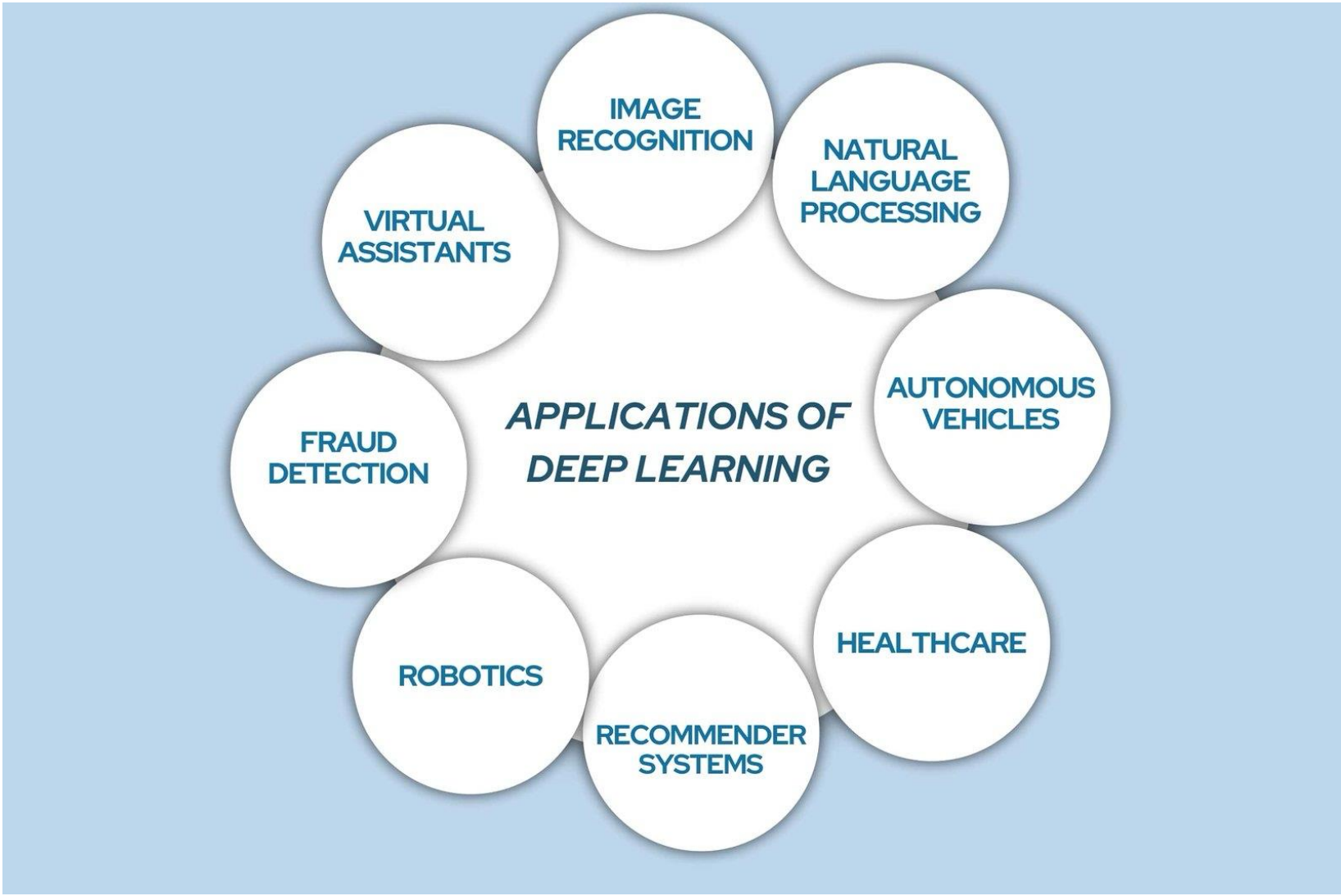
Performance and Power Modeling of ML Accelerators

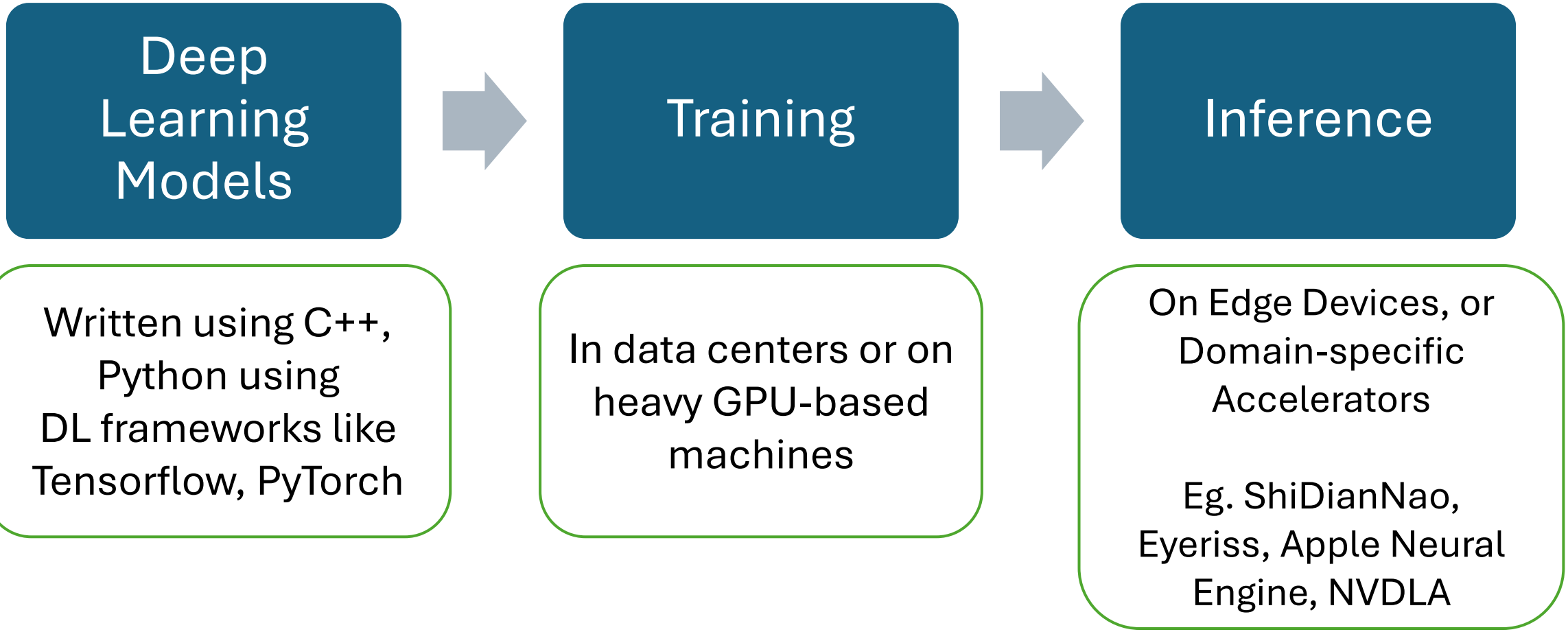
Ayushi Agarwal

Course: Synthesis of Digital Systems (COL719)

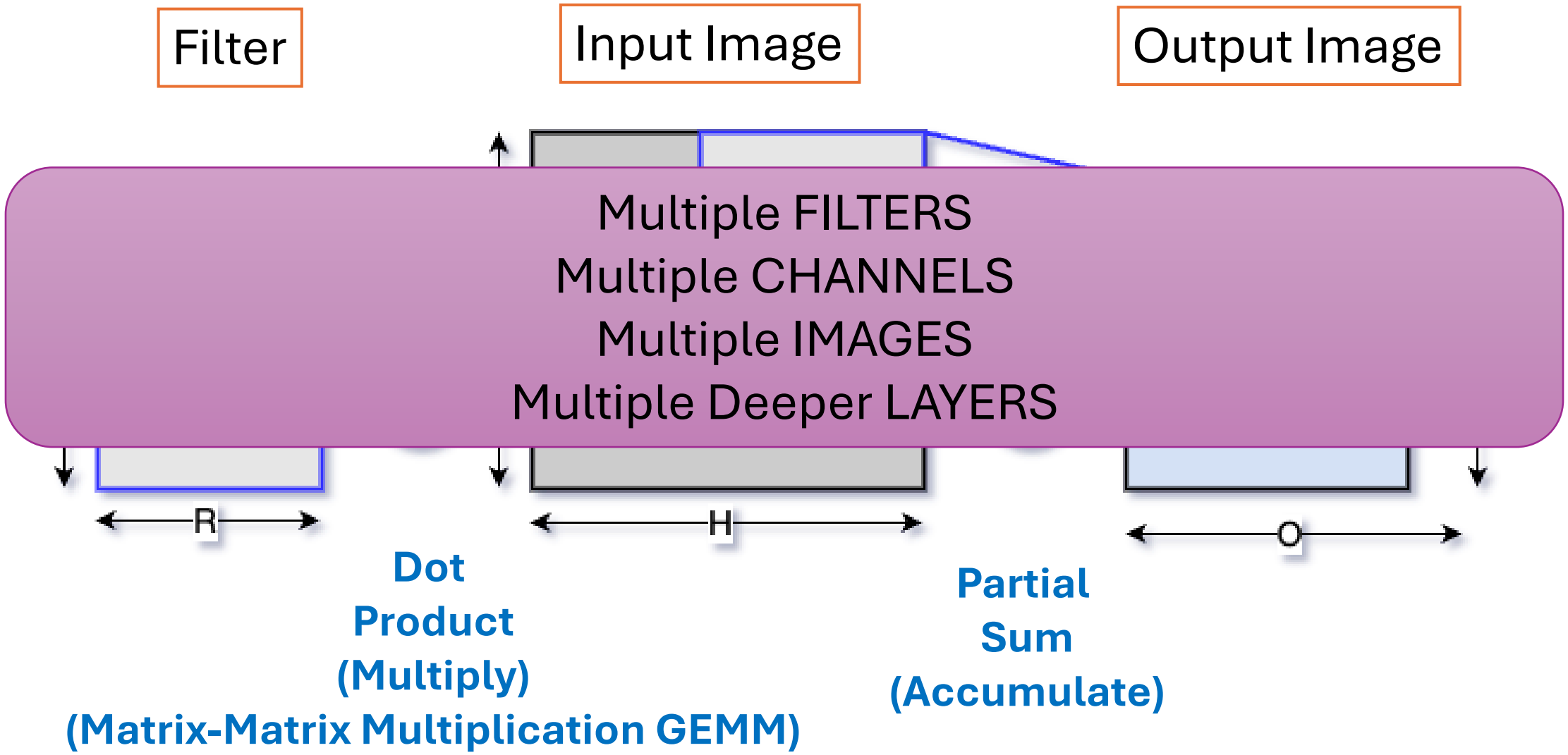
Deep Learning Applications



Deep Learning Application Pipeline



Deep Convolutional Neural Networks



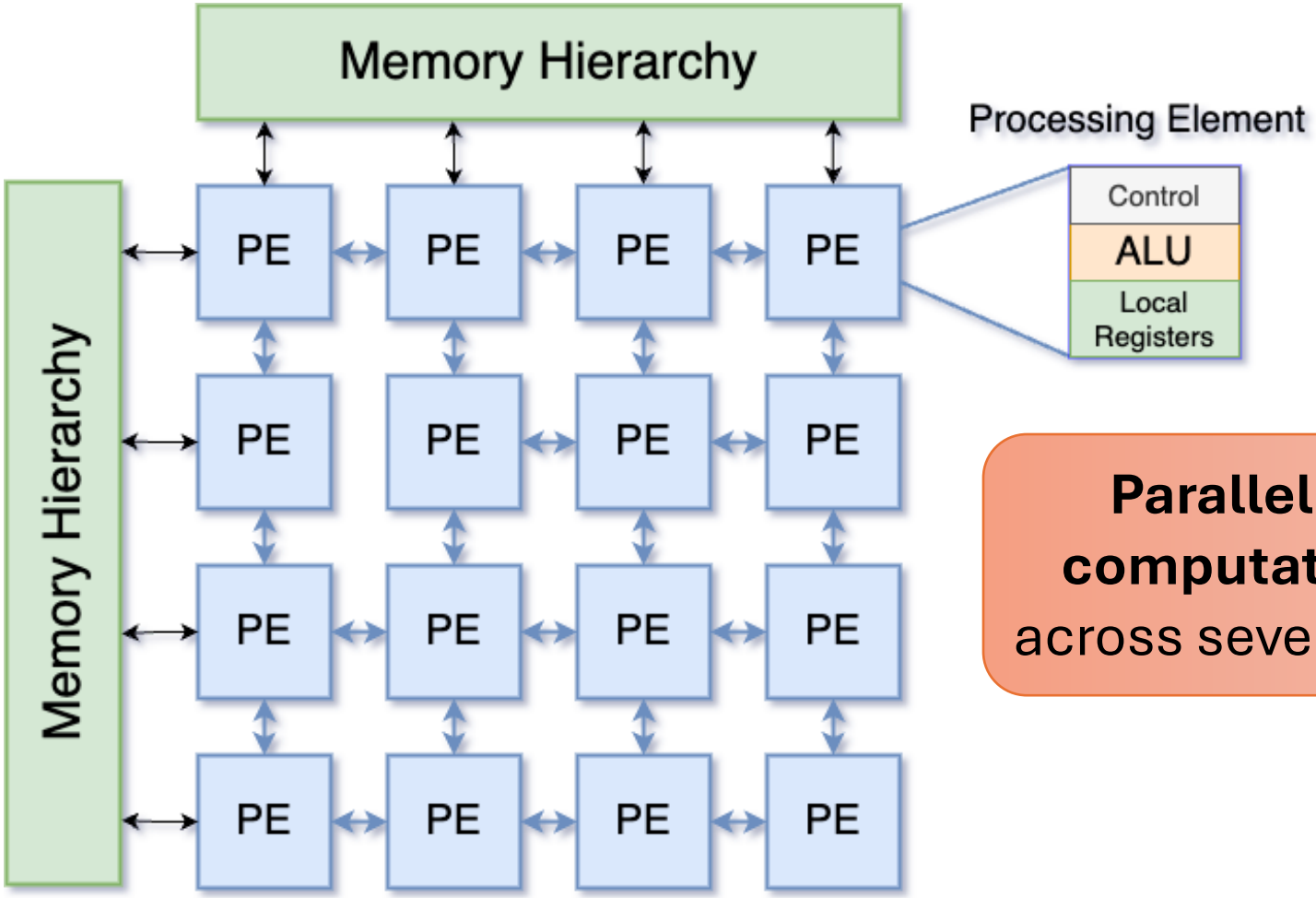
Challenges with DNN computations

- Large number of parameters in modern deep networks
  - Require many parallel computations ⟶ Not suitable for CPU execution
  - Require many data transfers from the memory to compute units ⟶ GPU computation is energy-inefficient

DNN Model	Number of Parameters (weights)
AlexNet (2012)	3.98M
VGGNet-16 (2014)	28.25M
GoogleNet (2015)	6.77M
ResNet-50 (2016)	23M
DLRM (2019)	540M

Domain-Specific Accelerators

Example: Eyeriss, Google TPU, NVDLA, ShiDianNao



Reuse Data for minimizing off-chip DRAM accesses

Parallelize computations across several PEs

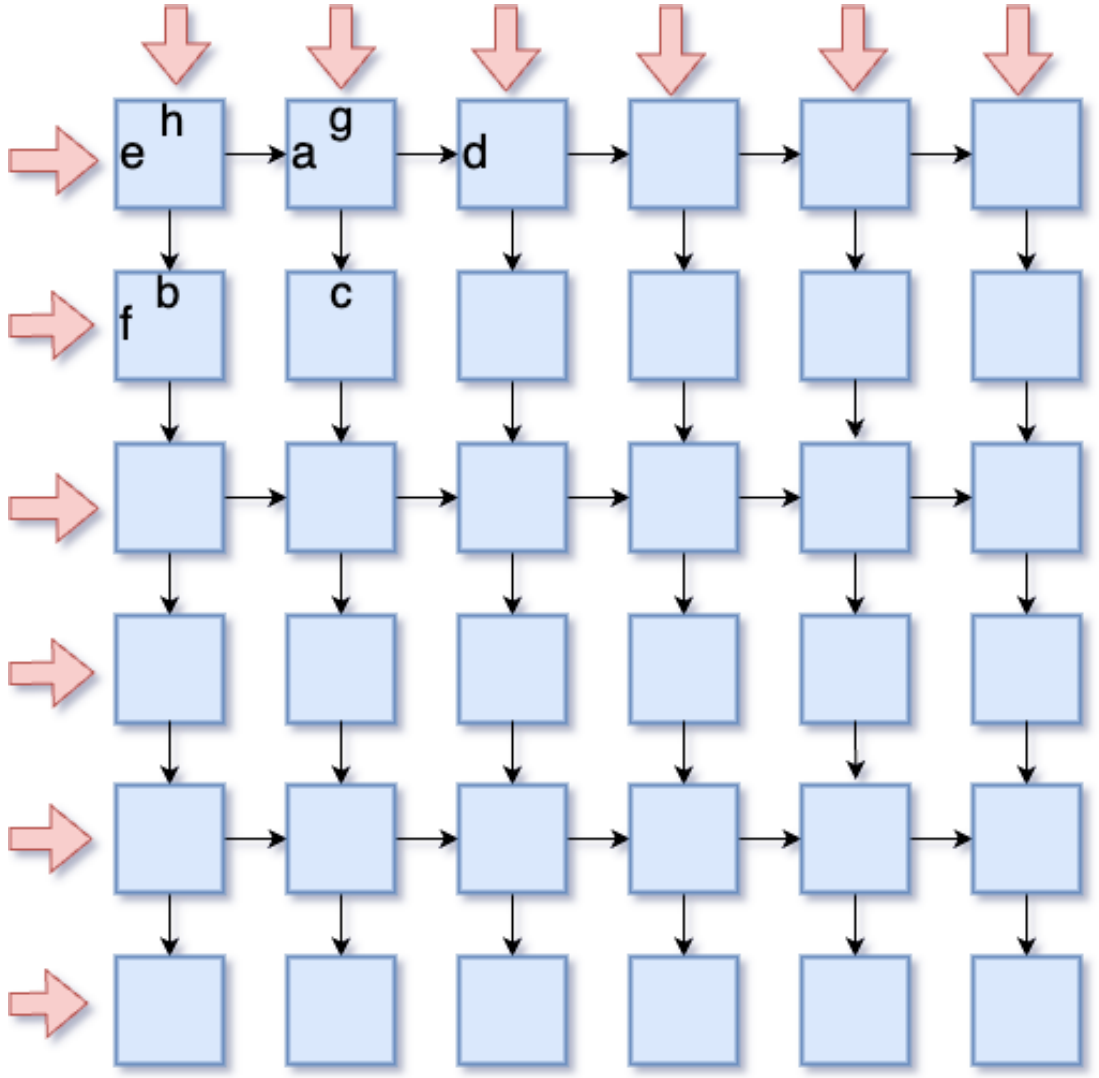
Systolic Array Architecture

- Designed for Efficient parallel computation
- Composed of identical PEs
- Simpler data routing and communication btw PEs
- Pipelined Dataflow

High Parallelism

Ease of Implementation

Efficient Data Reuse



Model GEMM using Systolic Array

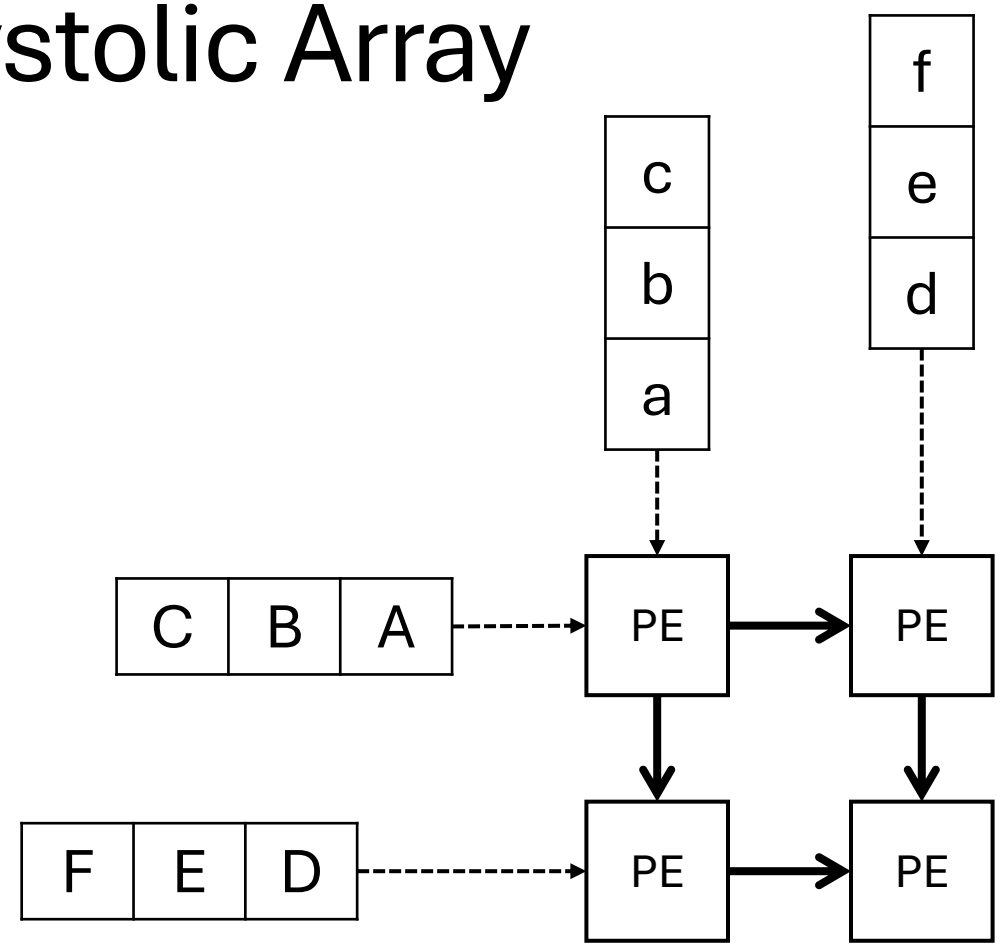
A	B	C
D	E	F

×

a	d
b	e
c	f

=

Aa + Bb + Cc	Ad + Be + Cf
Da + Eb + Fc	Dd + Ee + Ff



Model GEMM using Systolic Array

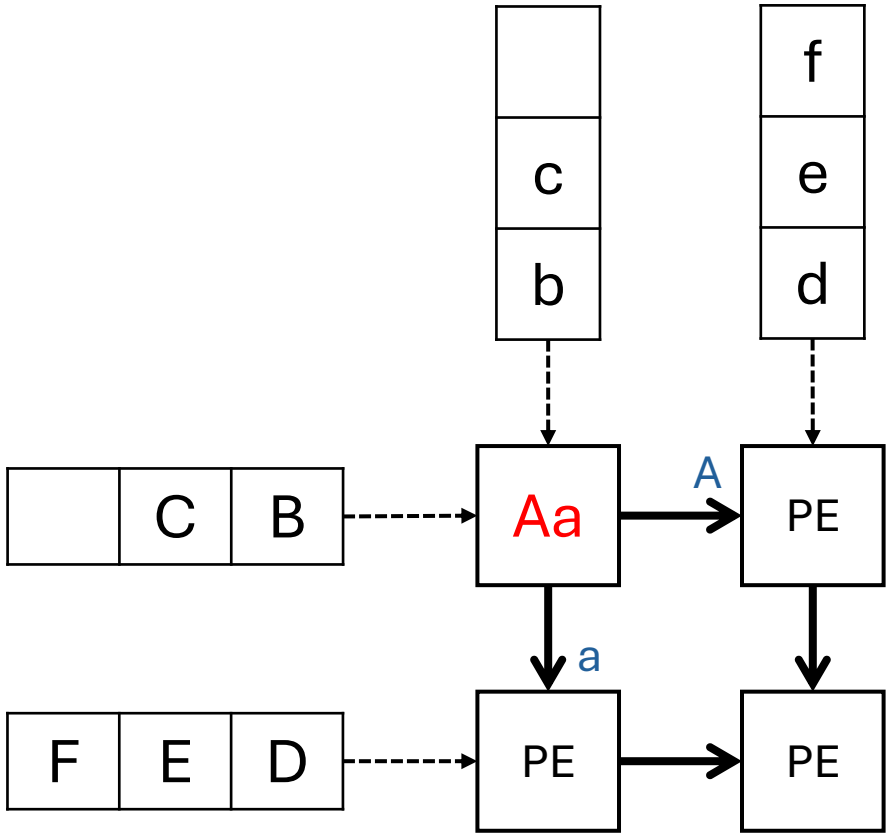
A	B	C
D	E	F

×

a	d
b	e
c	f

=

Aa + Bb + Cc	Ad + Be + Cf
Da + Eb + Fc	Dd + Ee + Ff



Model GEMM using Systolic Array

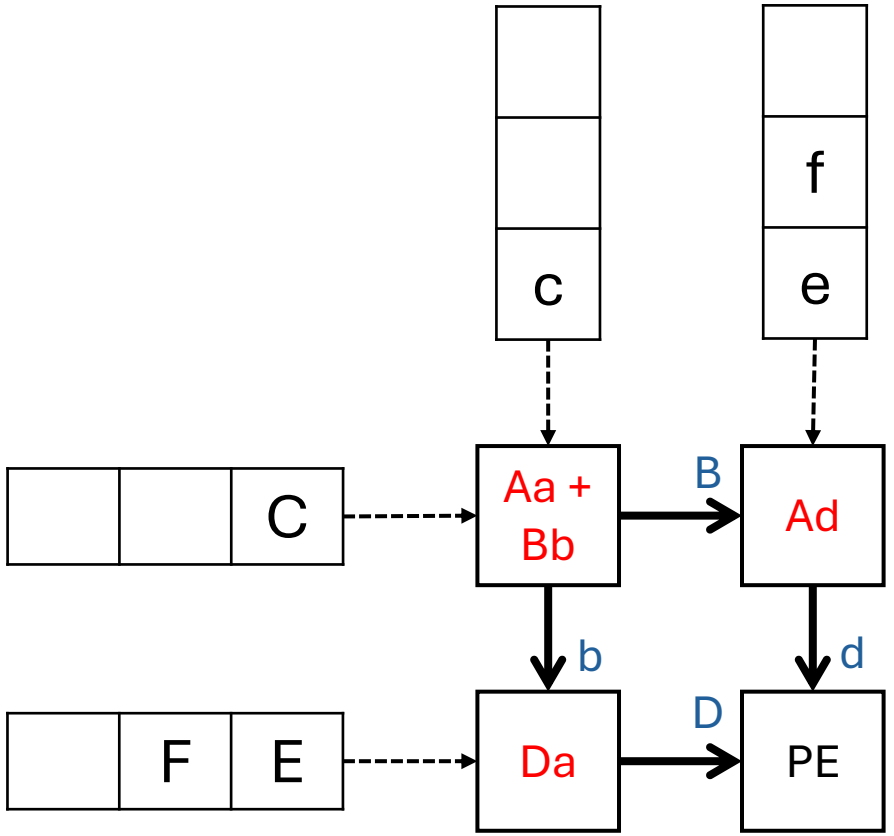
A	B	C
D	E	F

×

a	d
b	e
c	f

=

Aa + Bb + Cc	Ad + Be + Cf
Da + Eb + Fc	Dd + Ee + Ff



Model GEMM using Systolic Array

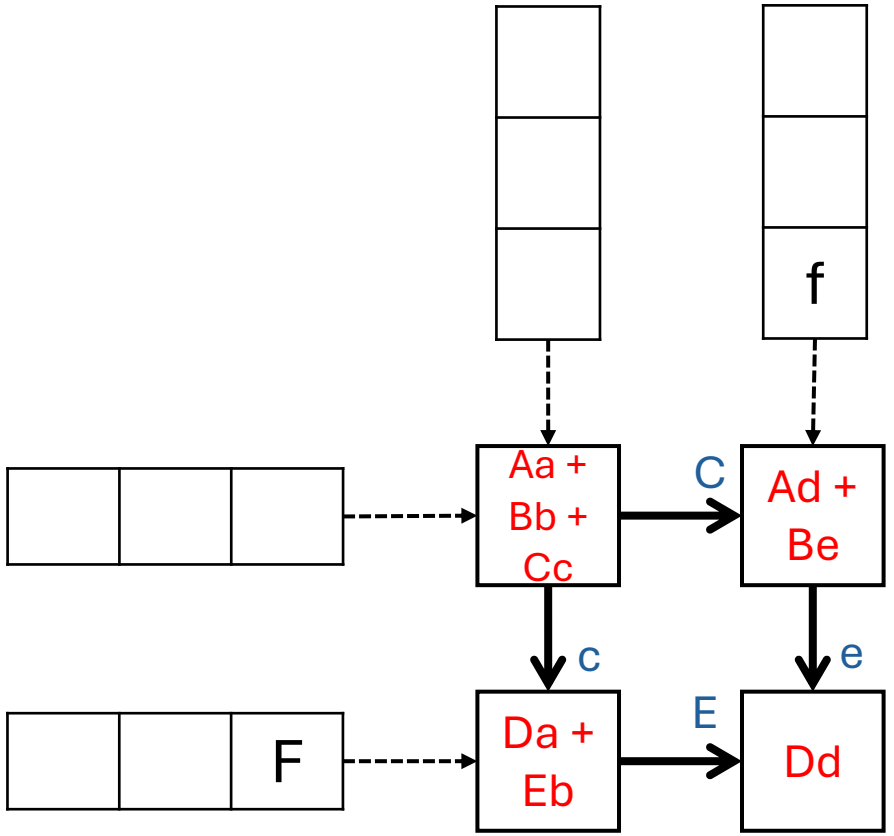
A	B	C
D	E	F

×

a	d
b	e
c	f

=

Aa + Bb + Cc	Ad + Be + Cf
Da + Eb + Fc	Dd + Ee + Ff



Model GEMM using Systolic Array

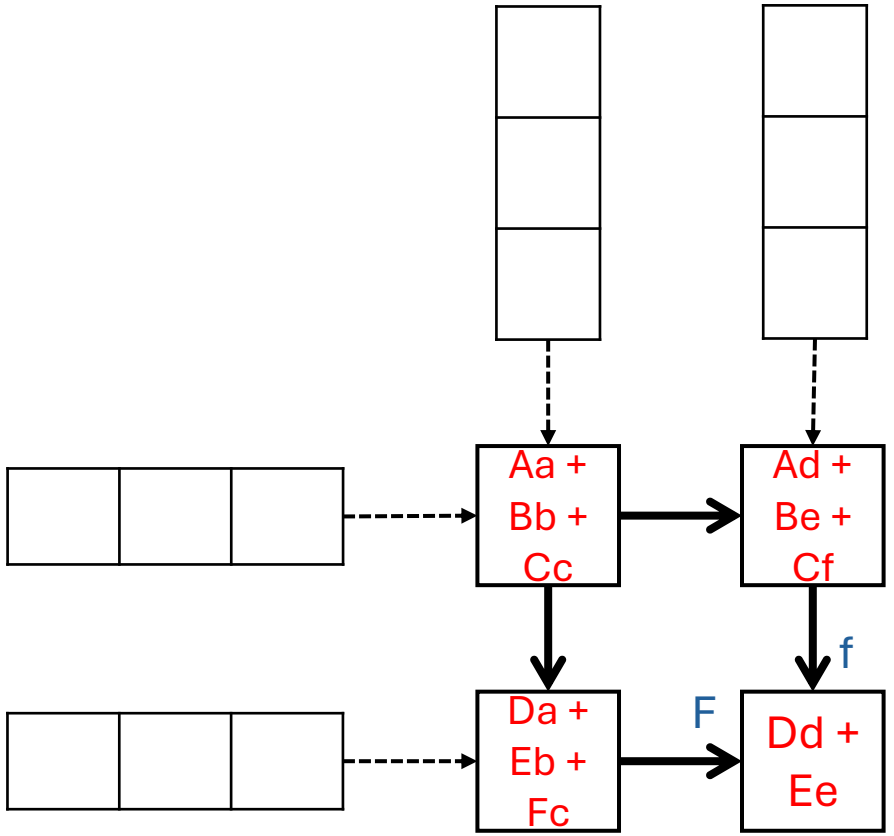
A	B	C
D	E	F

×

a	d
b	e
c	f

=

Aa + Bb + Cc	Ad + Be + Cf
Da + Eb + Fc	Dd + Ee + Ff



Model GEMM using Systolic Array

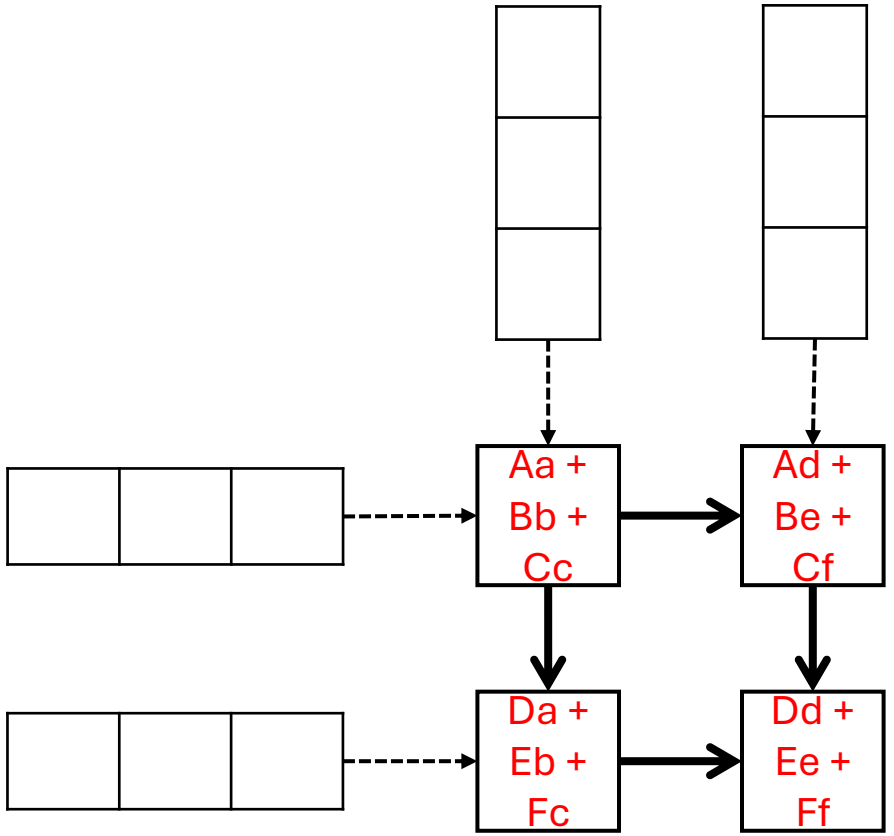
A	B	C
D	E	F

×

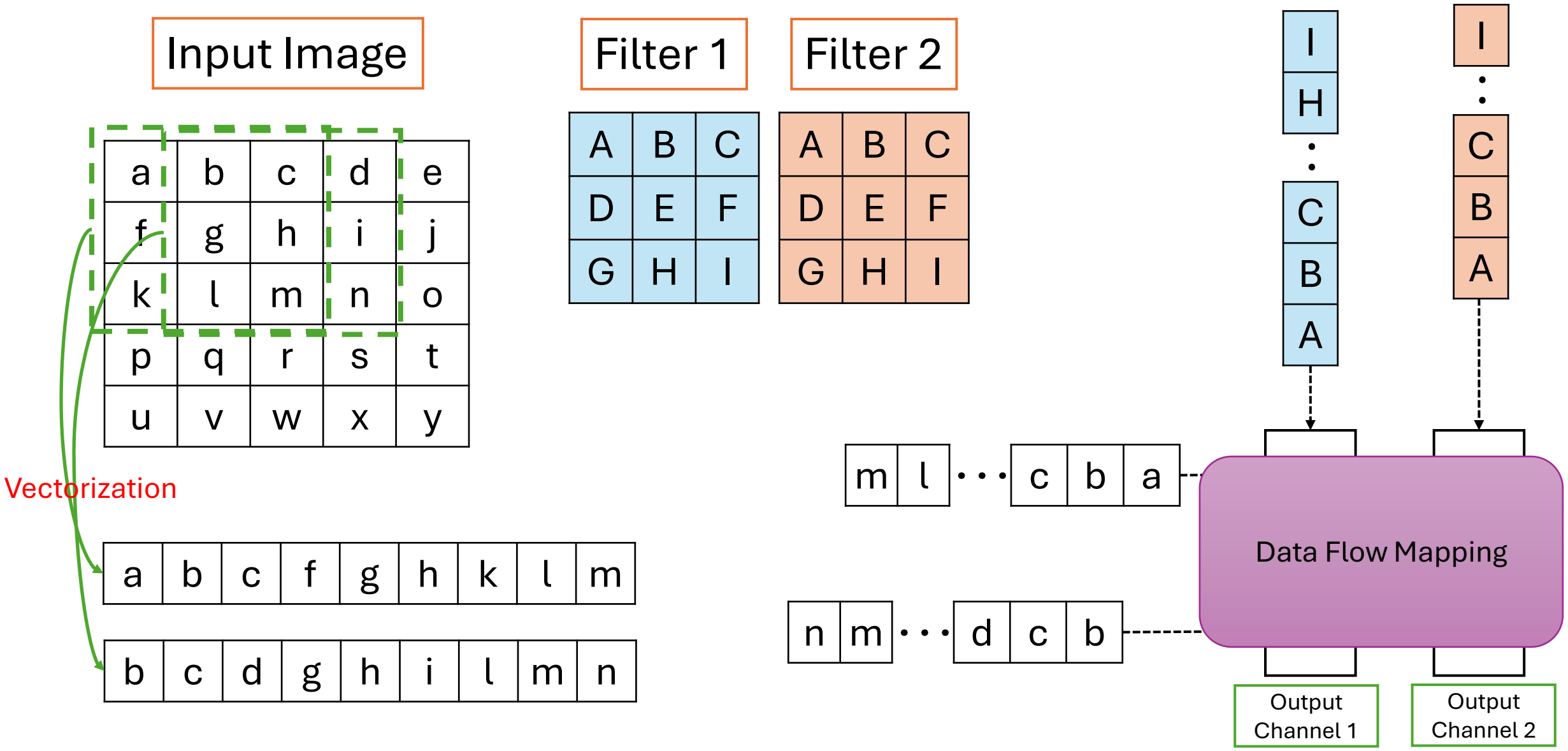
a	d
b	e
c	f

=

Aa + Bb + Cc	Ad + Be + Cf
Da + Eb + Fc	Dd + Ee + Ff



Modeling Convolutions



Machine Learning Accelerator Trace-Driven Simulator: Scale-Sim: Systolic CNN AcceLErator Simulator

- Analytical model for cycle-accurate computation timing, power/energy
  - This is for a specified accelerator hardware configuration and a neural network.
- Provides cycle-level traces for SRAM (on-chip accelerator memory) and off-chip DRAM.
- This can be used to analyze the access characteristics of several machine-learning models.
- Example Models: CNNs, RNNs, Transformers, Recommendation Models, etc.

Scale-Sim: Analytical Model of Accelerator

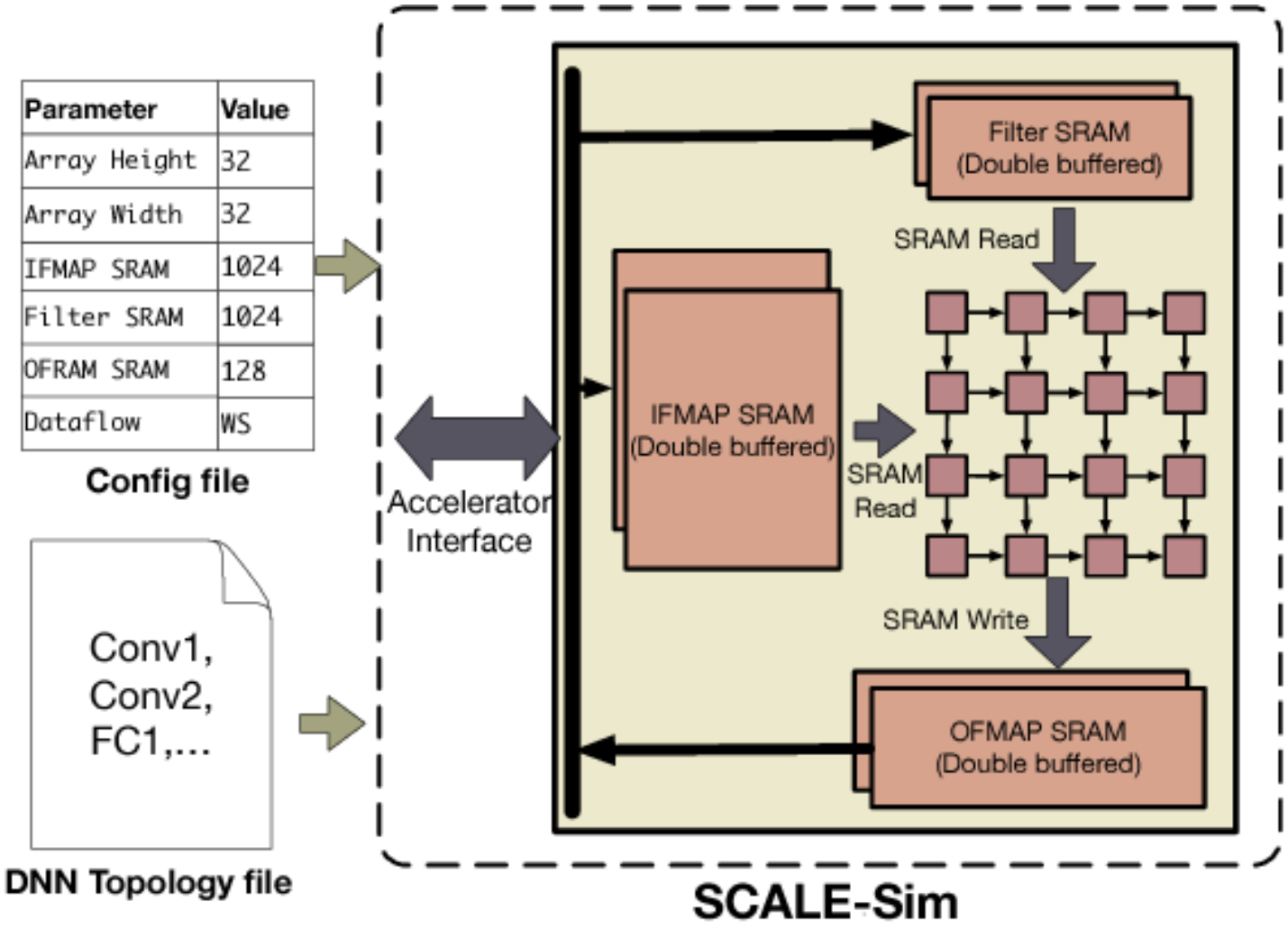
Inputs

Parameter	Value
Array Height	32
Array Width	32
IFMAP SRAM	1024
Filter SRAM	1024
OFRAM SRAM	128
Dataflow	WS

Config file

Conv1,
Conv2,
FC1,...

DNN Topology file



Outputs

Cycles,
Bandwidth,
Utilization
etc.

Simulation Summary

SRAM R/W
DRAM R/W

Cycle accurate traces

A Systematic Methodology for Characterizing Scalability of DNN Accelerators using SCALE-sim
Ananda Samajdar, Jan Moritz Joseph, Yuhao Zhu, Paul Whatmough, Matthew Mattina, and Tushar Krishna, ISPASS 2020

Scale-Sim: Analytical Model of Accelerator

- On-chip Memory Model
  - Double-buffered Memory
  - Model three sets of double-buffered memory for IFMAP, OFMAP, and FIL
- Compute Model
  - Using Systolic Array
- System Interface Model
  - Cycle-accurate DRAM traces that can be integrated into SoC
- Network Supported
  - Any network that can be mapped to Systolic Array as GEMM

Power Modeling of Accelerators

- $P = \alpha . C . V_{DD}^2 . f$ (α – Activity), $E = P . t$ (t – compute time)

- Design architecture-level power/energy estimation models.
- Provides accurate power/energy consumption analysis for accelerators.
- Predictions are based on factors like workload, architecture design, and operating conditions.
- Example of such models: Aladdin, Accelergy

Power Modeling Advantages

- Allows designers to optimize for energy efficiency during the design phase.
- Power Optimization
  - Identify design bottlenecks
  - Optimize parameters like clock frequency or memory hierarchy to minimize power consumption.
- Thermal management:
  - Predict power dissipation to analyze the temperature/thermal map of the Accelerator
  - Design efficient cooling methods for the accelerator

Energy Model using Scale-Sim

From Scale-Sim

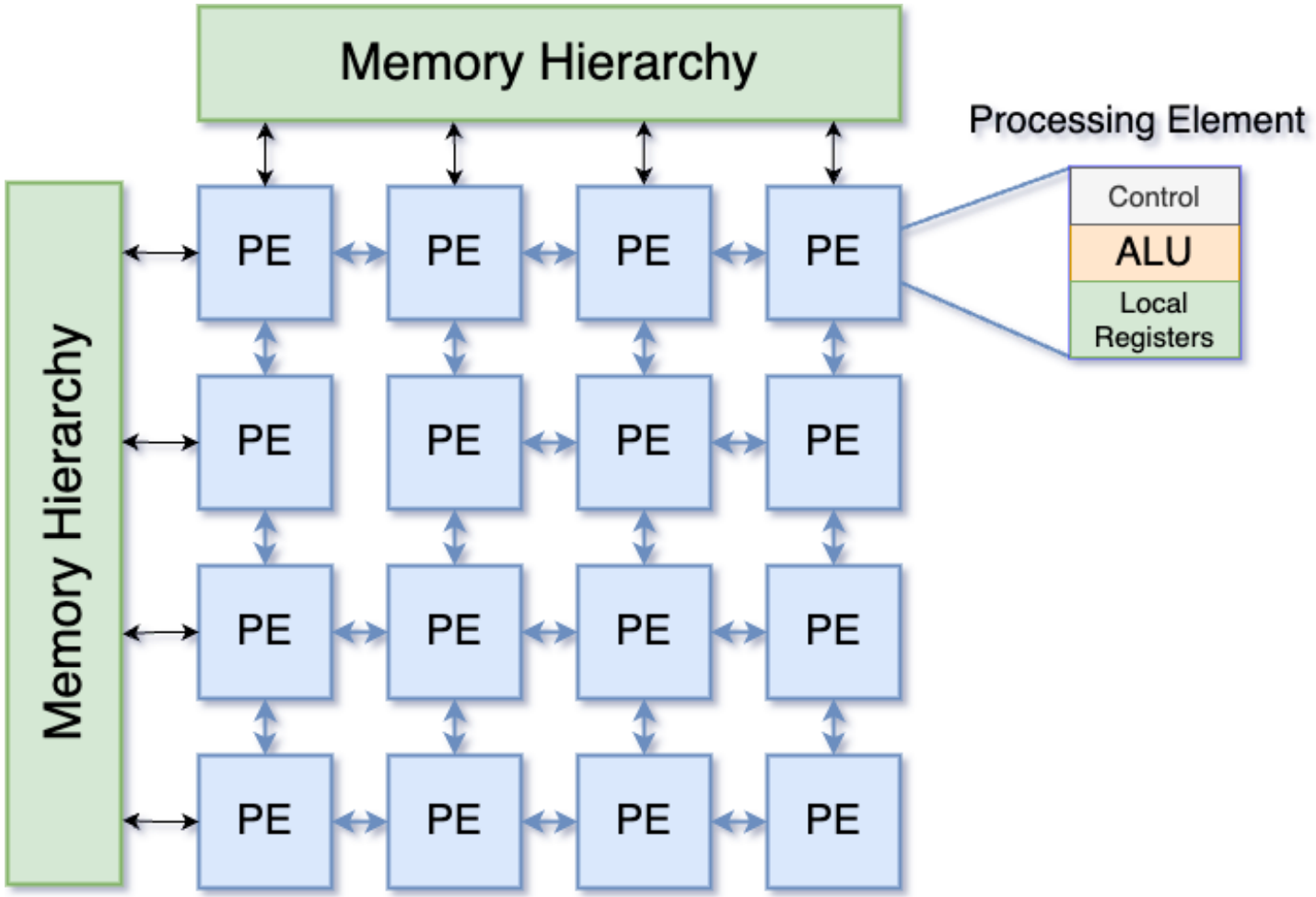
Access Counts for each SRAM buffer

×

Energy/Access (for SRAM)

=

Total Dynamic Energy of Memory



From Scale-Sim

Access Counts for each PE unit

×

Energy/Access (based on component)

=

Total Dynamic Energy of Each PE

Dynamic Voltage and Frequency Scaling (DVFS)

- DVFS is used to reduce energy consumption.

- Dynamic Scaling of the supply voltage and frequency of the chip
  - Scaling based on workload
  - Scaling done during idle periods or low computational intensity periods

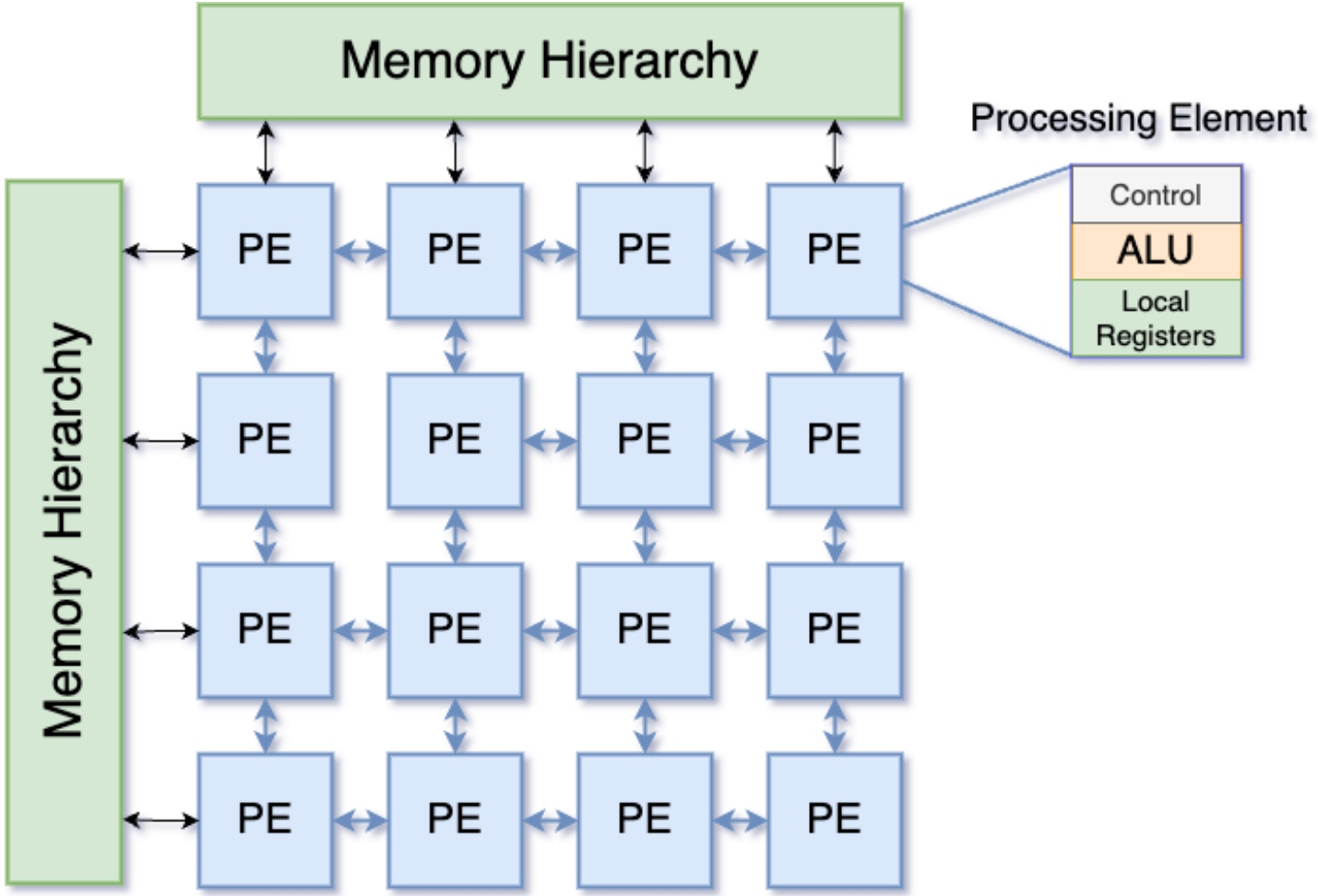
- Performance Trade-offs:
  - Accurate workload prediction – low and high compute intensity periods
    - Should not optimize high compute zones
  - Efficient voltage switching mechanisms
    - High switching latency – high overheads

DVFS Policy using Scale-Sim

From Scale-Sim

Interval-wise Access Counts for SRAM buffers

DVFS on Banks with low-activity



From Scale-Sim

Interval-wise Access Counts for each PE unit

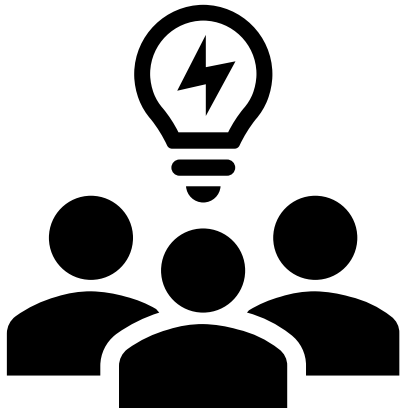
DVFS on PE's with low activity

Projects

- Design a DVFS policy for the accelerator, based on
  - Progress compared to a given deadline (time, in *ms*)
    - If AHEAD of the deadline, perform DVFS to optimize the accelerator's energy
  - or/and, Accelerator Activity during the computation
    - Based on activity in the accelerator, perform DVFS on components with low utilization or activity
- Design a Dynamic Thermal Management Policy for the Accelerator using Scale-Sim and Hotspot.
  - Collect interval-wise access counts for each component from Scale-Sim
  - Get a power trace from Accelergy
  - Use HotSpot to obtain a Thermal Map – identify the hot regions
  - Use DVFS to optimize temperature – maybe slow down the hot regions

Thank You

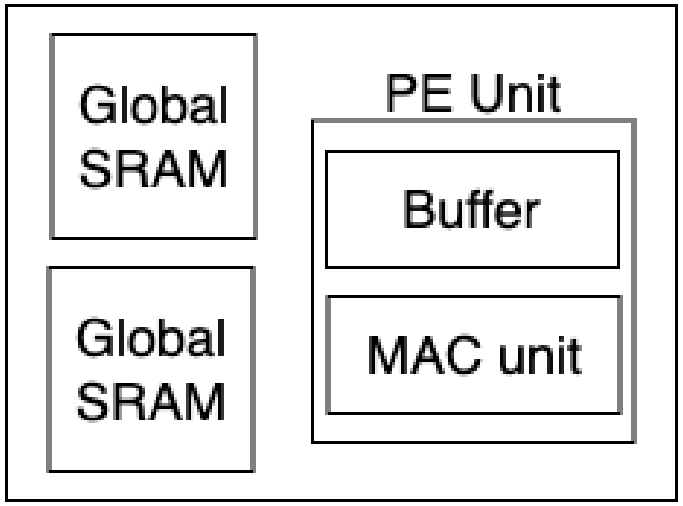
Questions?



Accelerator Energy Model – Accelergy tool

- Accelergy is an architecture-level energy estimator tool.
- It uses the action counts from the performance simulator to estimate the total energy.

Sample Accelerator



From Scale-Sim or Accelerator Performance Simulator

Module	Action	Counts
Global SRAM	access()	x1
Buffer	access()	x2
MAC unit	compute()	x3

Module	Energy/action
Global SRAM	y1
Buffer	y2
MAC unit	y3

From Tech libraries (22nm)

Energy Estimator = x1*y1 + x2*y2 + x3*y3

Total Accelerator Energy Estimate