http://synergy.ece.gatech.edu

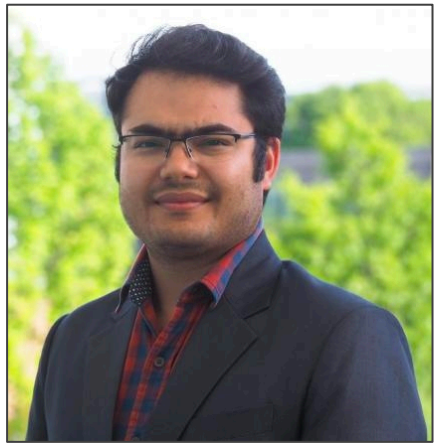
SCALE-Sim: Systolic CNN accelerator simulator

Tutorial @ ASPLOS 2021
April 16 2021

https://scalesim-project.github.io/
https://scalesim-project.github.io/tutorials-2021-asplos.html

Presenters

Tushar Krishna
Associate Professor
Georgia Tech



Ananda Samakdar
PhD Student
Georgia Tech

Jan Moritz Joseph
Post Doc Researcher
RWTH Aachen University

Paul Whatmough
Principal Research Engineer
ARM Research, Boston

Yuhao Zhu
Assistant Professor
Univ. of Rochester

Schedule (EST)

Time slot	Topic	
10:05 – 10:10	Welcome and Overview	Tushar
10:10 to 10:45	Introduction to DNNs and Accelerator Design	Paul & Tushar
10:45 to 11:15	Overview of SCALE-Sim	Anand + Moritz + Paul
11:15 to 11:50	Tutorial 1: Design Space Exploration using SCALE-Sim	Anand
12:00 to 12:40	Tutorial 2: Modifying SCALE-Sim to add custom features	Moritz
12:45 to 1:30	Tutorial 3: Using SCALE-Sim to build larger simulators	Anand
1:30 to 2:00	Discussion on future roadmap, planned features, and ideas from the community	Yuhao

Brief Q/A at the end of each talk.

Attention: Tutorial is being recorded!

Slides + Videos will be available on the SCALE-sim tutorial website

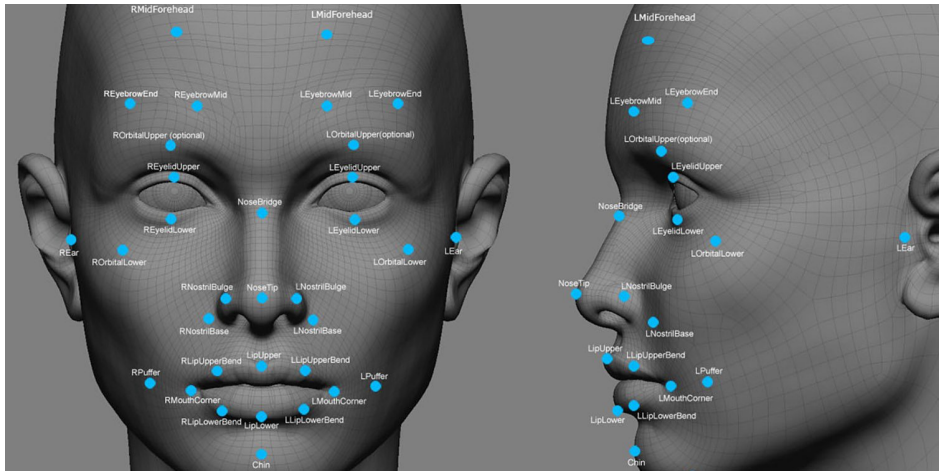
https://scalesim-project.github.io/tutorials-2021-asplos.html

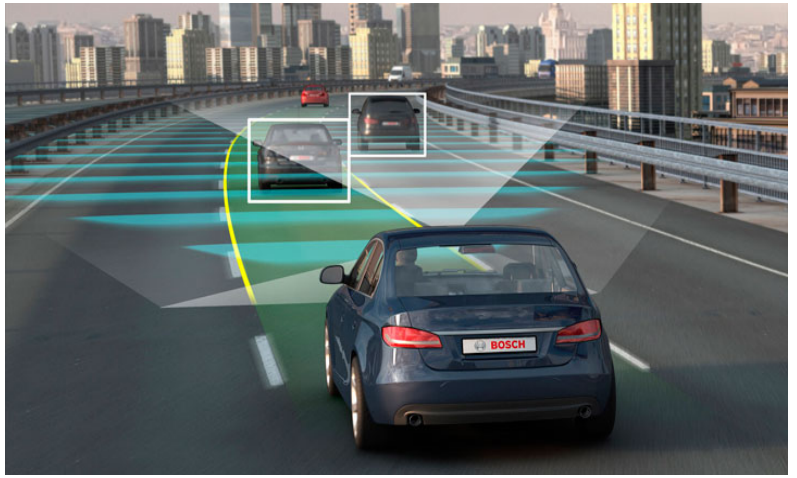
Outline

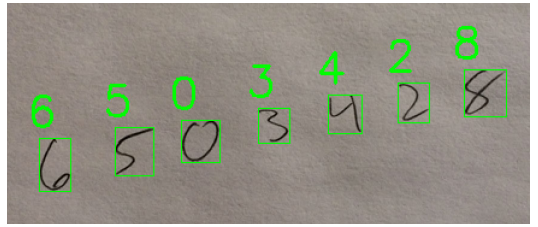
- **What this tutorial is about**
  - Role within the Deep Learning landscape
- **Resources**
  - SCALE-sim
  - Synthesis Lectures
- **Relevant Background**
  - Deep Learning Landscape
  - DNN Accelerators

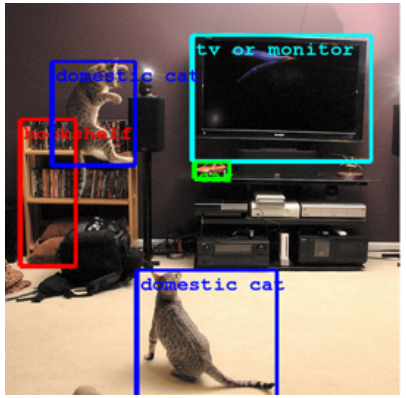
What this tutorial is about

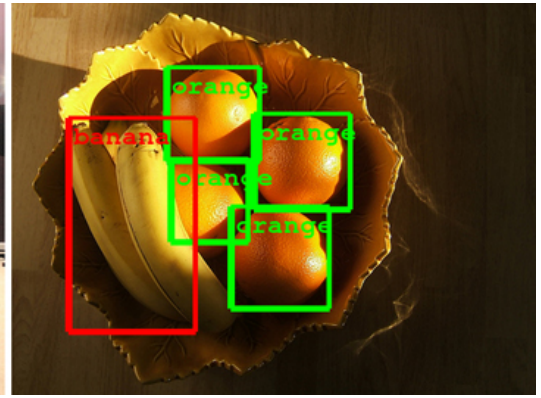
Deep Learning Applications

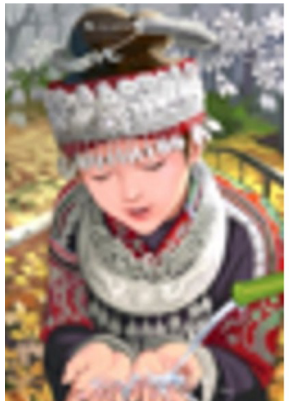


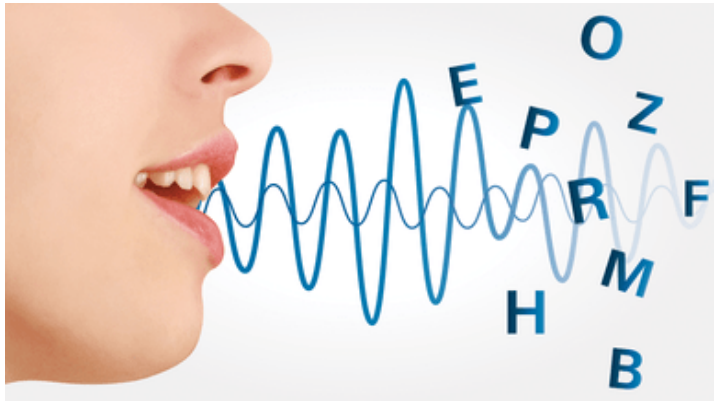




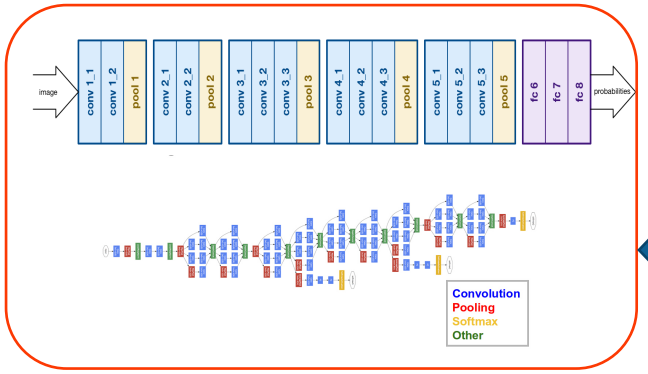








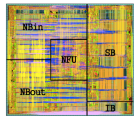
Deep Learning Landscape



Model Creation

Training

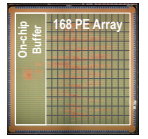
Inference



Apple Neural Engine

ShiDIanNao

NVDLA



ARM Trillum

CambriconX

Eyeriss

TensorFlow

Caffe

cuDNN

intel Software

MLSL

NVIDIA

TensorRT

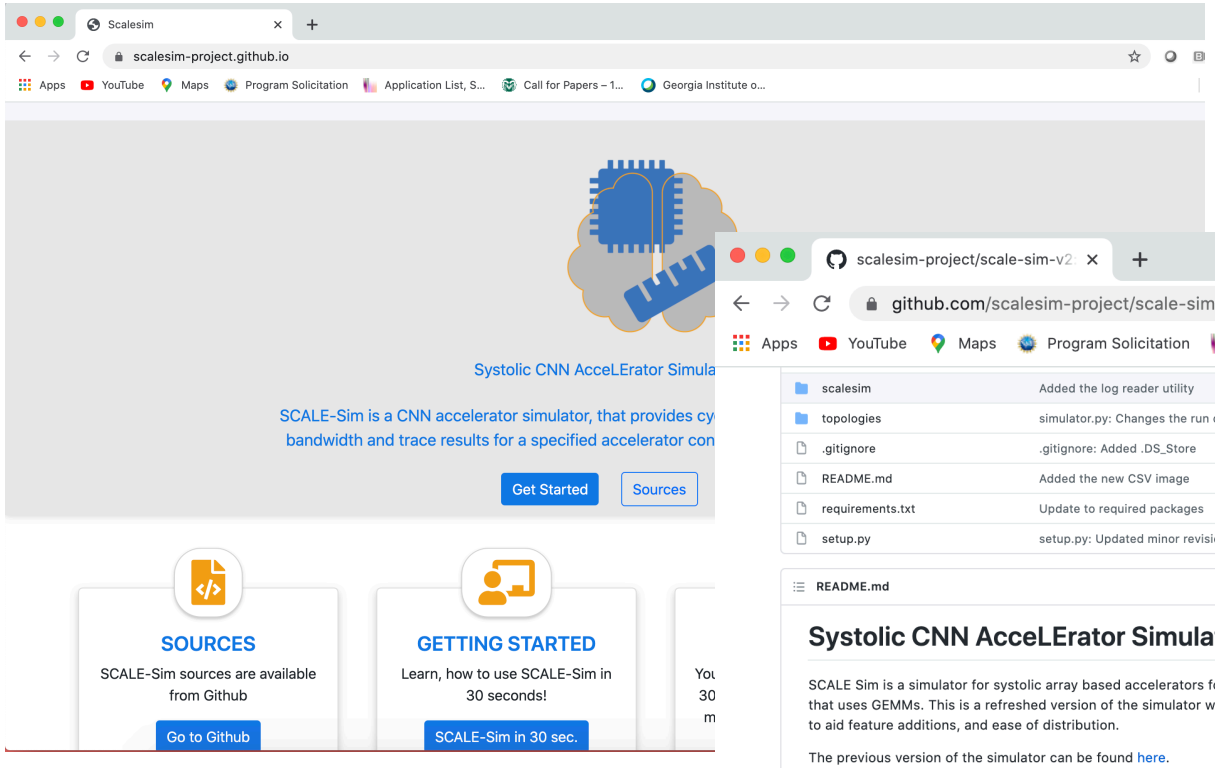
SCALE-sim

This Tutorial

Design Tools

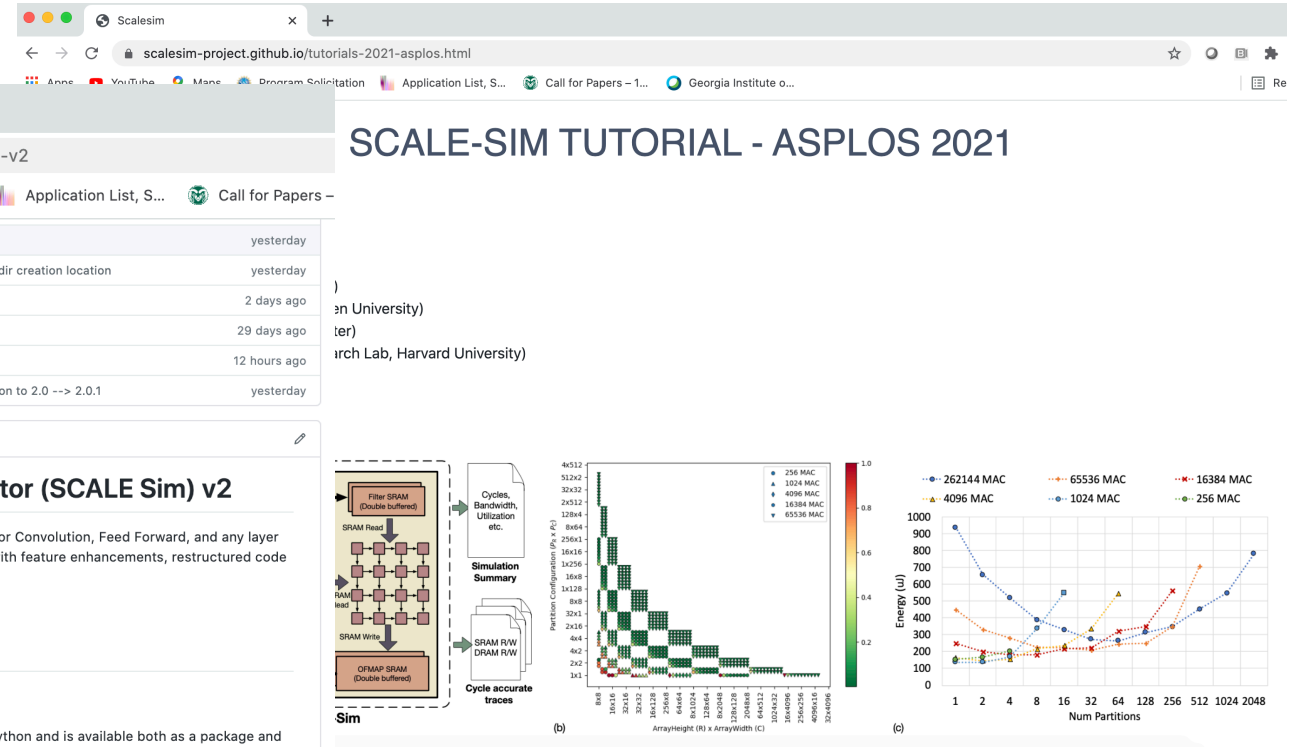
Resources

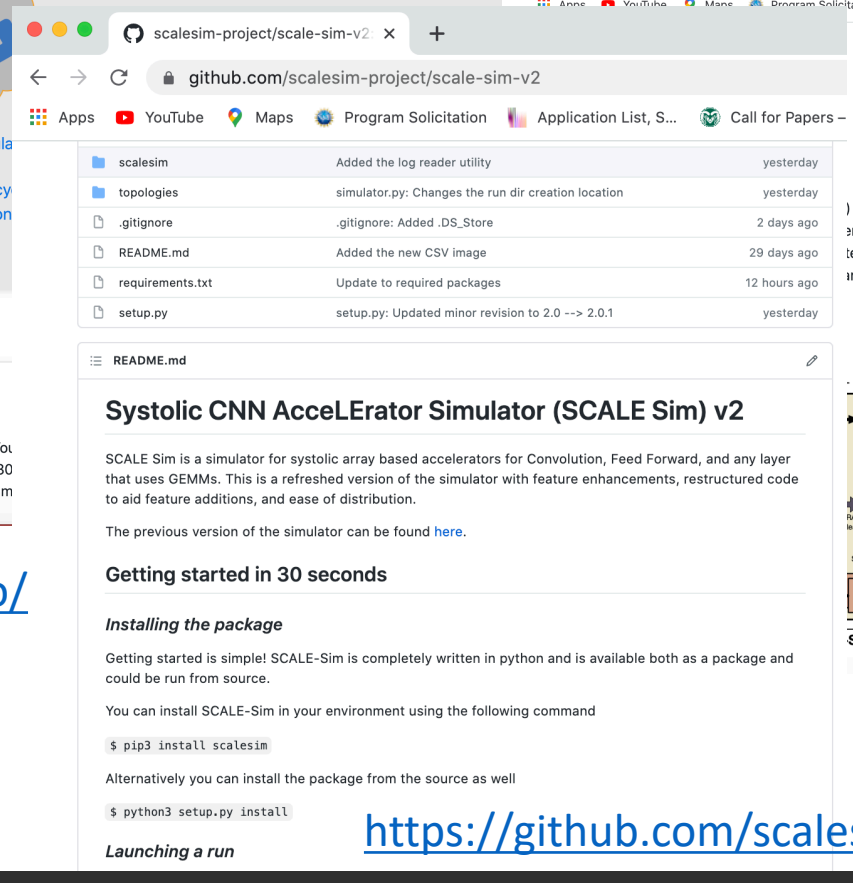
SCALE-sim Webpage, Github and Tutorial



https://scalesim-project.github.io/

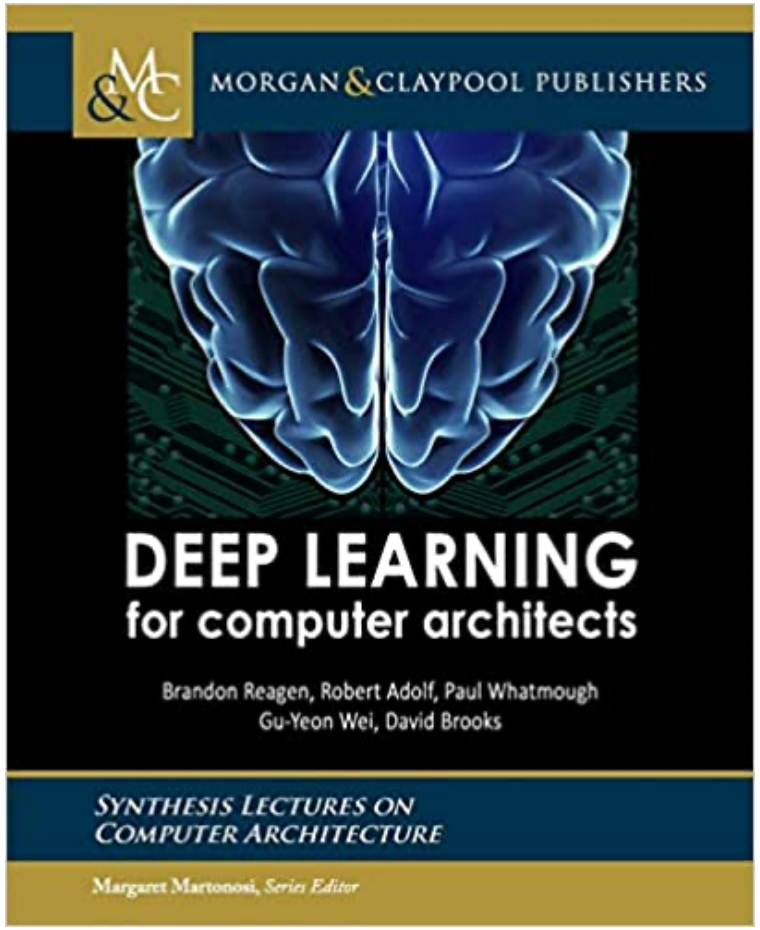
https://scalesim-project.github.io/tutorials-2021-asplos.html

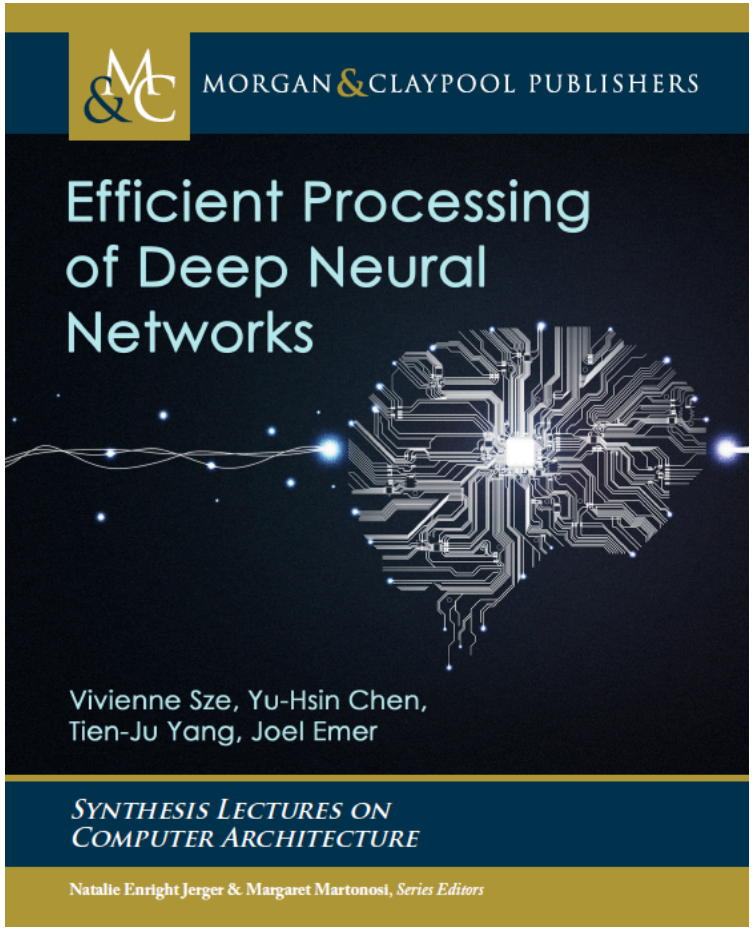


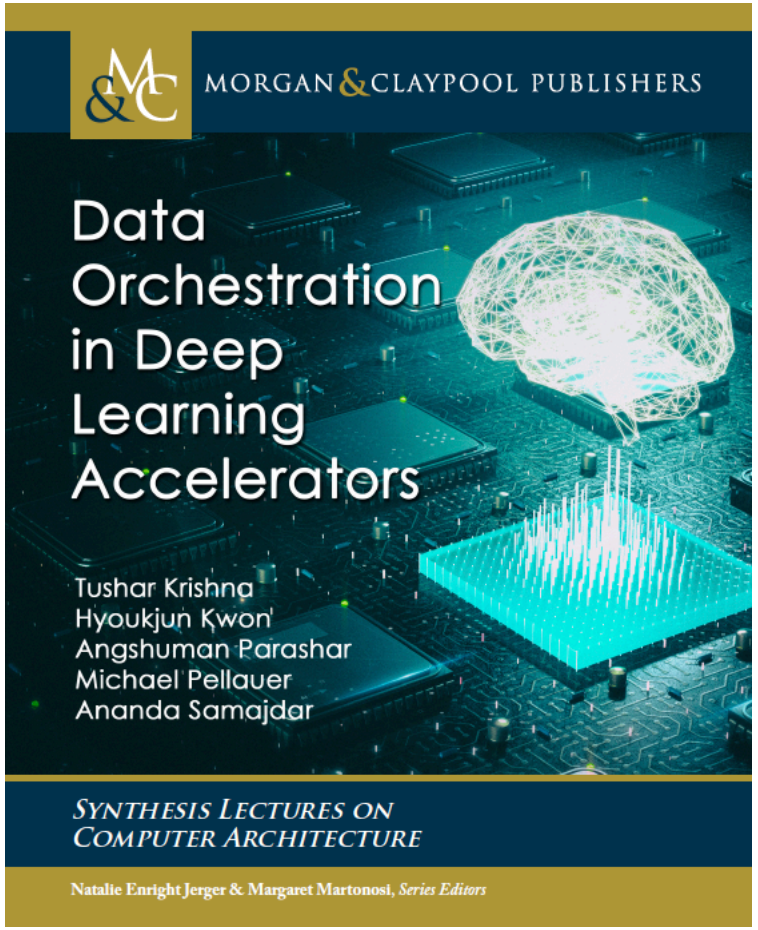


https://github.com/scalesim-project/scale-sim-v2

Synthesis Lectures







Welcome to the Tutorial!

Schedule (EST)

Time slot	Topic	
10:05 – 10:10	Welcome and Overview	Tushar
10:10 to 10:45	Introduction to DNNs and Accelerator Design	Paul & Tushar
10:45 to 11:15	Overview of SCALE-Sim	Anand + Moritz + Paul
11:15 to 11:50	Tutorial 1: Design Space Exploration using SCALE-Sim	Anand
12:00 to 12:40	Tutorial 2: Modifying SCALE-Sim to add custom features	Moritz
12:45 to 1:30	Tutorial 3: Using SCALE-Sim to build larger simulators	Anand
1:30 to 2:00	Discussion on future roadmap, planned features, and ideas from the community	Yuhao

Brief Q/A at the end of each talk.

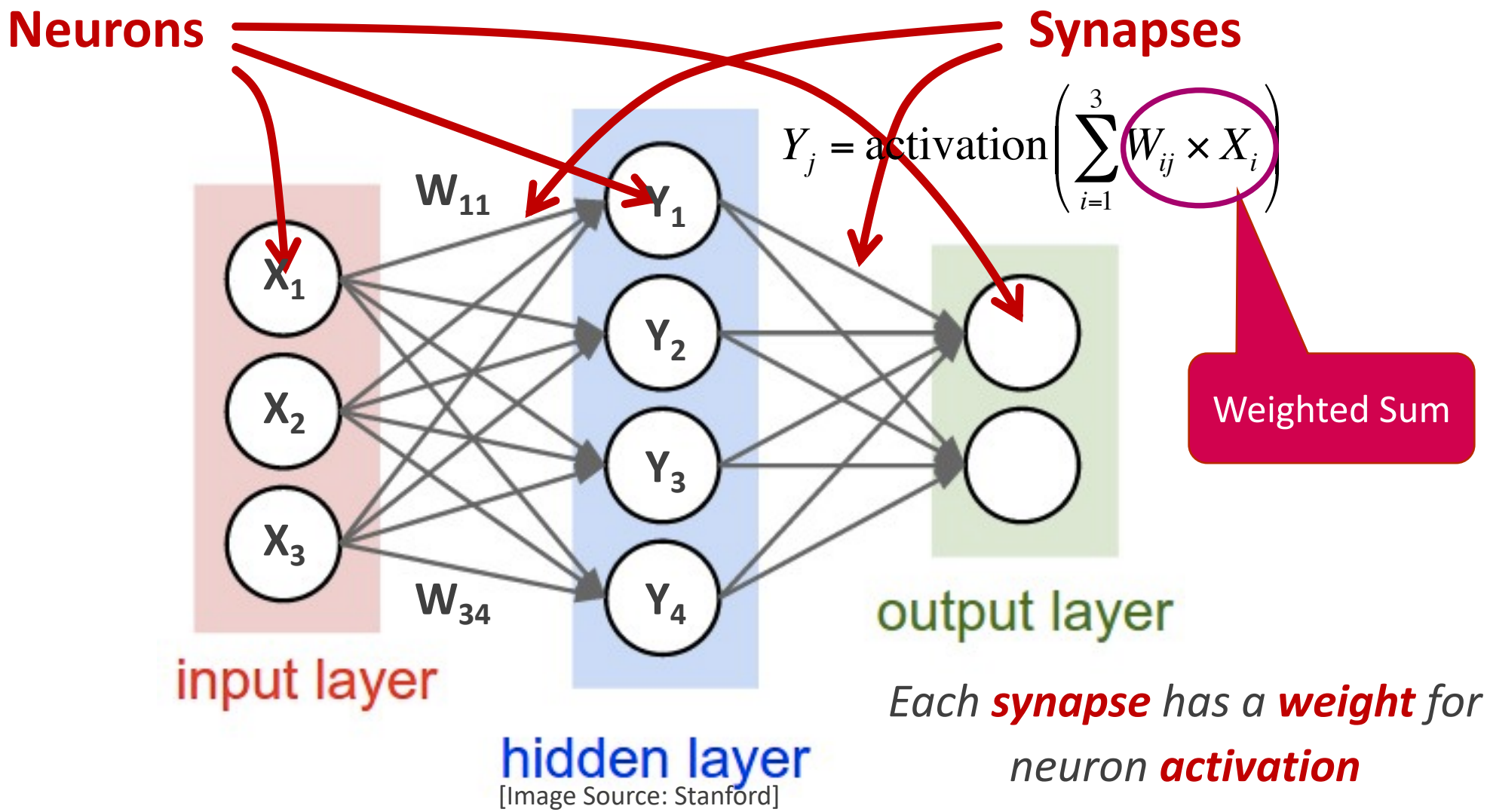
Attention: Tutorial is being recorded!

Slides + Videos will be available on the SCALE-sim tutorial website

https://scalesim-project.github.io/tutorials-2021-asplos.html

Relevant Background

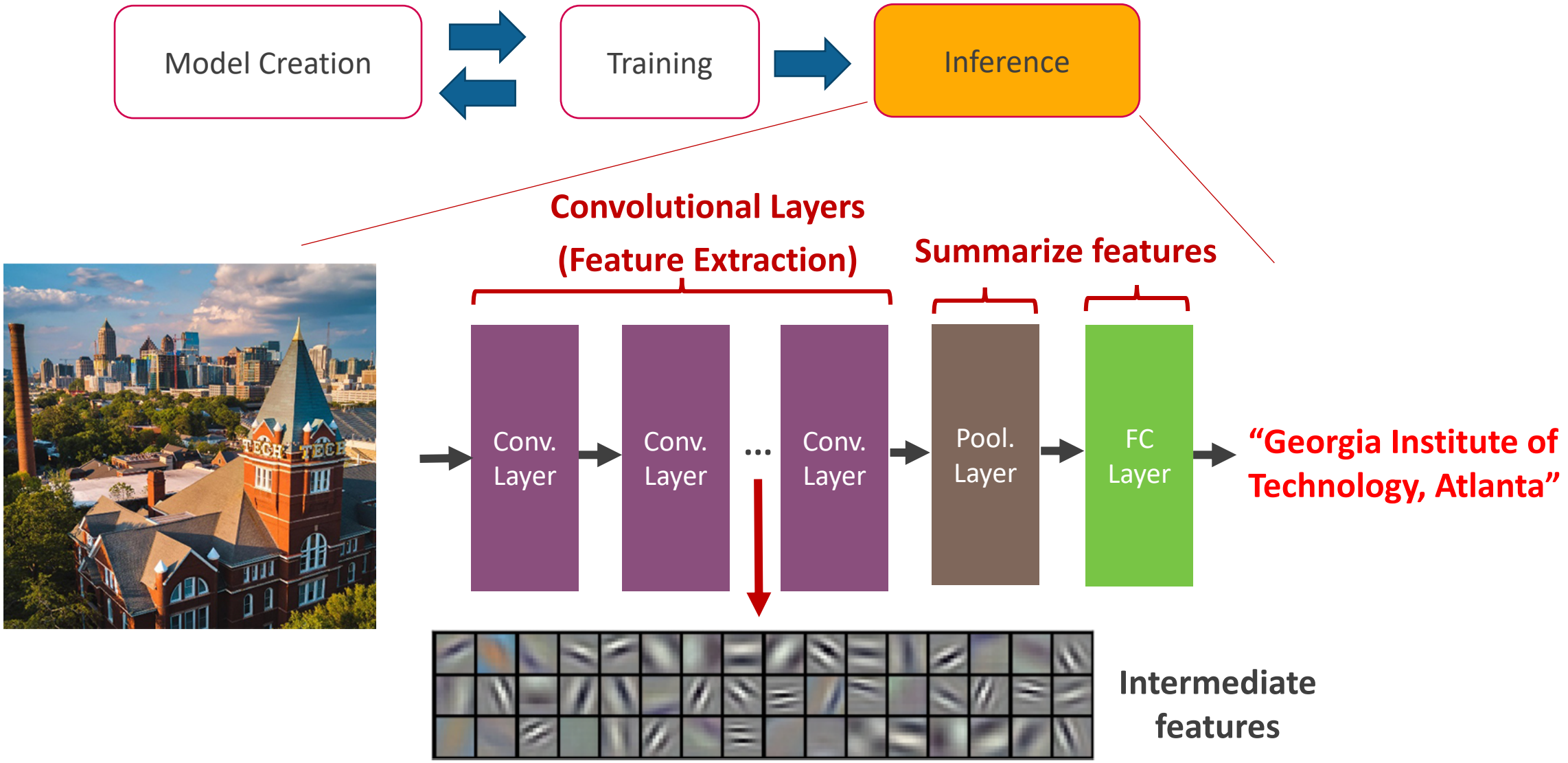
What is a Deep Neural Network?



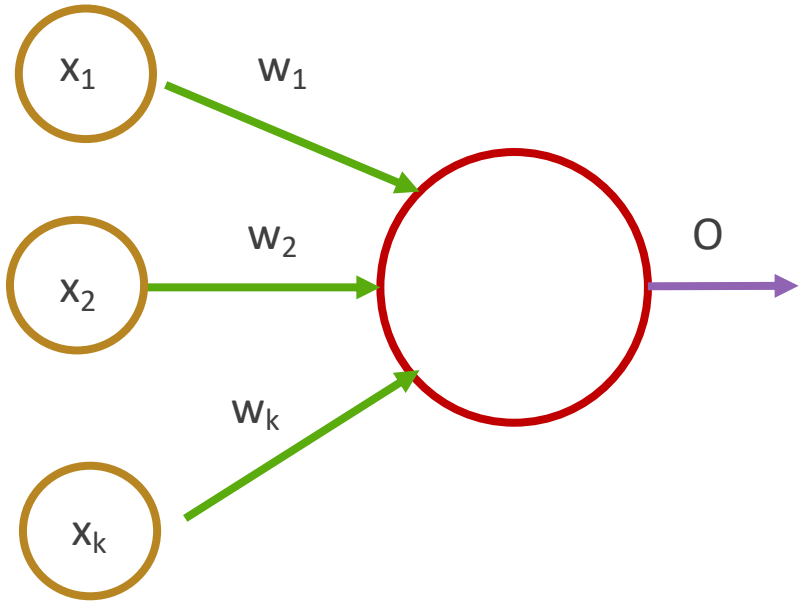
[Image Source: Stanford]

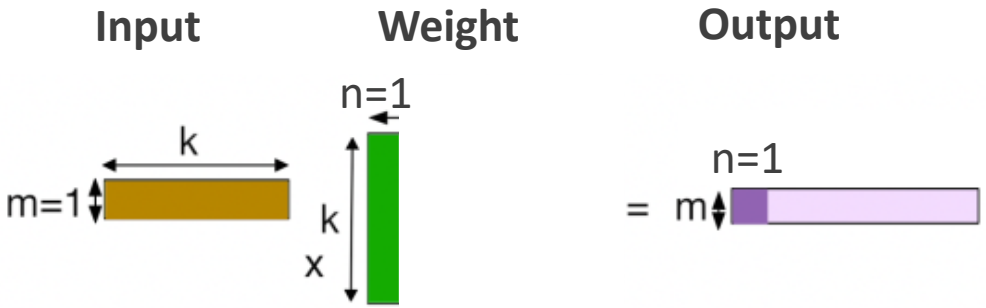
*Each **synapse** has a **weight** for neuron **activation***

Modern Deep Learning Landscape



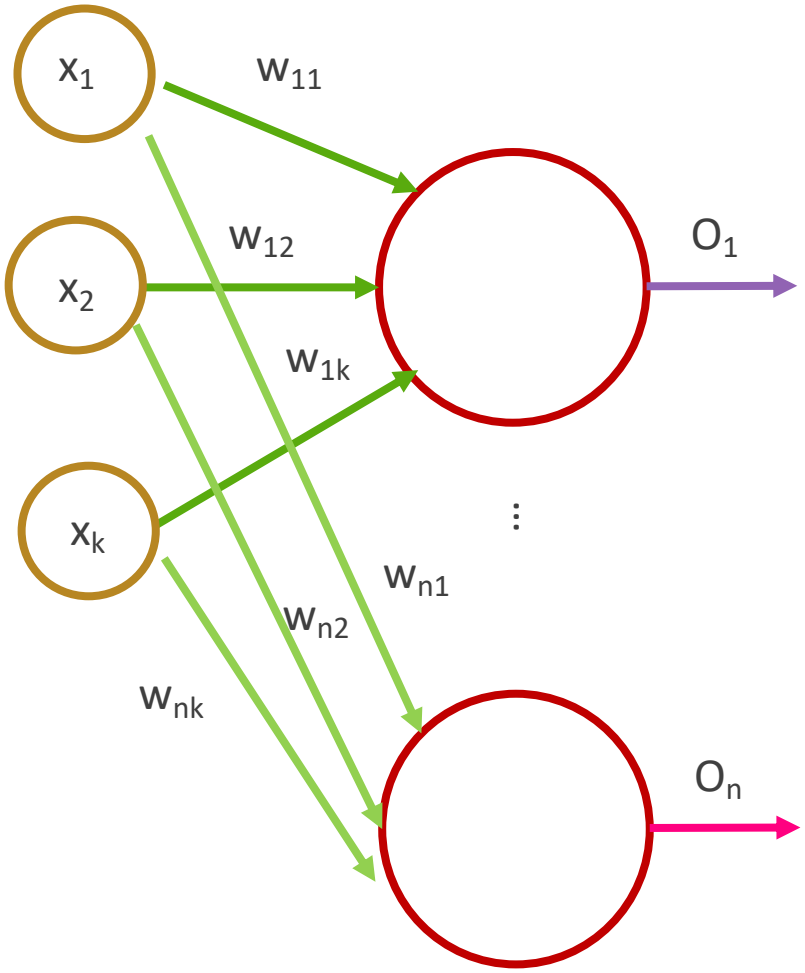
Computations in a DNN → Linear Algebra

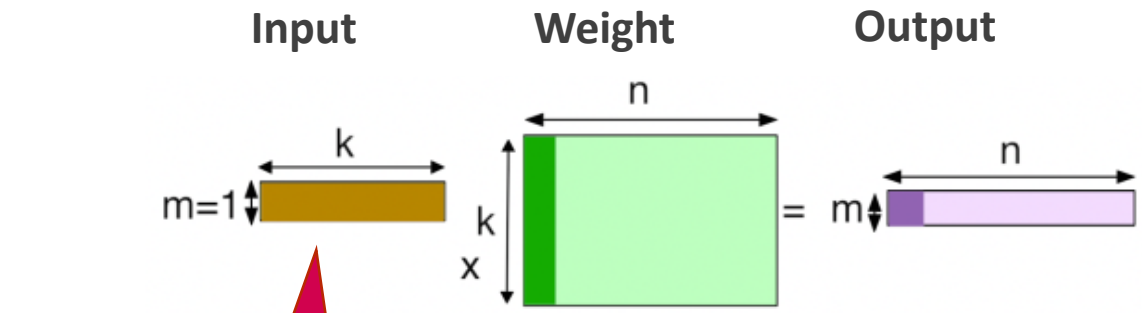




Neuron => Vector x Vector

Computations in a DNN → Linear Algebra

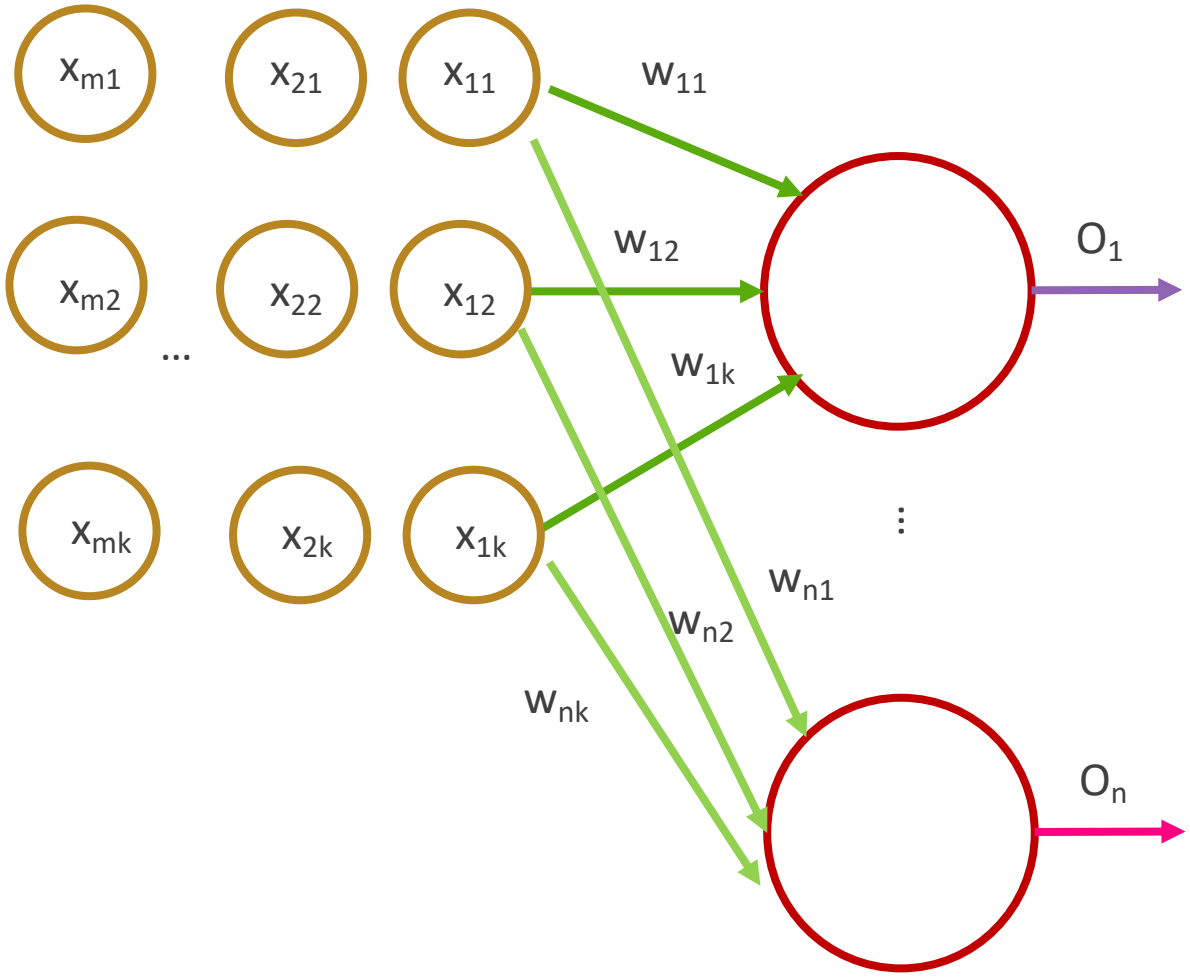


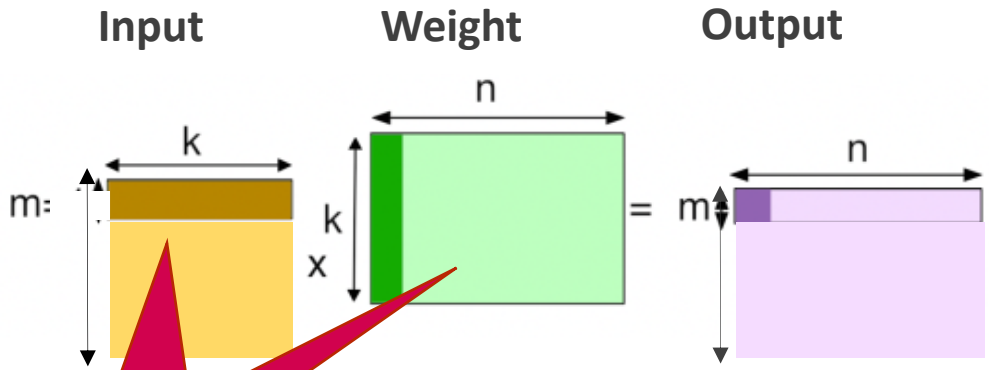


DNN Layer => Vector x Matrix

Data "Reuse"

Computations in a DNN → Linear Algebra



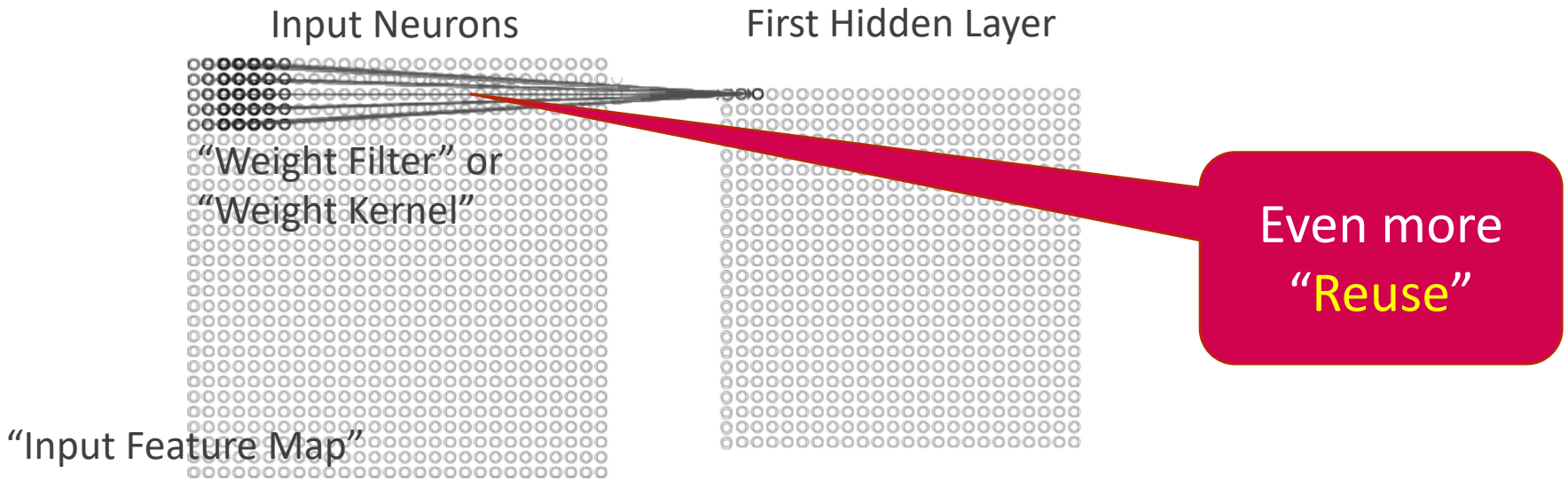


Batching => Matrix x Matrix

Data "Reuse"

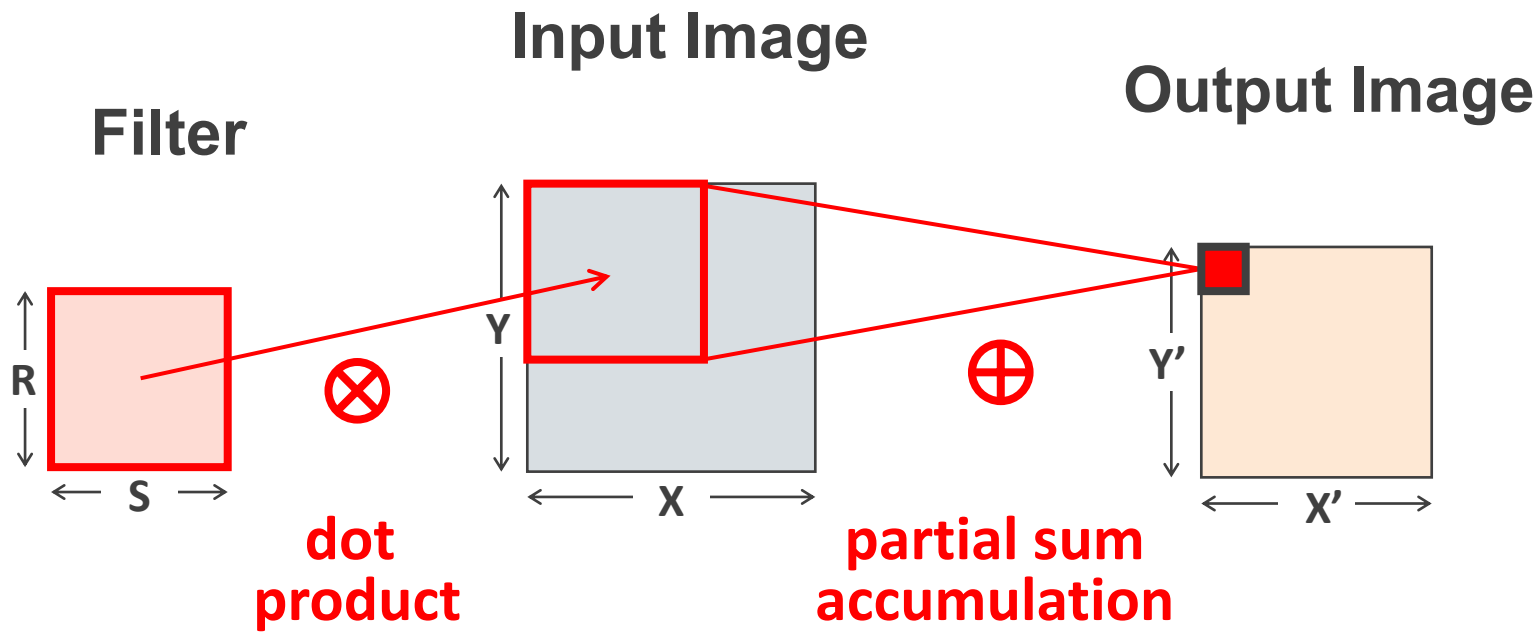
GEMM

Convolutional Neural Networks

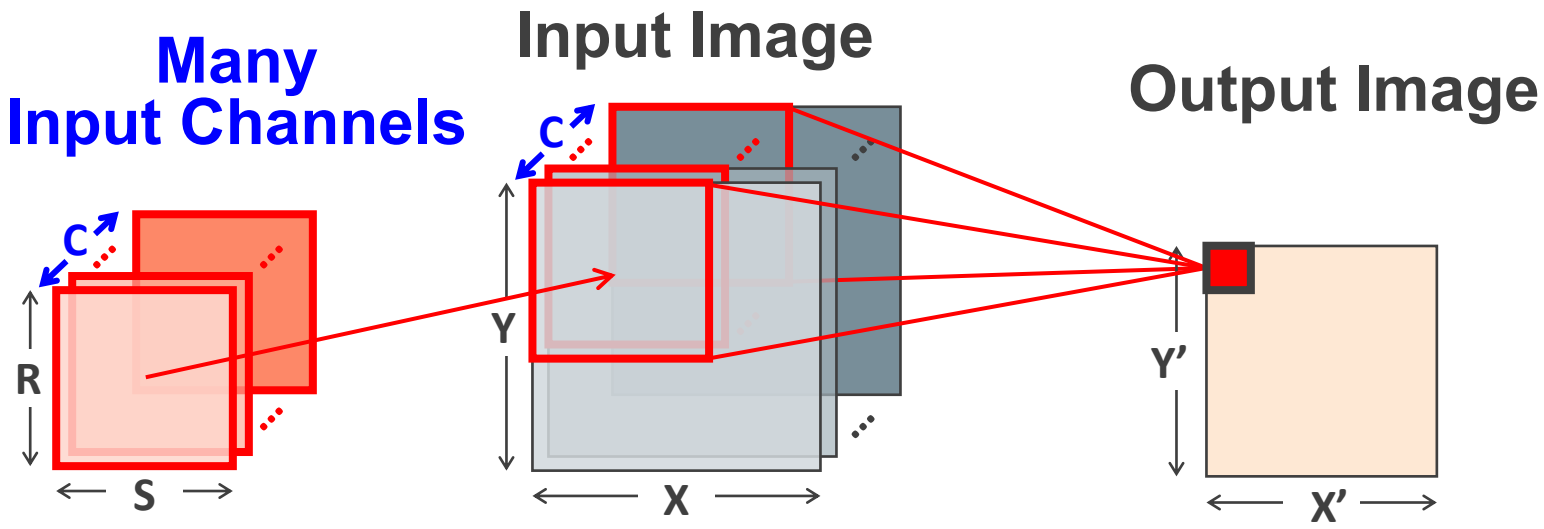


Shared Weights:
All neurons use the *same* filter weights

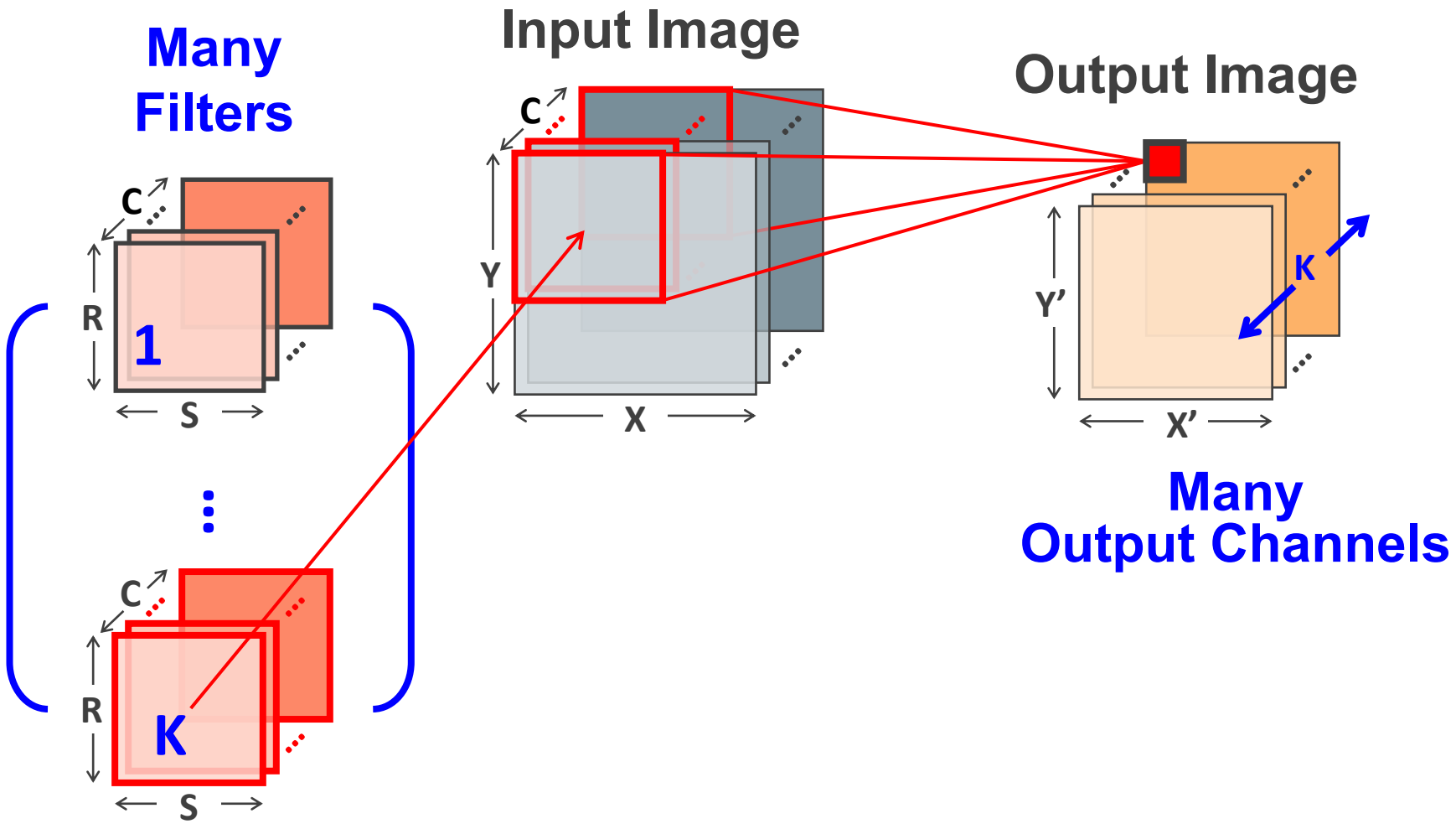
Convolution in CNN



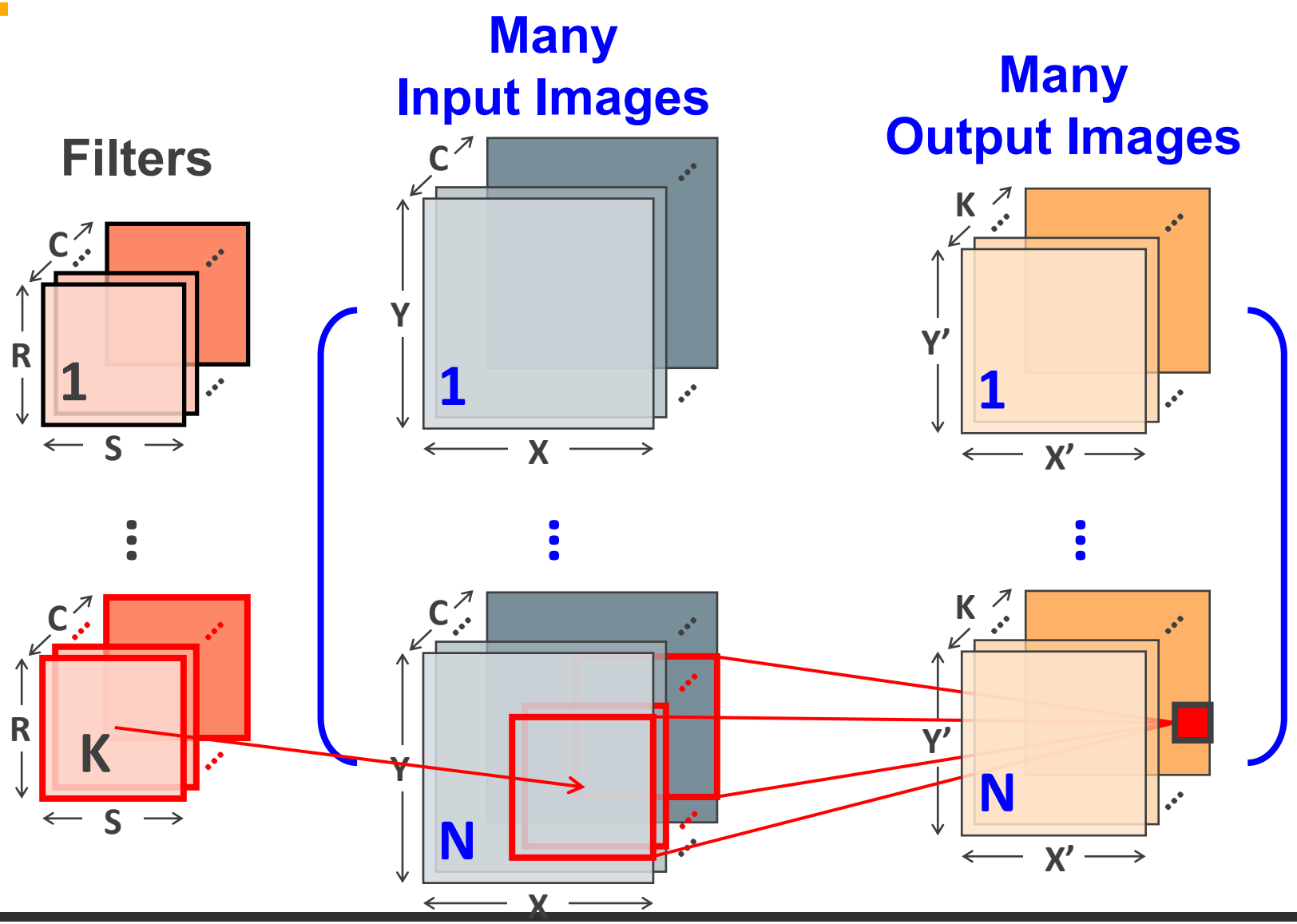
Convolution in CNN



Convolution in CNN



Convolution in CNN



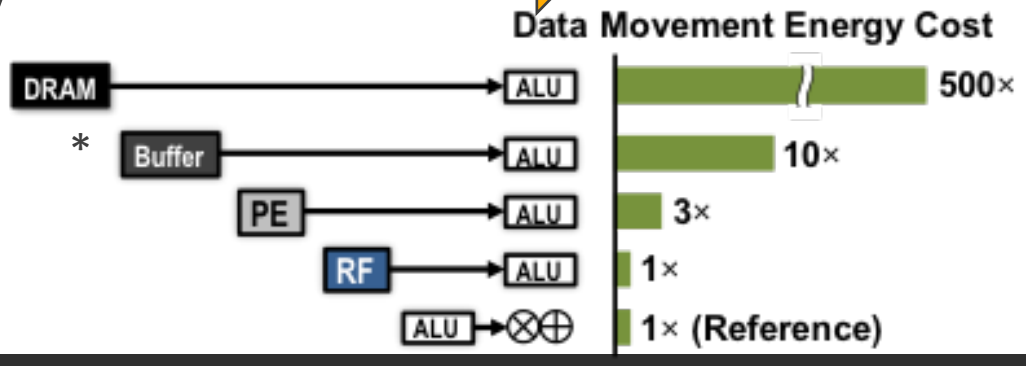
Challenges with DNN Computations

- **Millions of Parameters (i.e., weights)**
  - Billions of computations → **Need lots of parallel computations**

DNN Topology	Number of Weights
AlexNet (2012)	3.98M
VGGnet-16 (2014)	28.25M
GoogleNet (2015)	6.77M
Resnet-50 (2016)	23M
DLRM (2019)	540M
Megatron (2019)	8.3B

This makes CPUs inefficient

  - Heavy data movement → **Need to reduce energy**

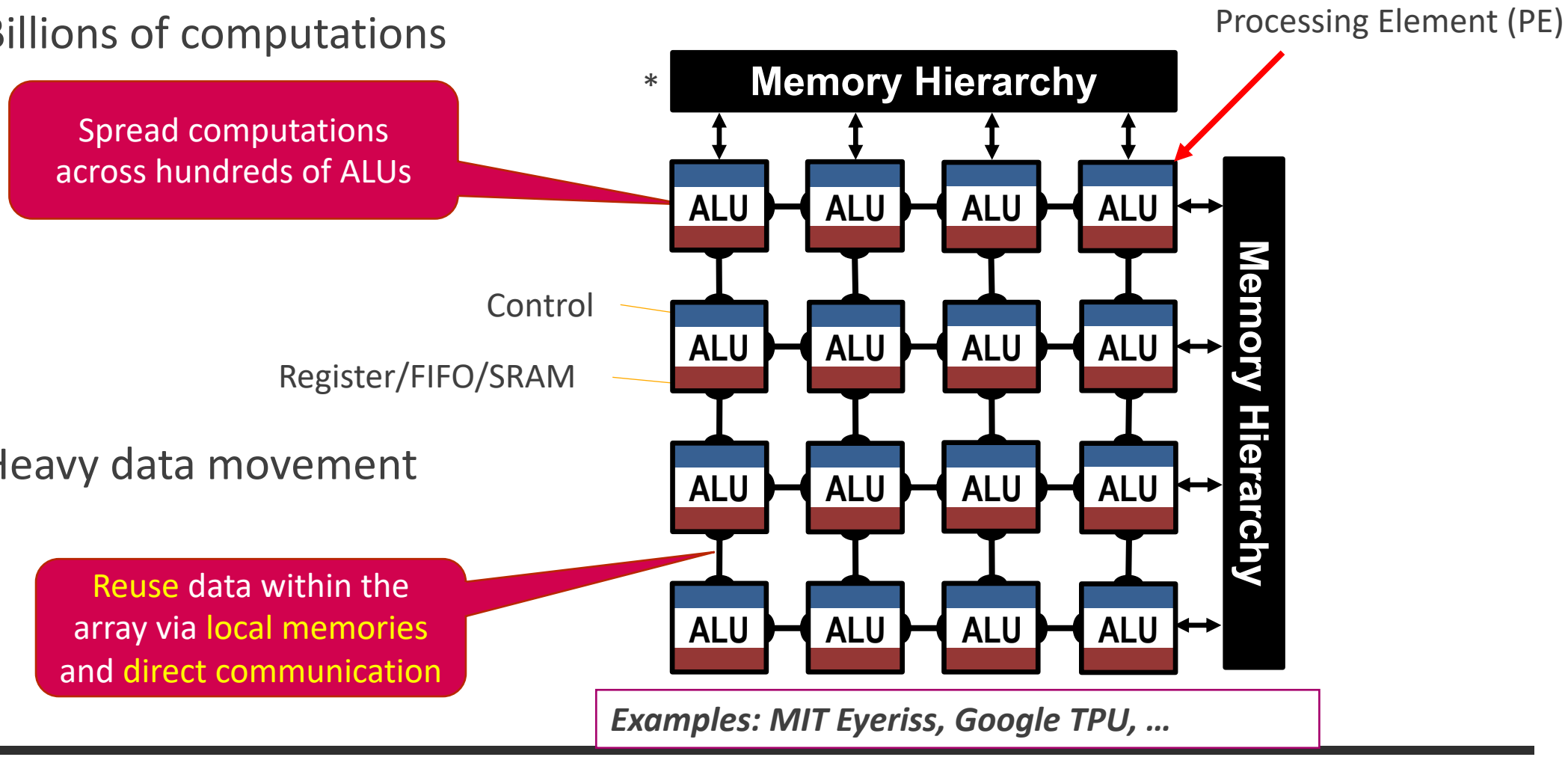


This makes GPUs inefficient

*Y. Chen et. al., "Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks," ISCA, 2016.

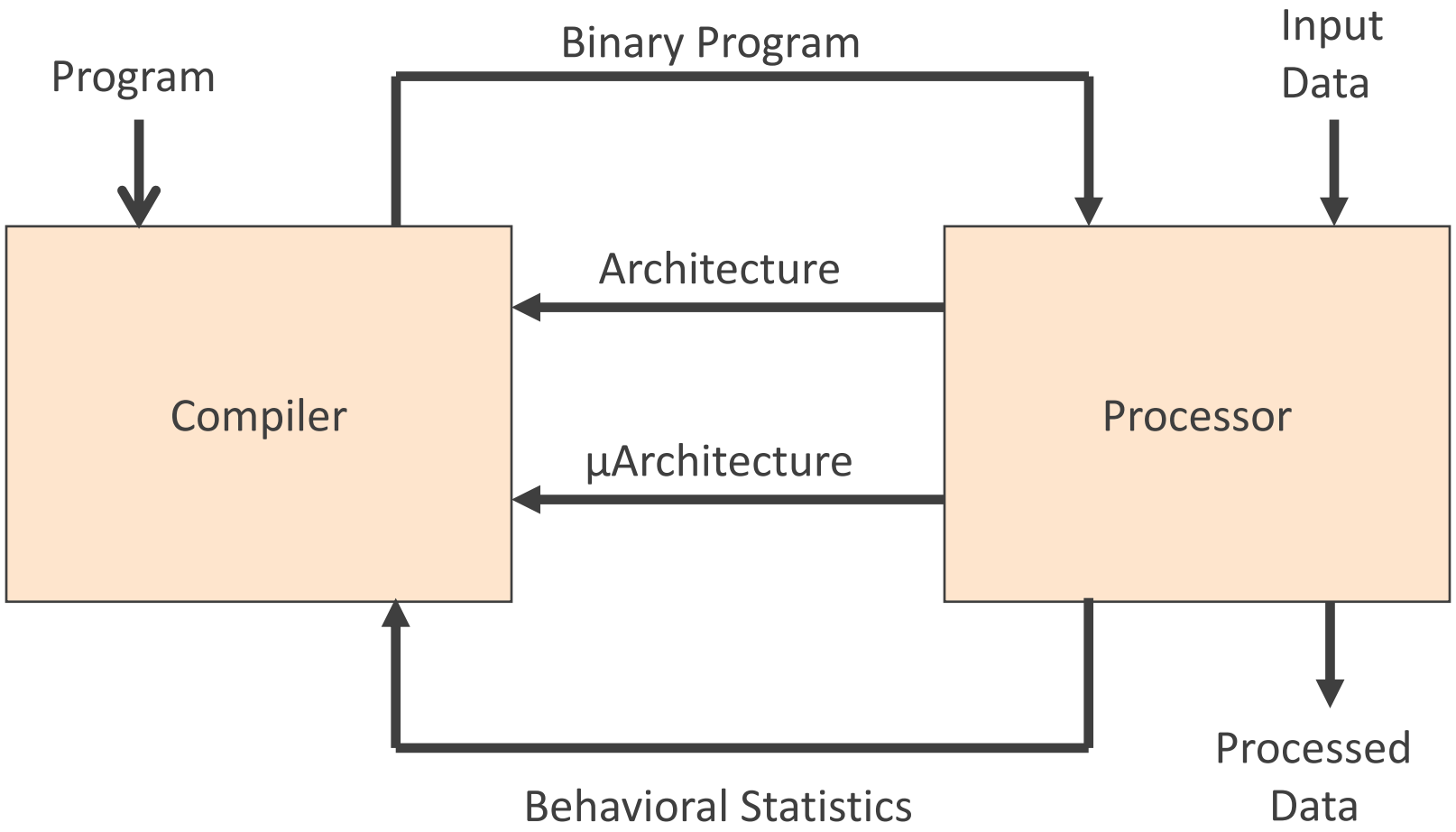
Spatial (or Dataflow) Accelerators

- **Millions of Parameters (i.e., weights)**
  - Billions of computations
  - Heavy data movement



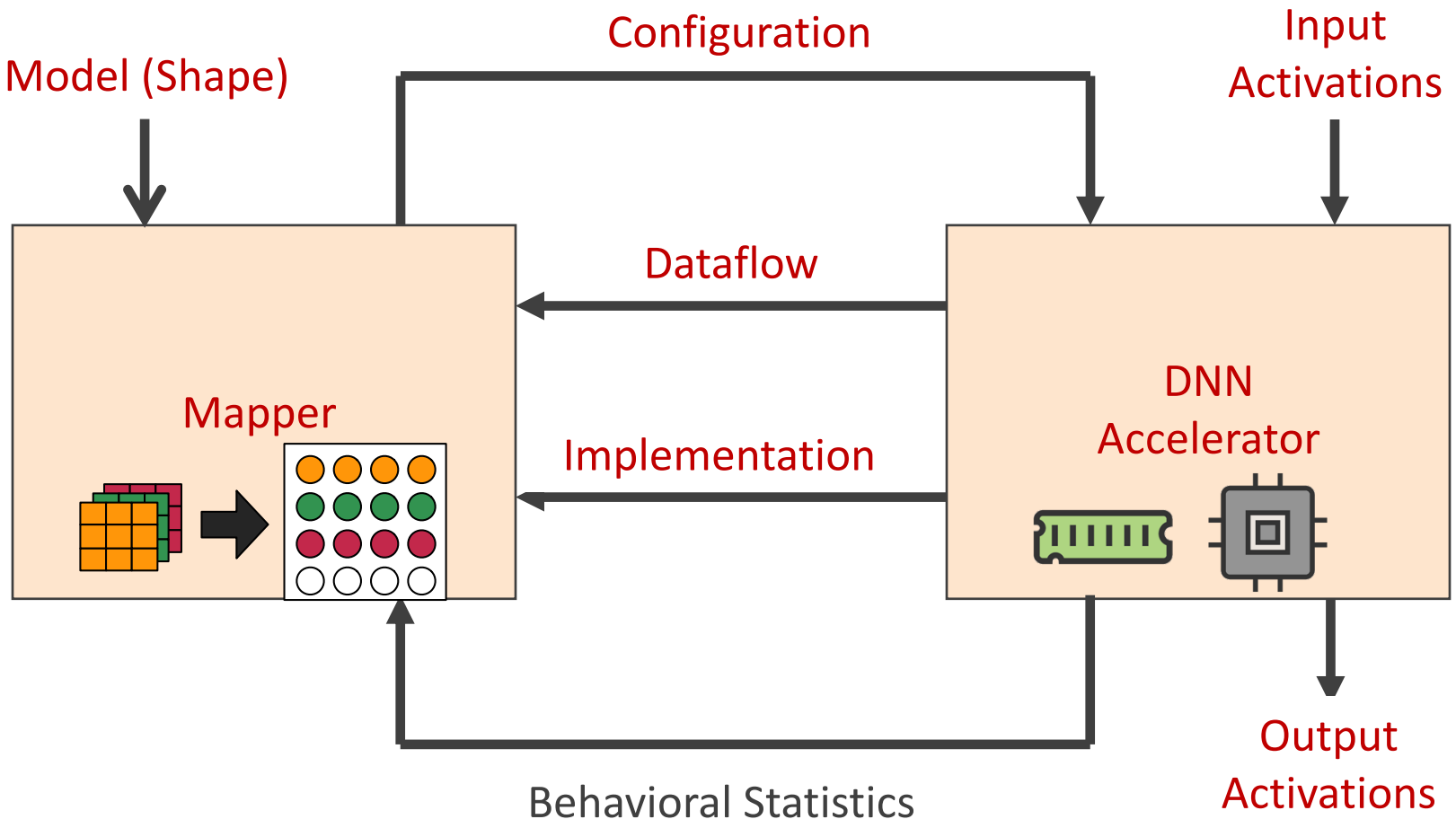
*Y. Chen et. al., "Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks," ISCA, 2016.

CPU Compute Model



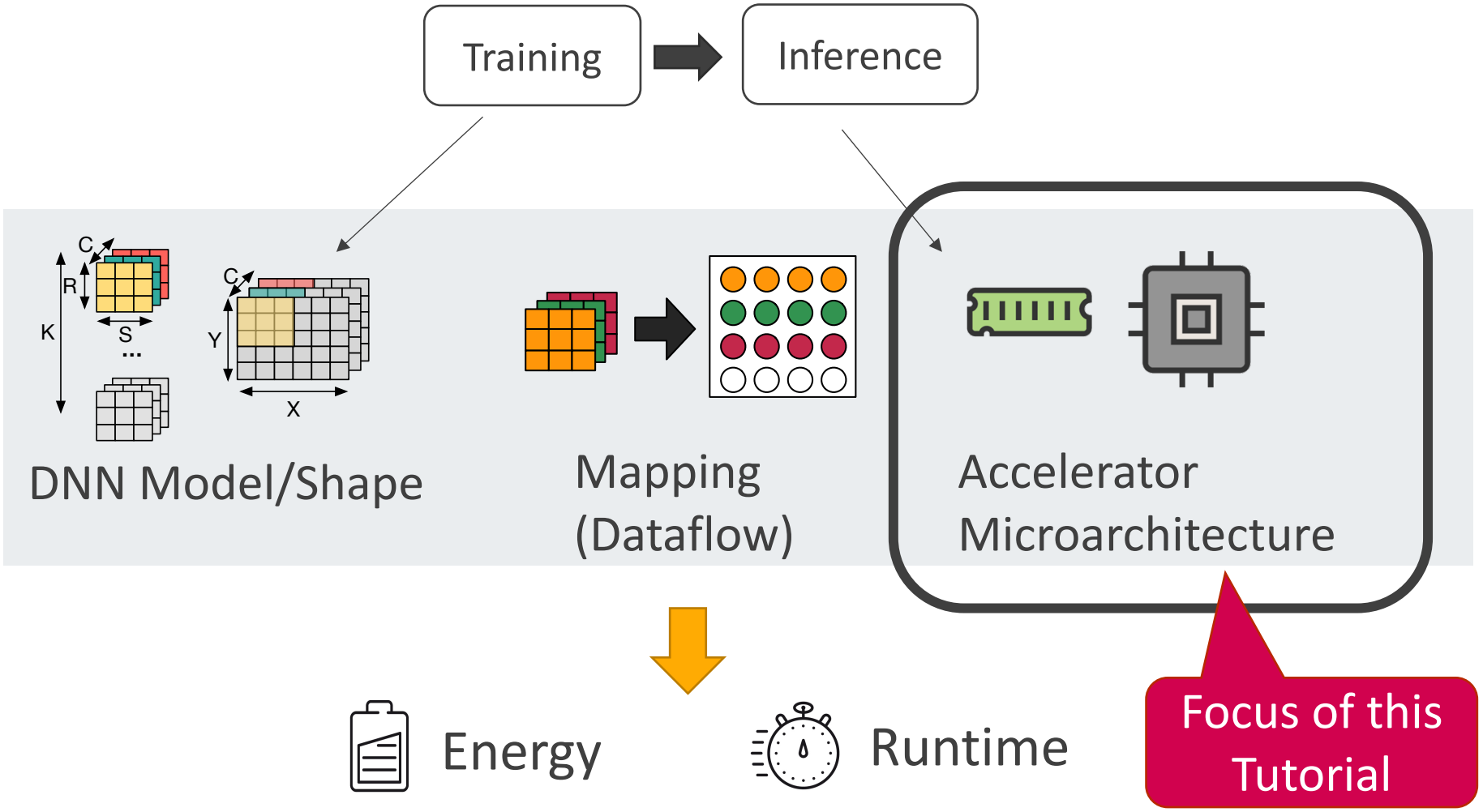
Slide Courtesy: Joel Emer and Vivienne Sze

DNN Compute Model

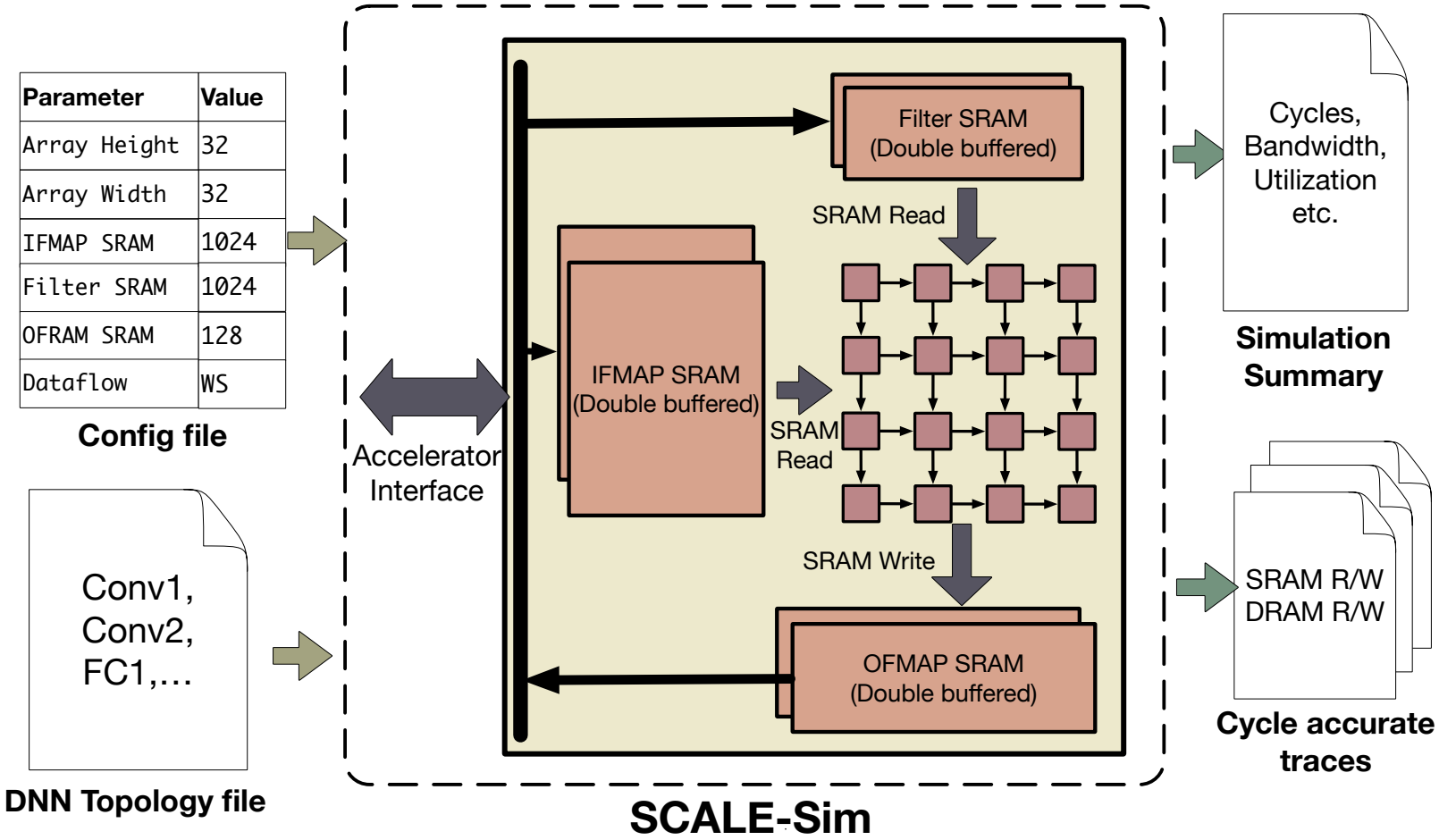


Slide Courtesy: Joel Emer and Vivienne Sze

Challenges in Design and Deployment



SCALE-sim



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

Schedule (EST)

Time slot	Topic	
10:05 – 10:10	Welcome and Overview	Tushar
10:10 to 10:45	Introduction to DNNs and Accelerator Design	Paul & Tushar
10:45 to 11:15	Overview of SCALE-Sim	Anand + Moritz + Paul
11:15 to 11:50	Tutorial 1: Design Space Exploration using SCALE-Sim	Anand
12:00 to 12:40	Tutorial 2: Modifying SCALE-Sim to add custom features	Moritz
12:45 to 1:30	Tutorial 3: Using SCALE-Sim to build larger simulators	Anand
1:30 to 2:00	Discussion on future roadmap, planned features, and ideas from the community	Yuhao

Brief Q/A at the end of each talk.

Attention: Tutorial is being recorded!

Slides + Videos will be available on the SCALE-sim tutorial website

https://scalesim-project.github.io/tutorials-2021-asplos.html