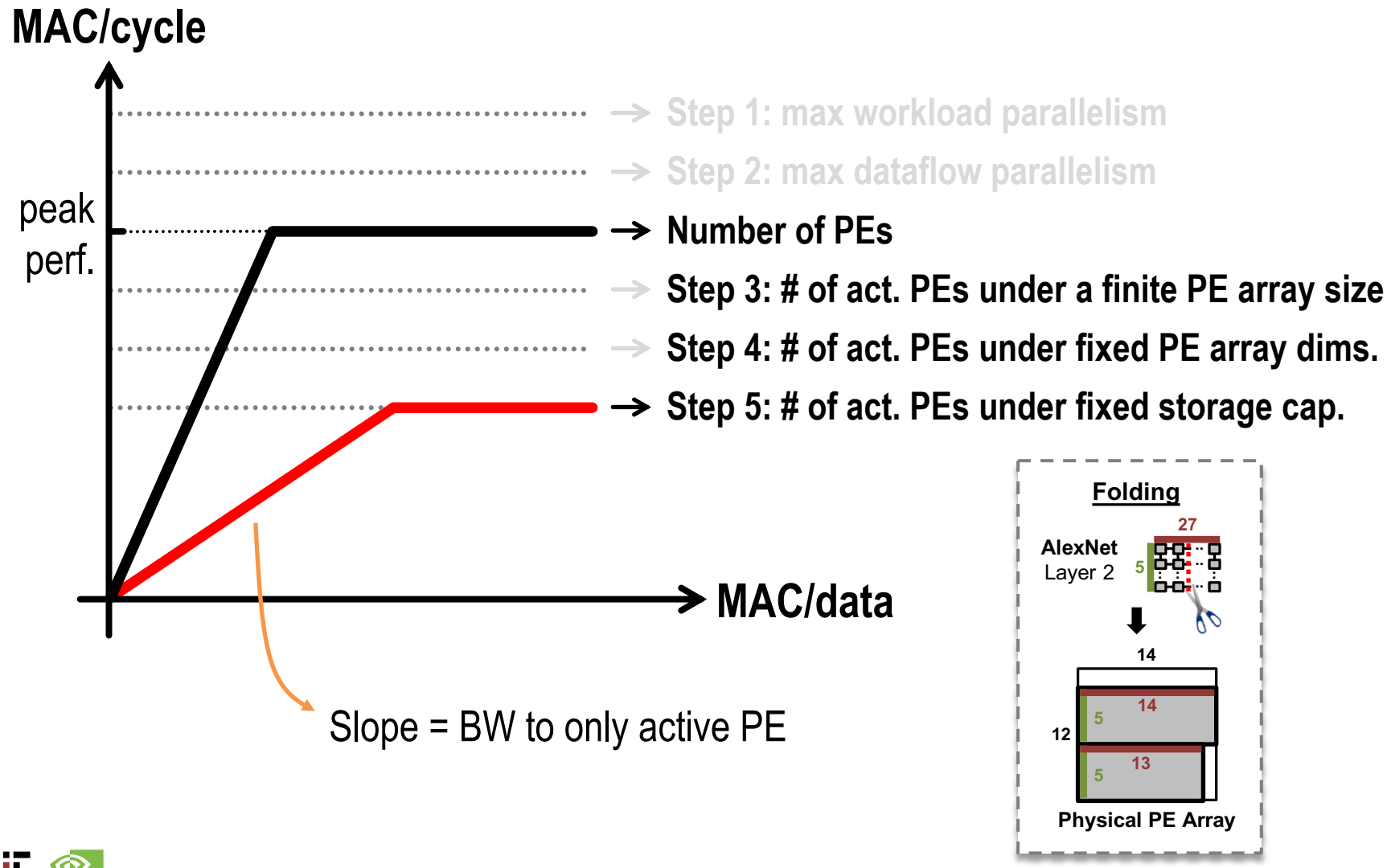
Eyexam: Performance Eval Framework



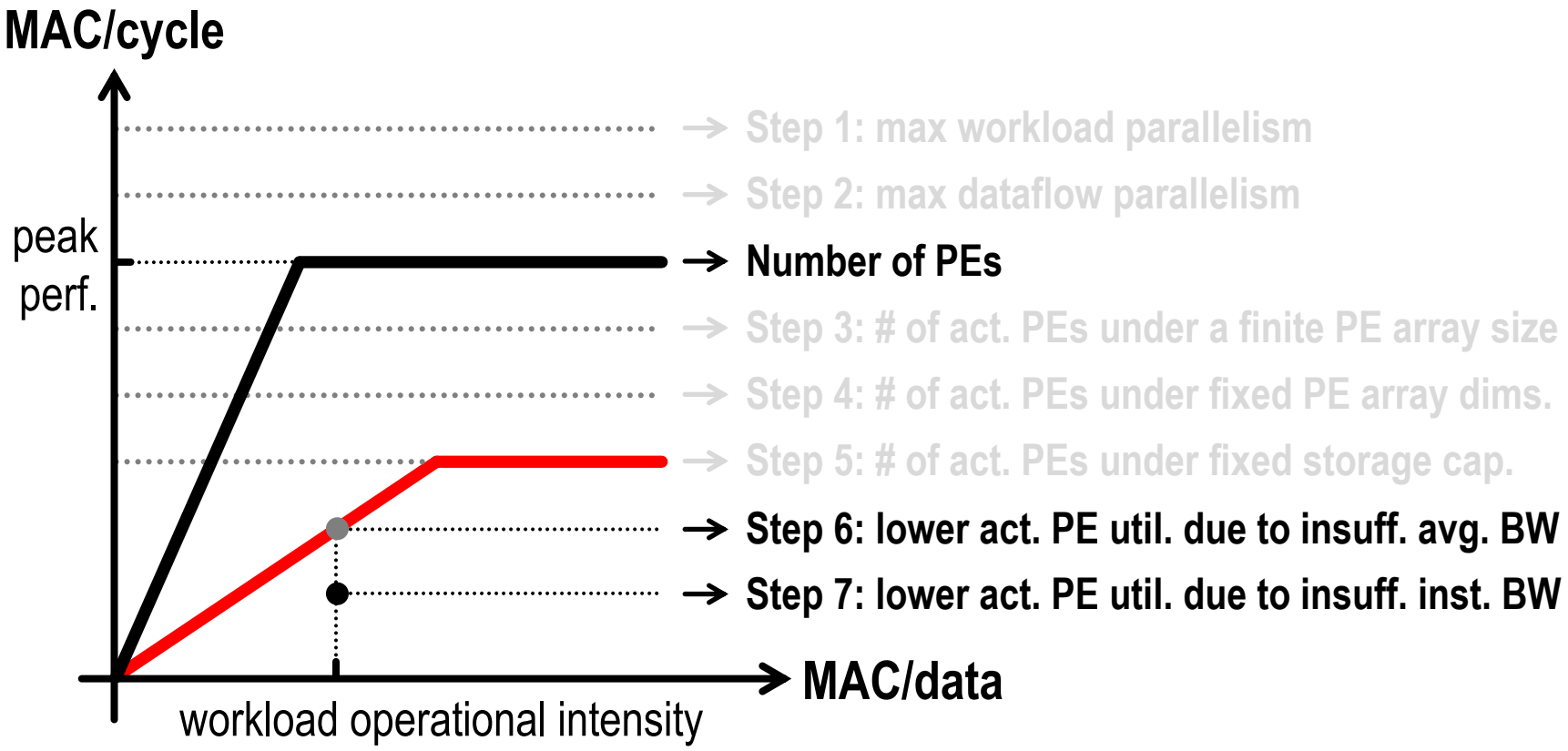
Eyexam: Performance Eval Framework



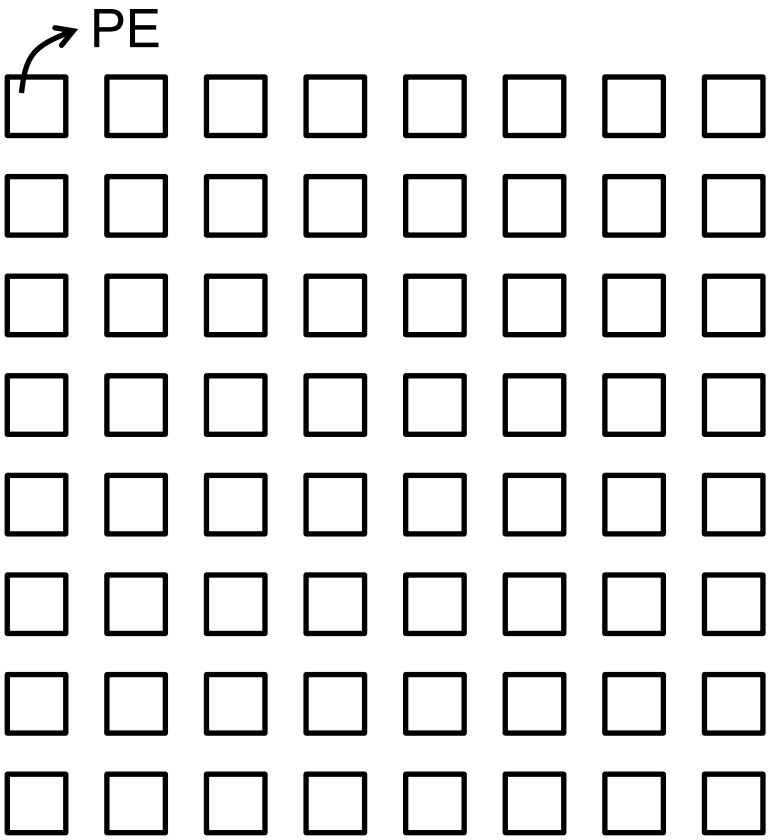
Eyexam: Performance Eval Framework



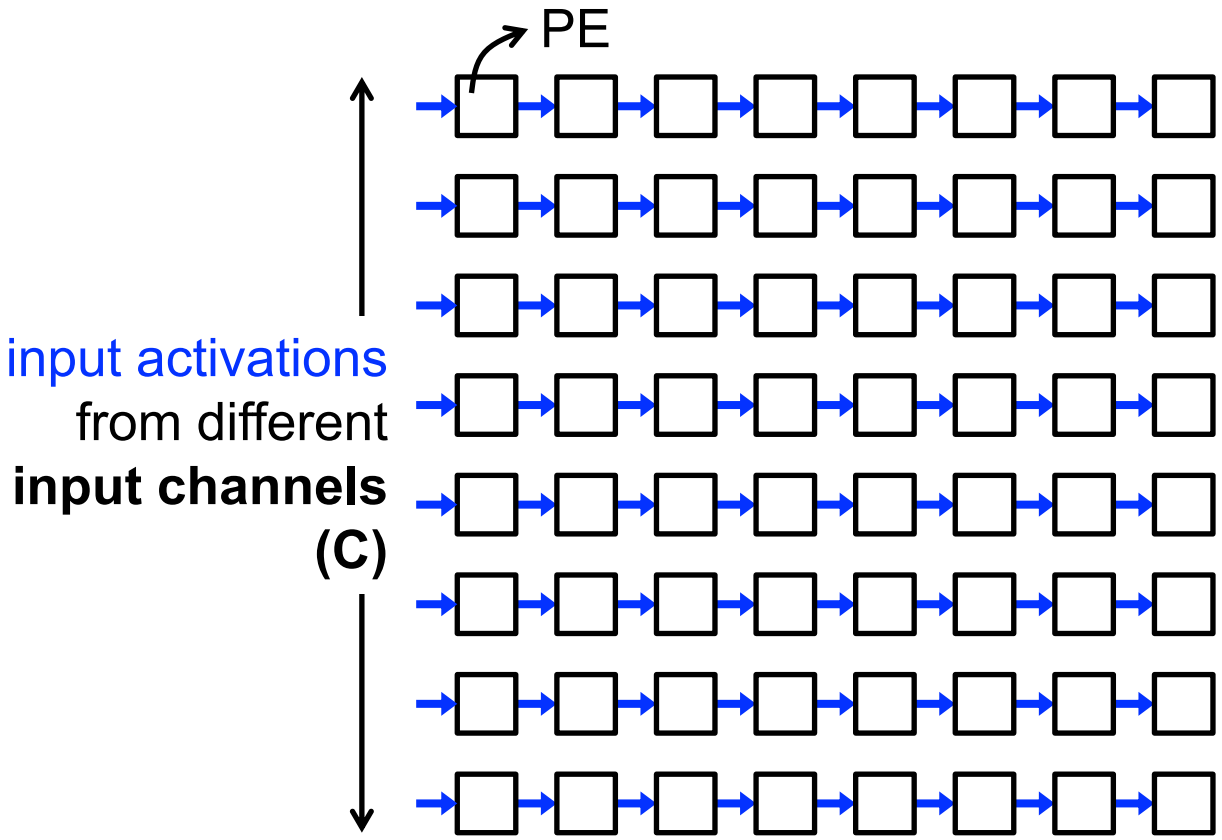
Eyexam: Performance Eval Framework



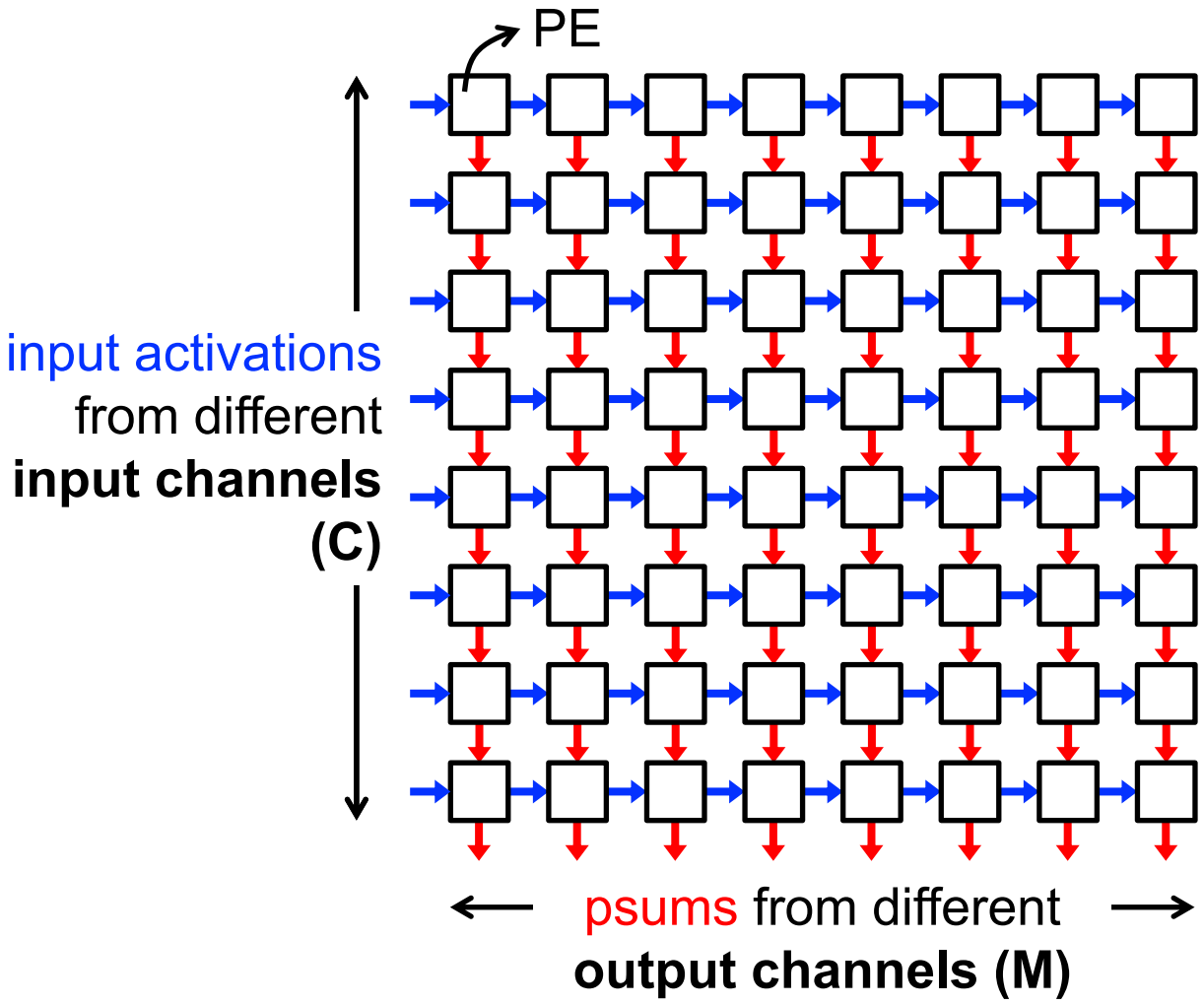
Example: A Common Design Pattern



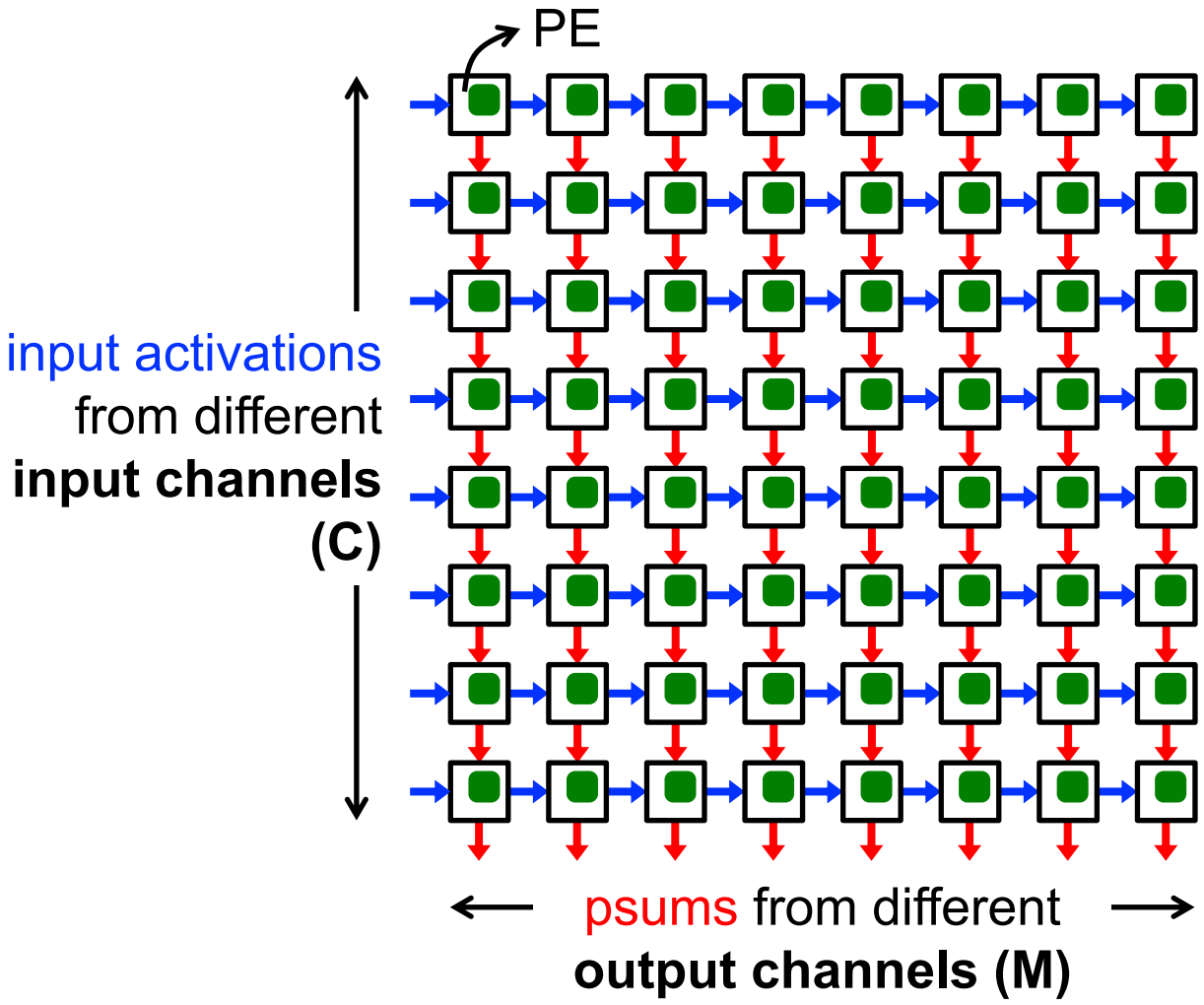
Example: A Common Design Pattern



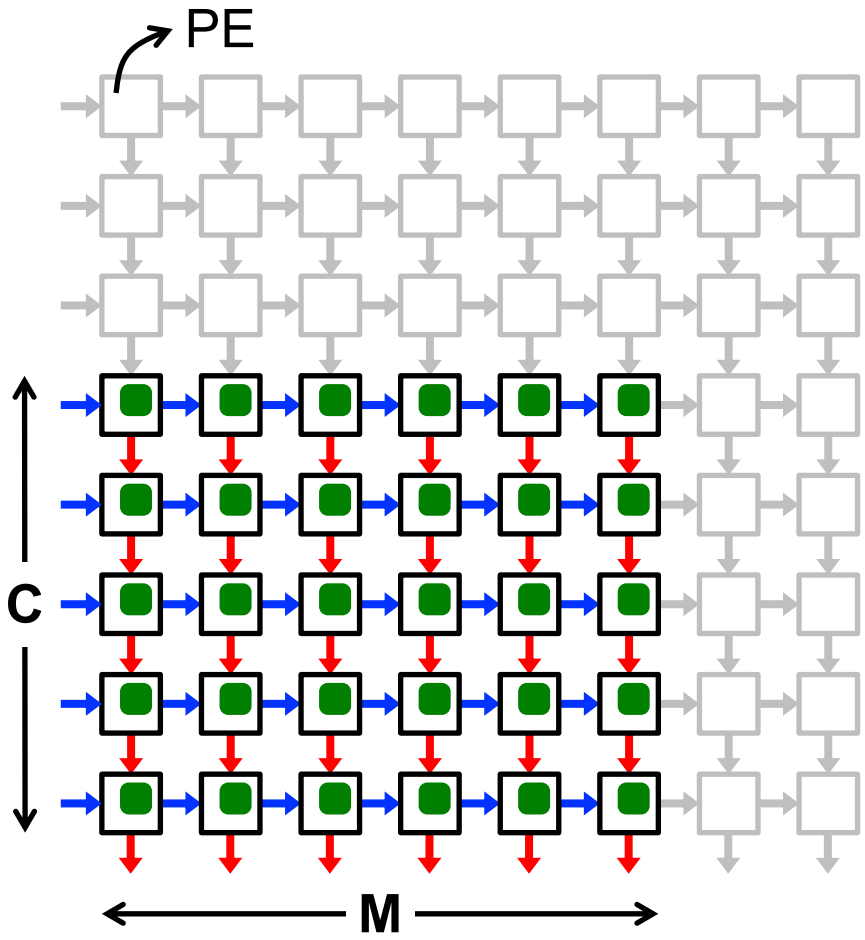
Example: A Common Design Pattern



Example: A Common Design Pattern



Example: A Common Design Pattern

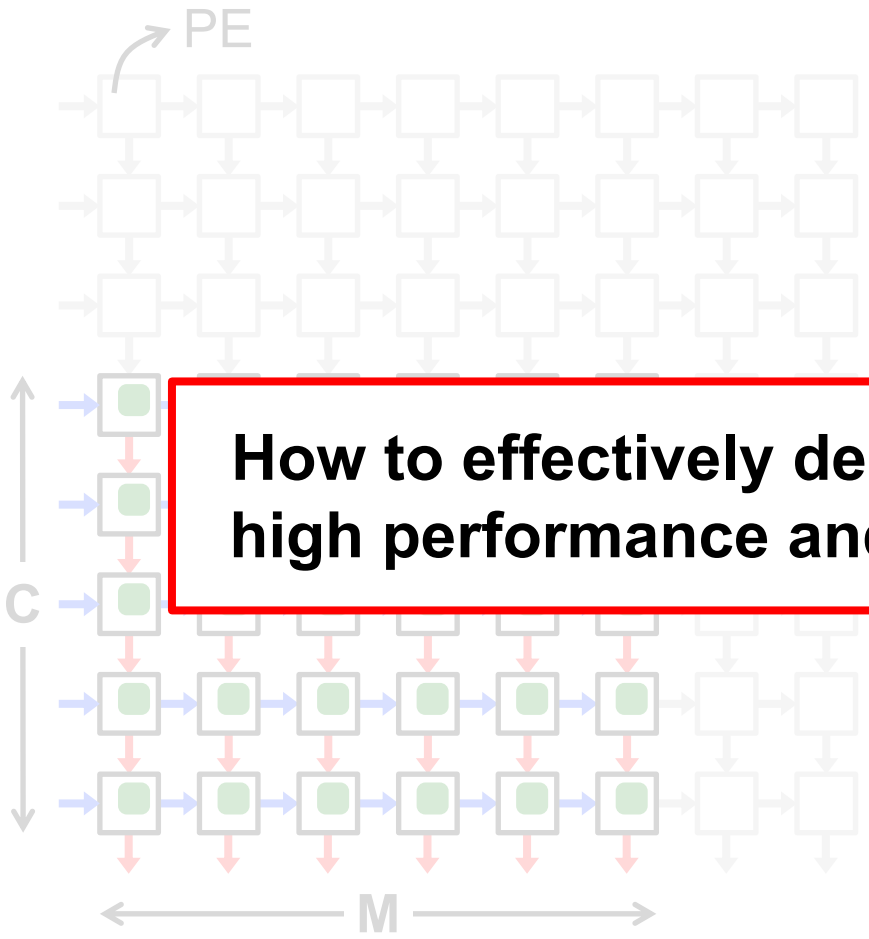


1. PE array **underutilized** If there are fewer input and/or output channels than the array dimensions

2. Effective data delivery BW also becomes lower → further impact performance

3. Not scalable →utilization will be worse at larger scales

Example: A Common Design Pattern



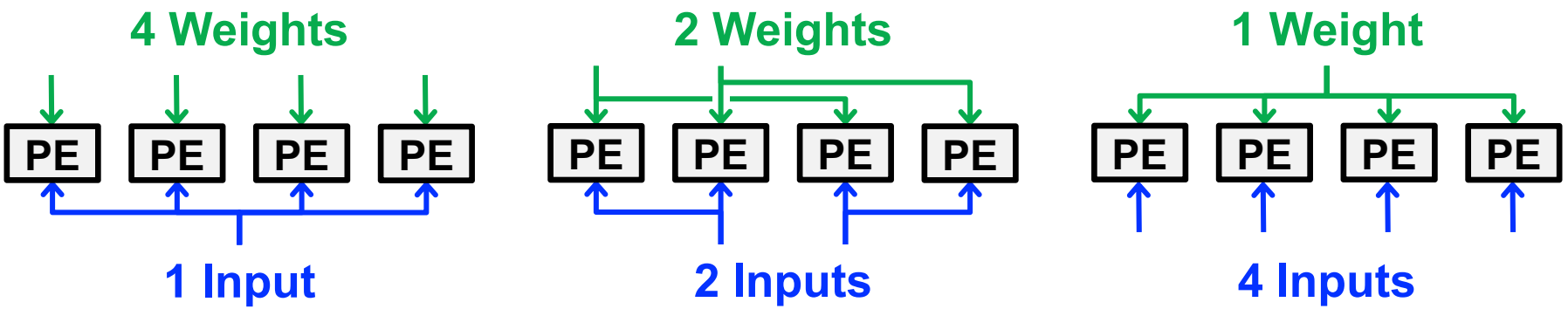
1. PE array **underutilized** If there are fewer input and/or output channels than the array dimensions

How to effectively deliver data to achieve high performance and energy efficiency?

3. Not scalable →utilization will be worse at larger scales

A More Flexible Data Delivery Strategy

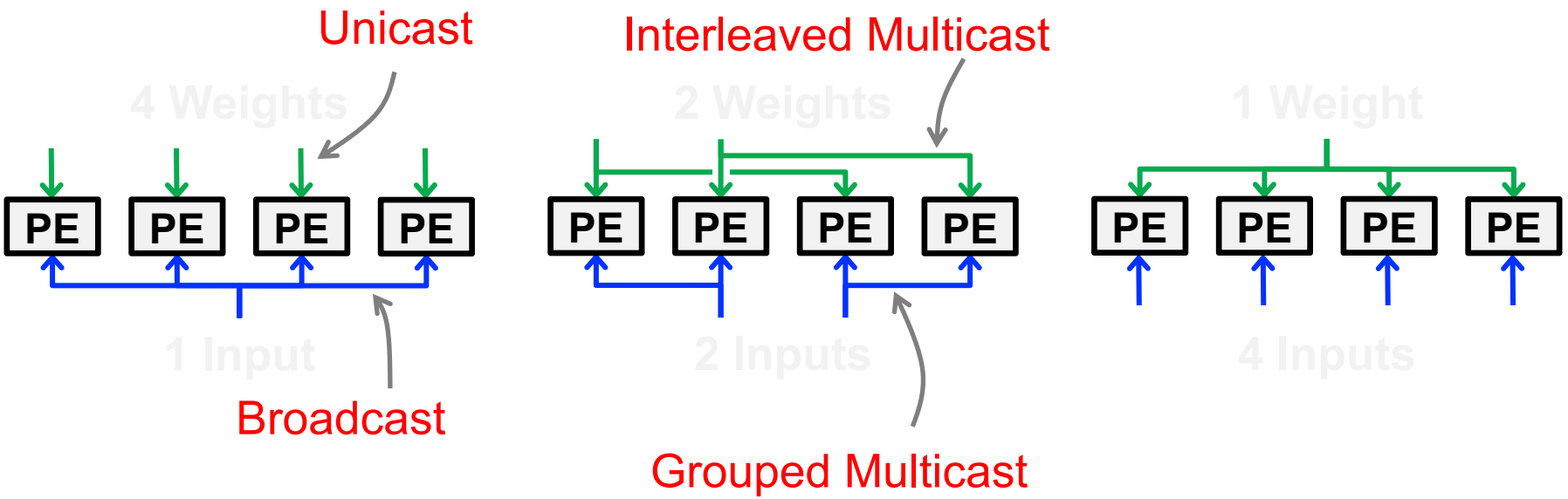
Adapt to the reuse and bandwidth requirements



A More Flexible Data Delivery Strategy

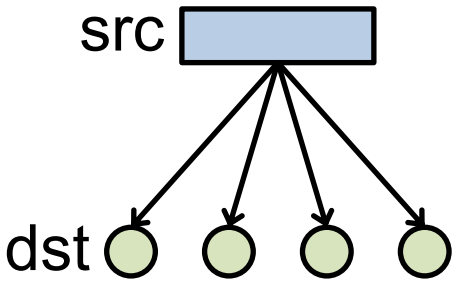
Adapt to the reuse and bandwidth requirements

4 Data Delivery Patterns



On-Chip Network (NoC) is the Bottleneck

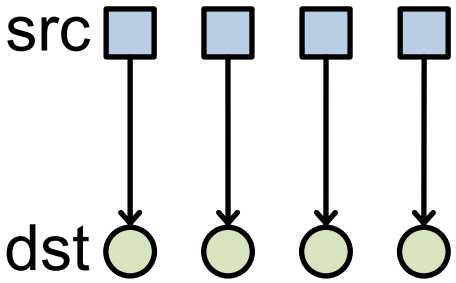
Broadcast Network
(Eyeriss v1 NoC)



High Reuse

Low Bandwidth

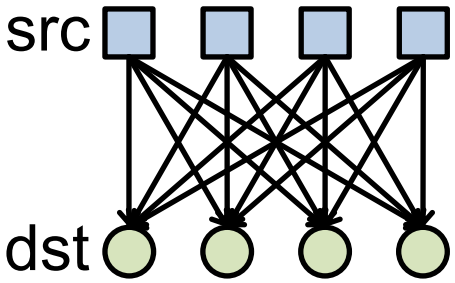
Unicast Networks



Low Reuse

High Bandwidth

All-to-All Networks

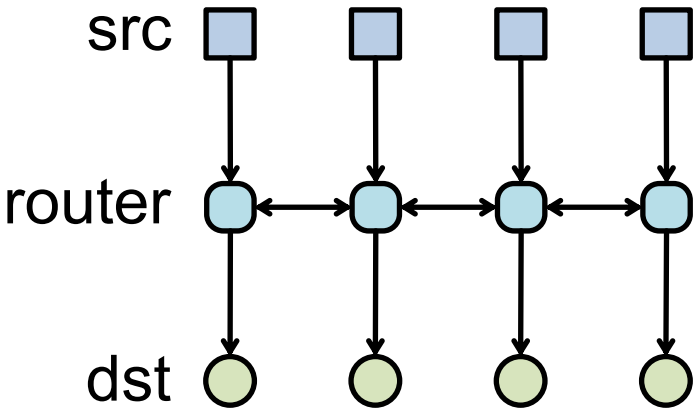


High Reuse

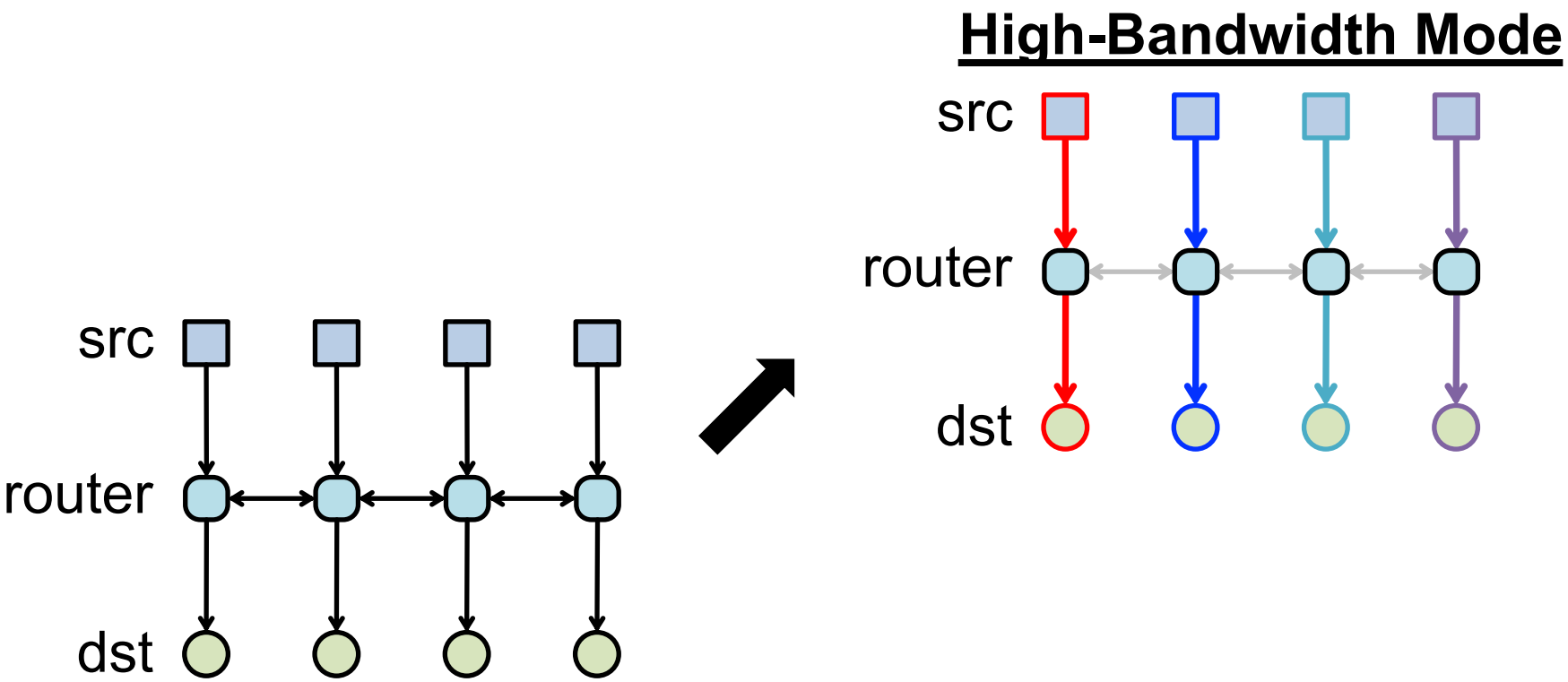
High Bandwidth

Hard to Scale

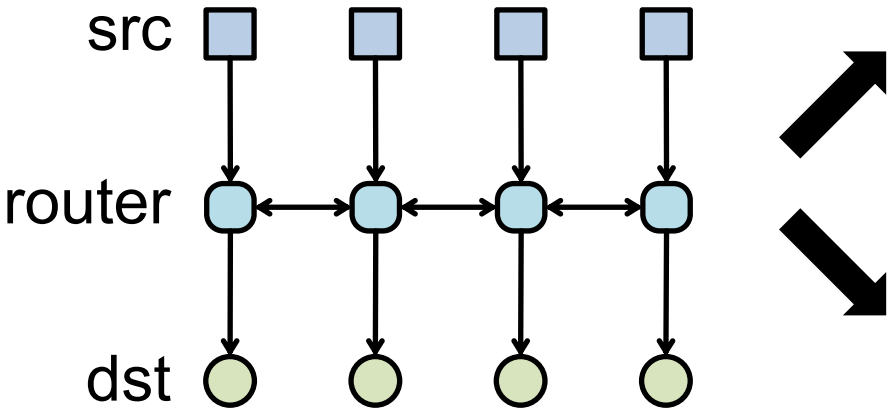
Mesh Network – Best of Both Worlds



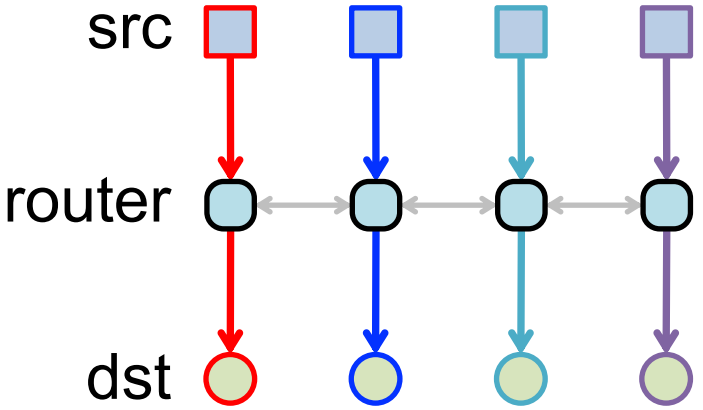
Mesh Network – Best of Both Worlds



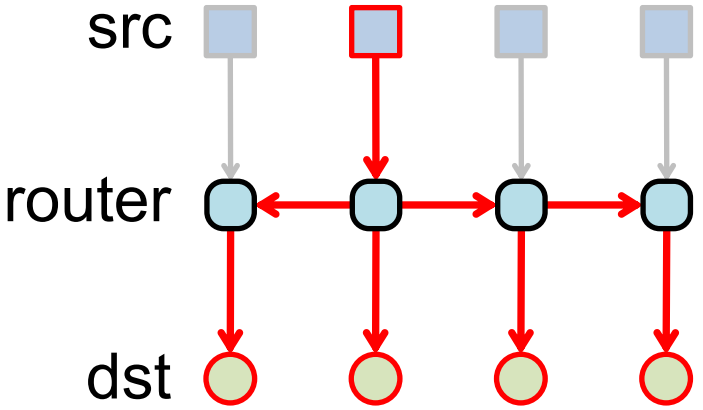
Mesh Network – Best of Both Worlds



High-Bandwidth Mode

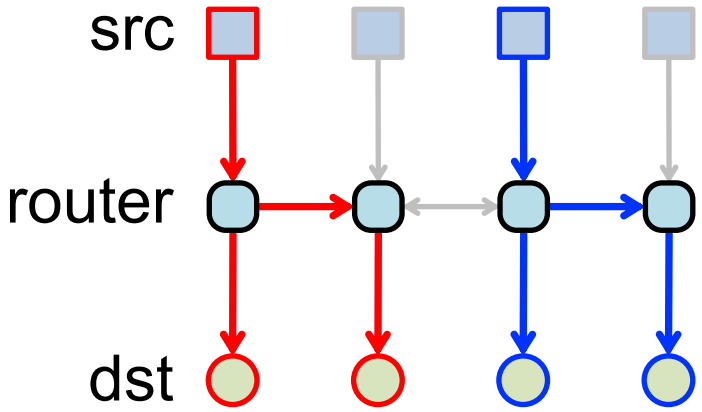


High-Reuse Mode



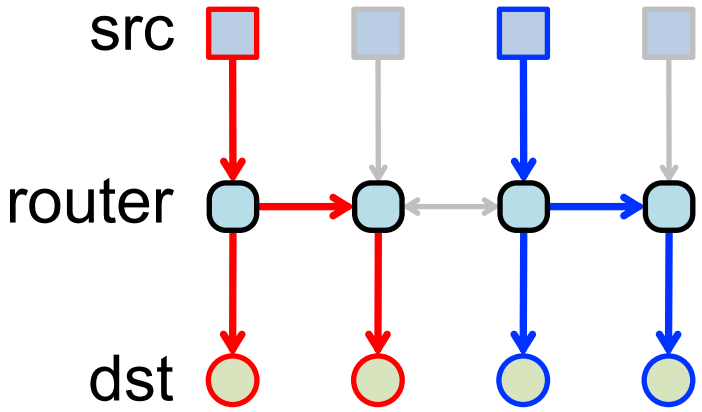
Mesh Network – More Complicated Cases

Grouped-Multicast Mode

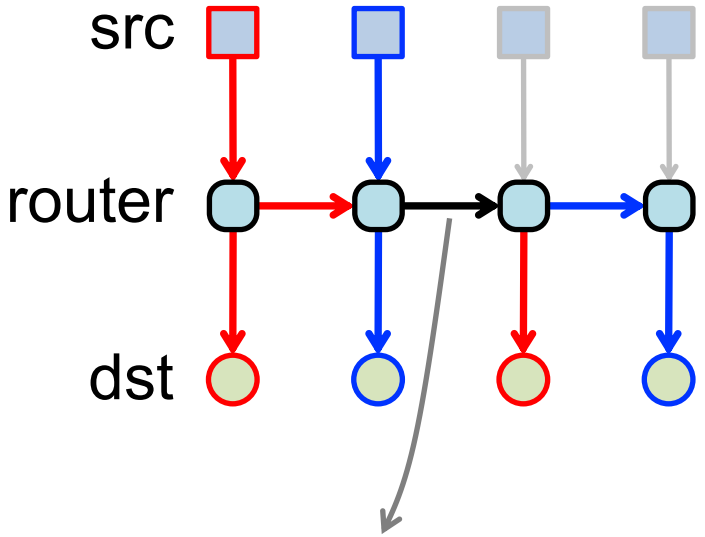


Mesh Network – More Complicated Cases

Grouped-Multicast Mode



Interleaved-Multicast Mode



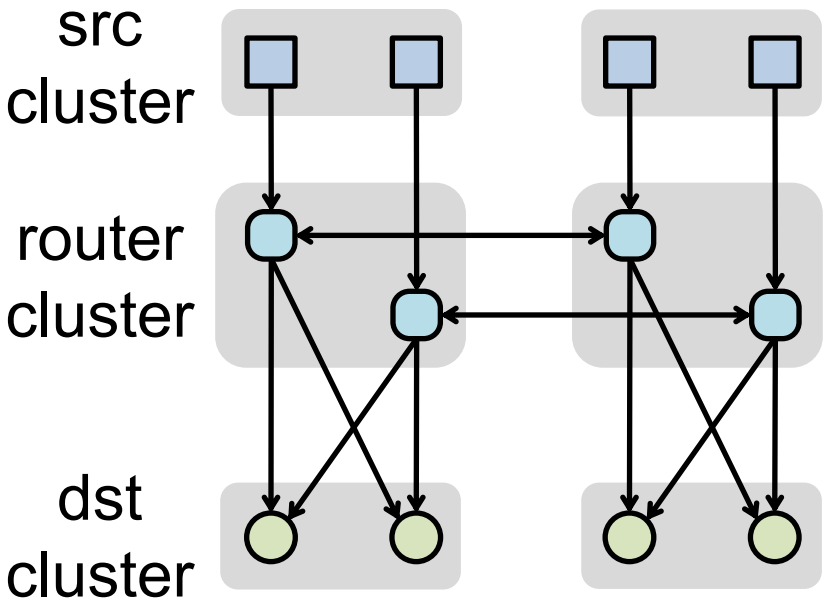
Bandwidth-limited route
(flow control required)

Hierarchical Mesh Network

- Flexible to support patterns ranging from high reuse to high bandwidth scenarios
- Can be easily scaled at a low cost

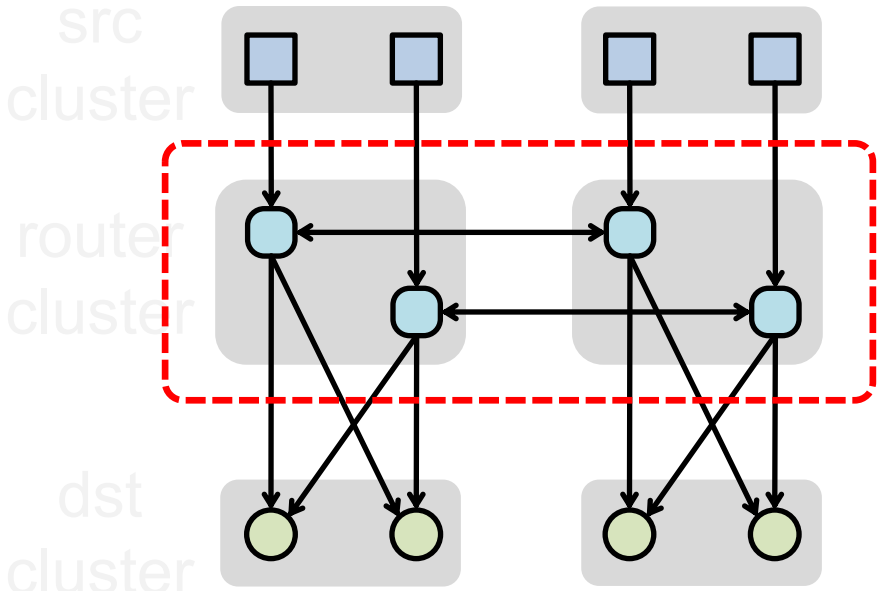
[Chen et al., JETCAS 2019]

Design of Hierarchical Mesh Network

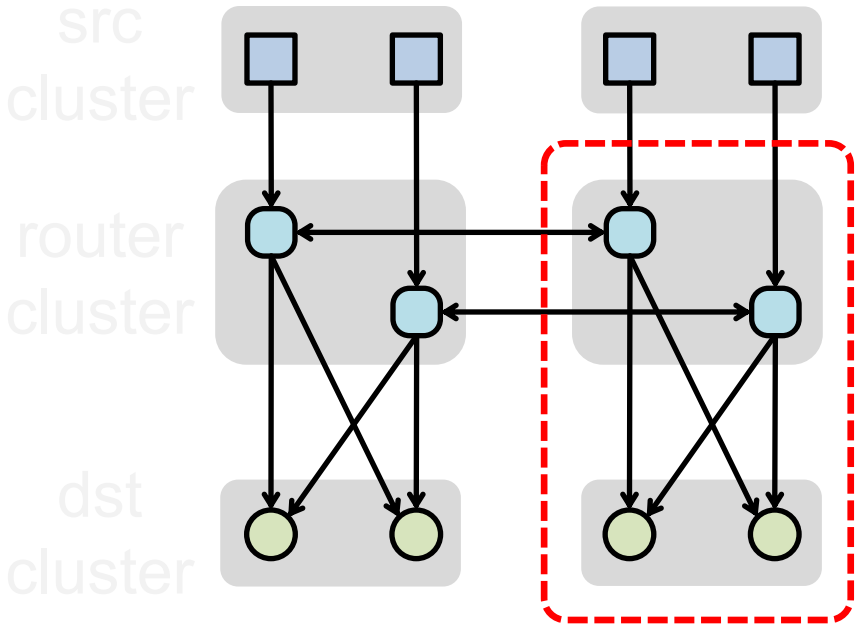


Design of Hierarchical Mesh Network

Mesh Network for **inter-cluster** connections



Design of Hierarchical Mesh Network

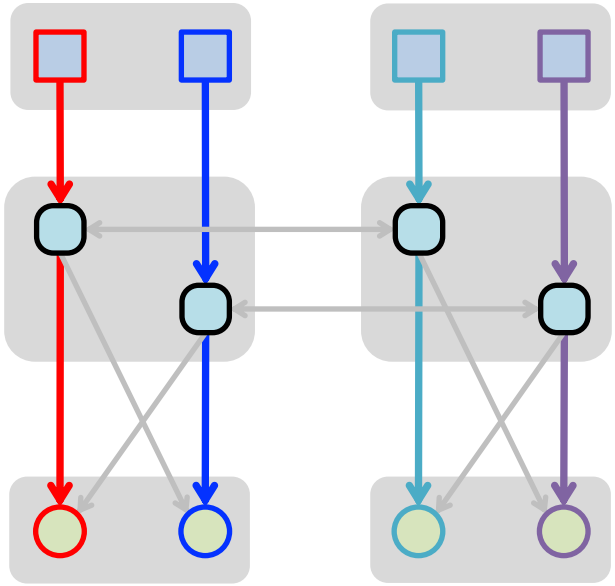


All-to-All Network for **intra-cluster** connections

Complexity is contained within a cluster

Design of Hierarchical Mesh Network

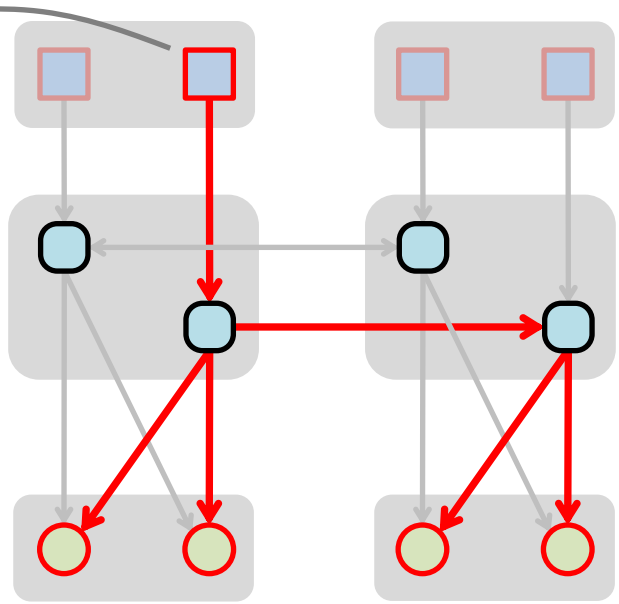
High-Bandwidth Mode



Design of Hierarchical Mesh Network

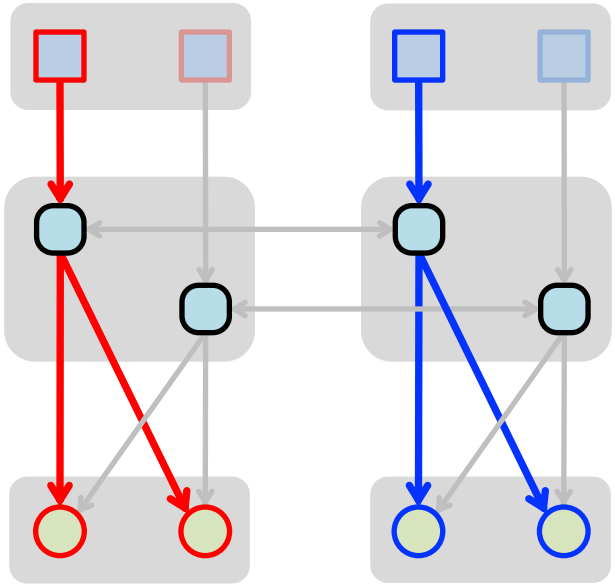
High-Reuse Mode

from any one src



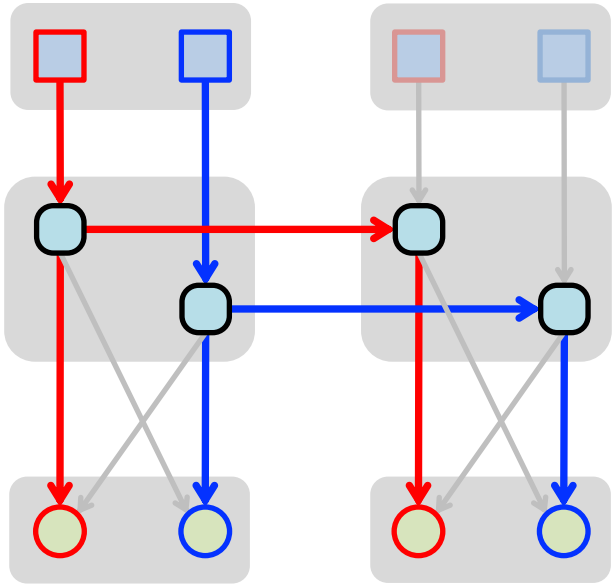
Design of Hierarchical Mesh Network

Grouped-Multicast Mode



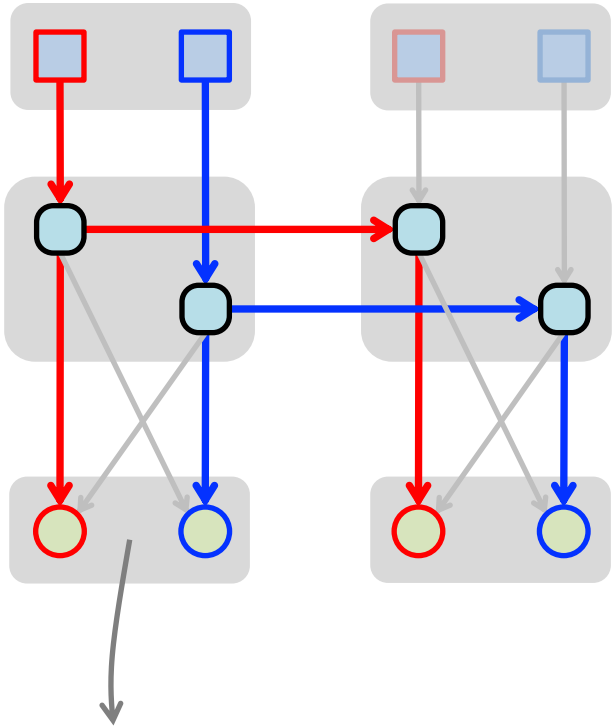
Design of Hierarchical Mesh Network

Interleaved-Multicast Mode



Design of Hierarchical Mesh Network

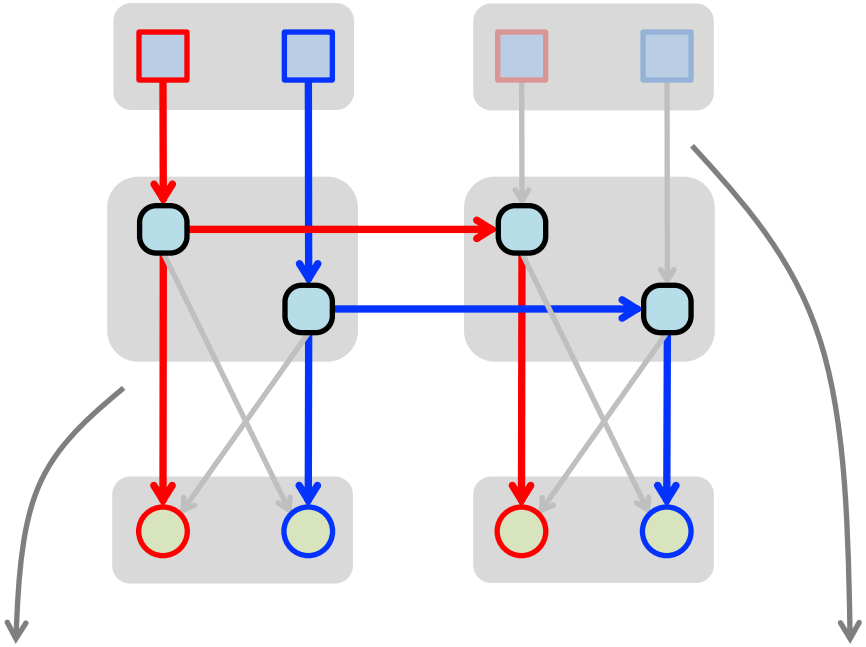
Interleaved-Multicast Mode



Can interleave more by scaling up the cluster size

Design of Hierarchical Mesh Network

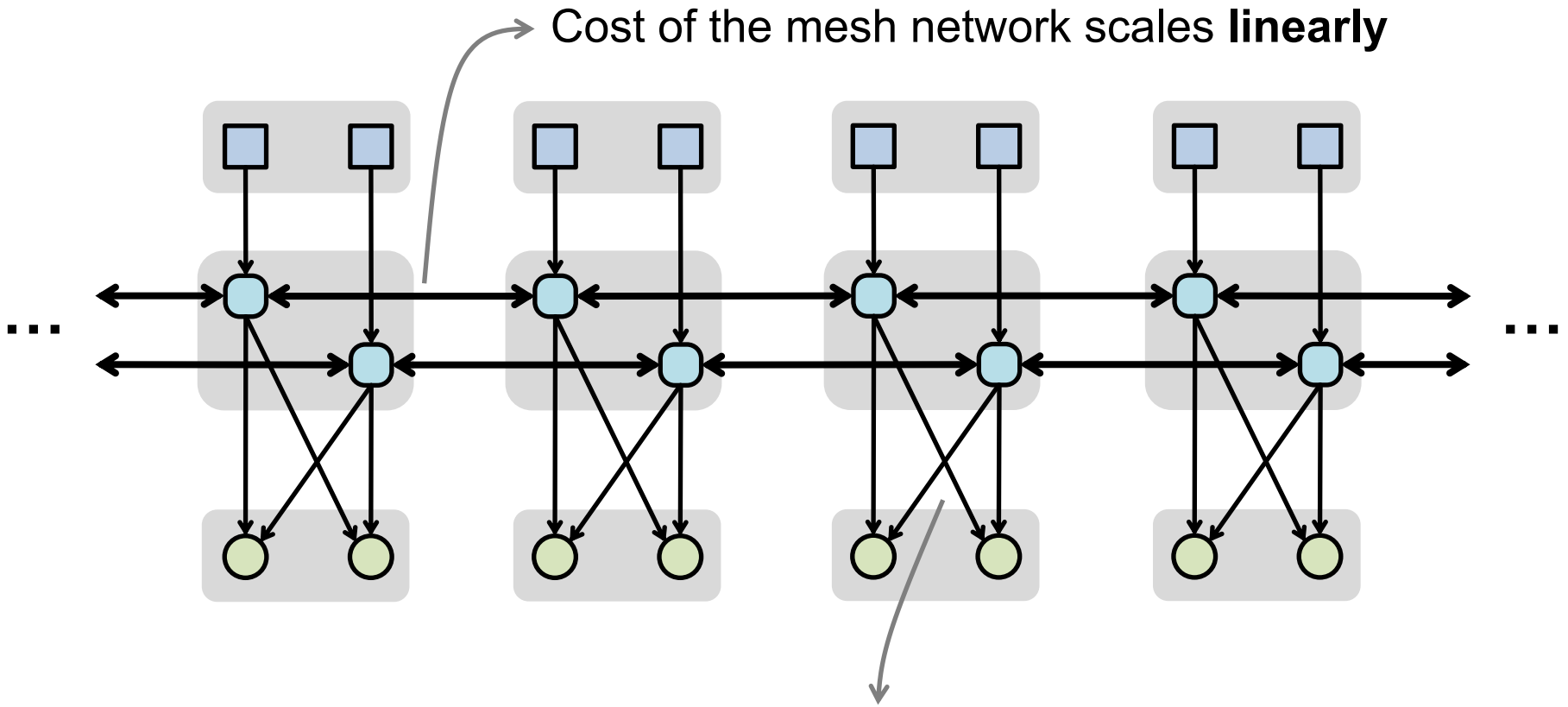
Interleaved-Multicast Mode



All routes are determined **at configuration time**

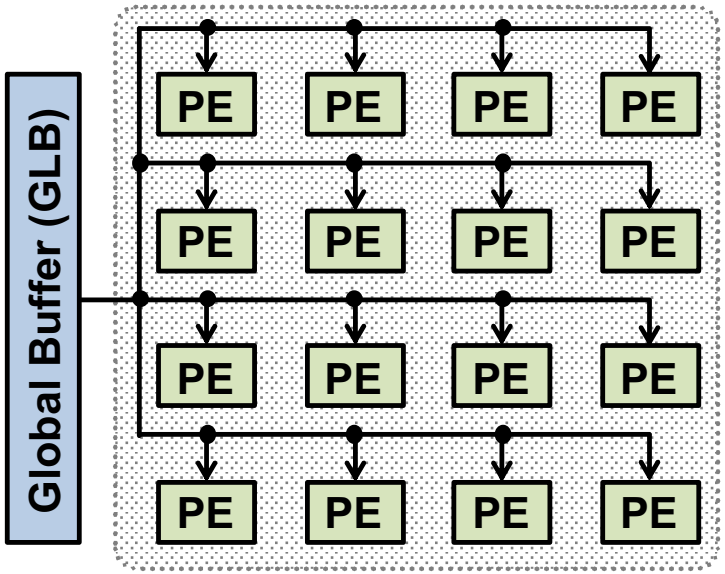
→ **Routers** are **circuit-switched** (only **MUXes**)

Scaling the Hierarchical Mesh Network

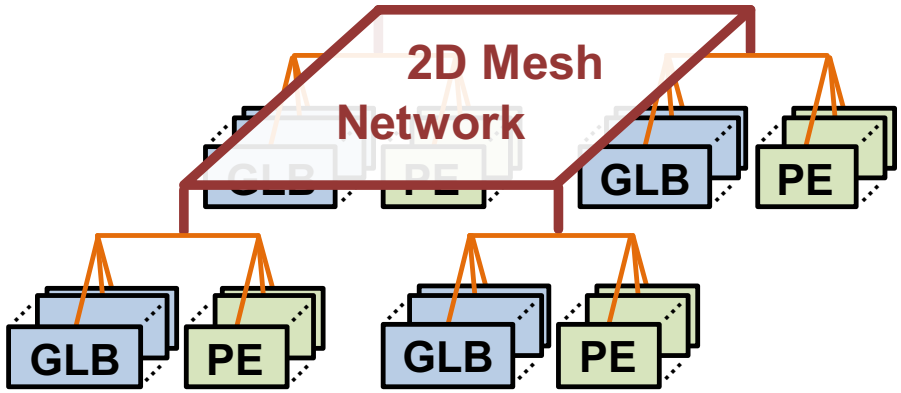


Eyeriss with Hierarchical Mesh Network

Eyeriss v1

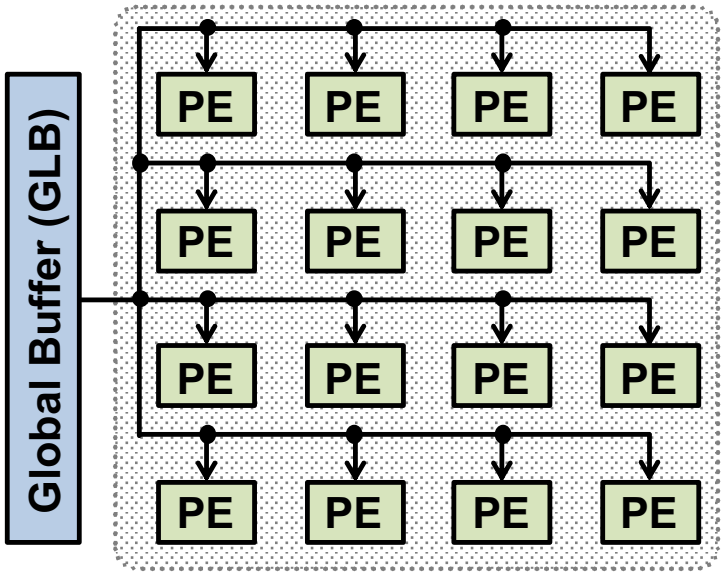


Eyeriss v2

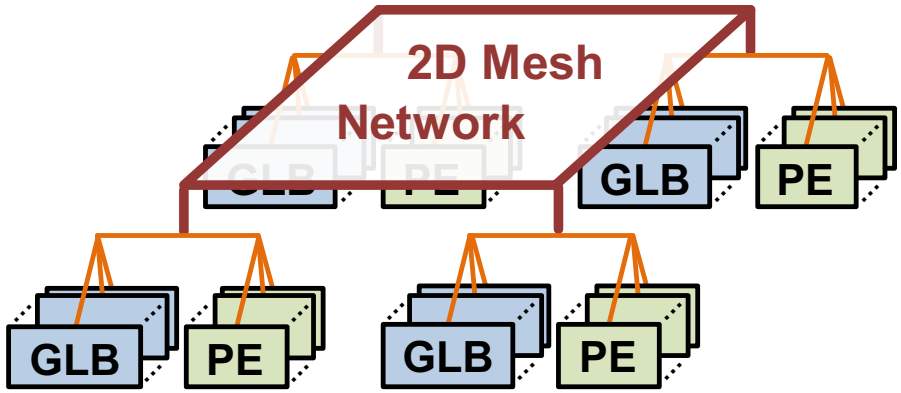


Eyeriss with Hierarchical Mesh Network

Eyeriss v1



Eyeriss v2



	Speedup	Energy Efficiency
AlexNet	6.9×	2.6×
MobileNet	5.6×	1.8×

* not including the benefits from the sparsity features in v2

Summary

- **Data reuse** is the key to achieving **high energy efficiency**.
- High PE utilization with **adaptive on-chip networks** is the key to achieving **high performance**
- Co-design of **dataflow** and **hardware** is critical for the optimization of **performance**, **energy efficiency** and **flexibility** for DNN accelerators.