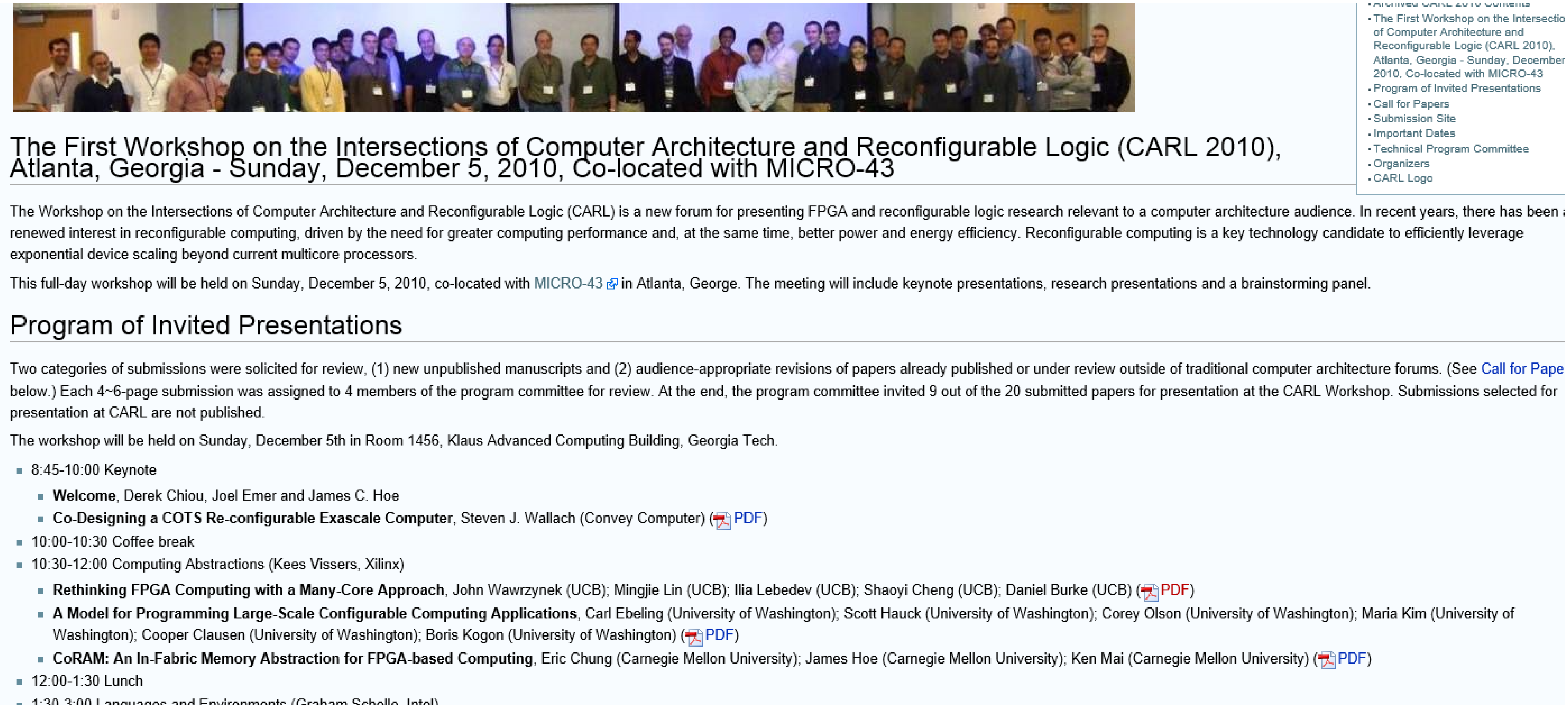
It’s Great to be Back ☺



- Archived CARL 2010 Contents
- The First Workshop on the Intersectio of Computer Architecture and Reconfigurable Logic (CARL 2010), Atlanta, Georgia - Sunday, December 2010, Co-located with MICRO-43
- Program of Invited Presentations
- Call for Papers
- Submission Site
- Important Dates
- Technical Program Committee
- Organizers
- CARL Logo

The First Workshop on the Intersections of Computer Architecture and Reconfigurable Logic (CARL 2010), Atlanta, Georgia - Sunday, December 5, 2010, Co-located with MICRO-43

The Workshop on the Intersections of Computer Architecture and Reconfigurable Logic (CARL) is a new forum for presenting FPGA and reconfigurable logic research relevant to a computer architecture audience. In recent years, there has been a renewed interest in reconfigurable computing, driven by the need for greater computing performance and, at the same time, better power and energy efficiency. Reconfigurable computing is a key technology candidate to efficiently leverage exponential device scaling beyond current multicore processors.

This full-day workshop will be held on Sunday, December 5, 2010, co-located with MICRO-43 in Atlanta, George. The meeting will include keynote presentations, research presentations and a brainstorming panel.

Program of Invited Presentations

Two categories of submissions were solicited for review, (1) new unpublished manuscripts and (2) audience-appropriate revisions of papers already published or under review outside of traditional computer architecture forums. (See Call for Pape below.) Each 4~6-page submission was assigned to 4 members of the program committee for review. At the end, the program committee invited 9 out of the 20 submitted papers for presentation at the CARL Workshop. Submissions selected for presentation at CARL are not published.

The workshop will be held on Sunday, December 5th in Room 1456, Klaus Advanced Computing Building, Georgia Tech.

- 8:45-10:00 Keynote
  - **Welcome**, Derek Chiou, Joel Emer and James C. Hoe
  - **Co-Designing a COTS Re-configurable Exascale Computer**, Steven J. Wallach (Convey Computer) (PDF)
- 10:00-10:30 Coffee break
- 10:30-12:00 Computing Abstractions (Kees Vissers, Xilinx)
  - **Rethinking FPGA Computing with a Many-Core Approach**, John Wawrzynek (UCB); Mingjie Lin (UCB); Ilia Lebedev (UCB); Shaoyi Cheng (UCB); Daniel Burke (UCB) (PDF)
  - **A Model for Programming Large-Scale Configurable Computing Applications**, Carl Ebeling (University of Washington); Scott Hauck (University of Washington); Corey Olson (University of Washington); Maria Kim (University of Washington); Cooper Clausen (University of Washington); Boris Kogon (University of Washington) (PDF)
  - **CoRAM: An In-Fabric Memory Abstraction for FPGA-based Computing**, Eric Chung (Carnegie Mellon University); James Hoe (Carnegie Mellon University); Ken Mai (Carnegie Mellon University) (PDF)
- 12:00-1:30 Lunch
- 1:30-3:00 Languages and Environments (Graham Schelle, Intel)

Microsoft Research

Microsoft

Accelerating Deep Convolutional Neural Networks Using Specialized Hardware in the Datacenter

Kalin Ovtcharov, Olatunji Ruwase, Joo-Young Kim,
Jeremy Fowers, Karin Strauss, Eric S. Chung

Microsoft

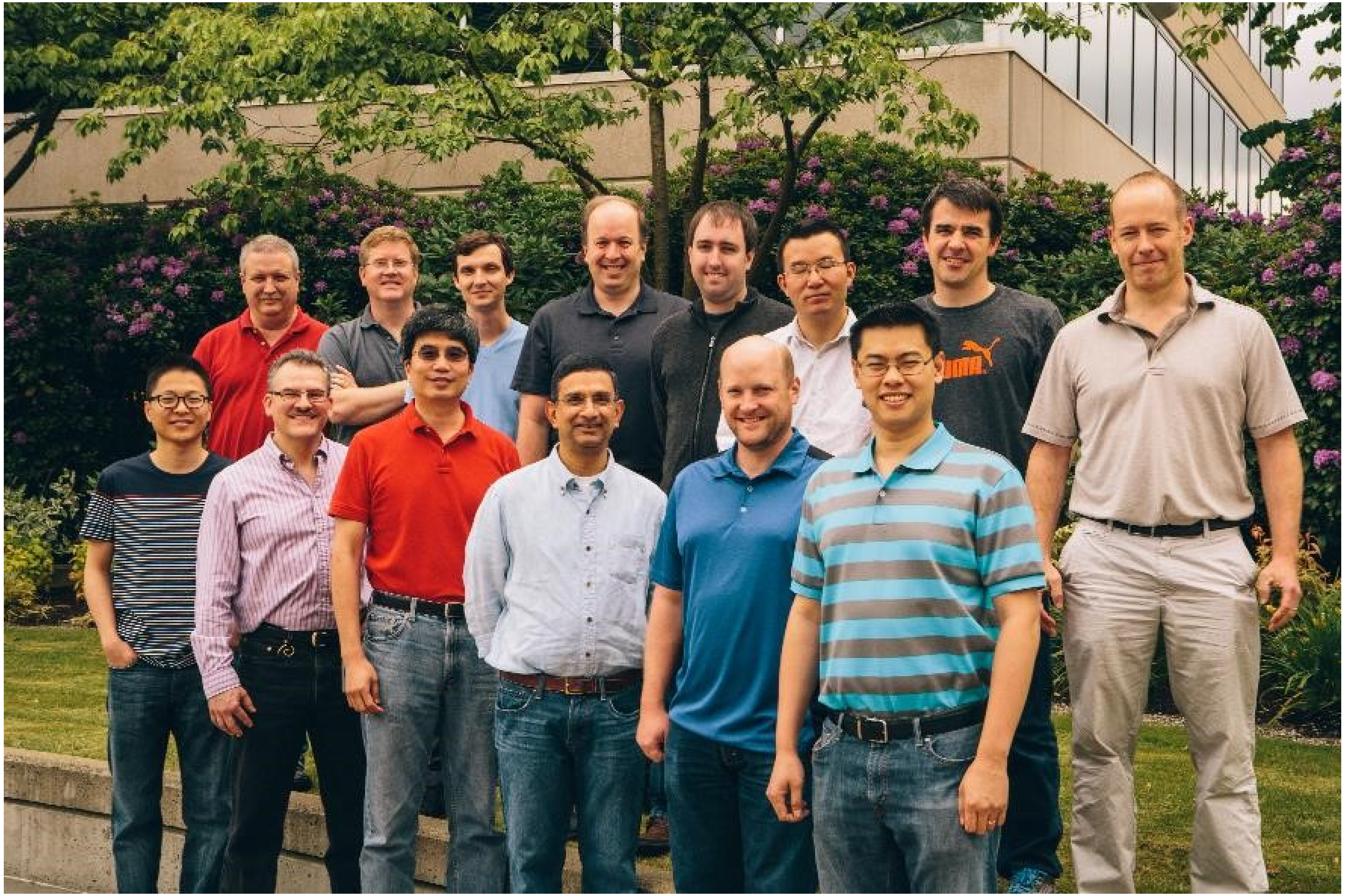
Top Row: Eric Peterson, Scott Hauck, Aaron Smith, Jan Gray, Adrian M. Caulfield, Phillip Yi Xiao, Michael Haselman, Doug Burger

Bottom Row: Joo-Young Kim, Stephen Heil, Derek Chiou, Sitaram Lanka, Andrew Putnam, Eric S. Chung

Not Pictured: Kypros Constantinides, John Demme, Hadi Esmaeilzadeh, Jeremy Fowers, Gopi Prashanth Gopal, Amir Hormati, James Larus, Simon Pope, Jason Thong

Huge thanks to our partners at

ALTERA

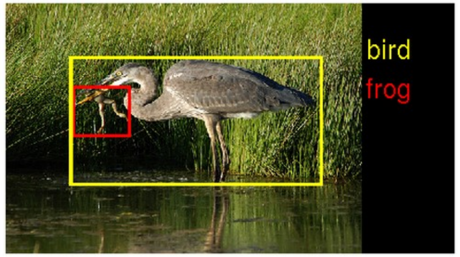


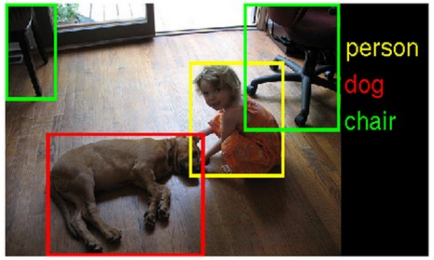
Agenda

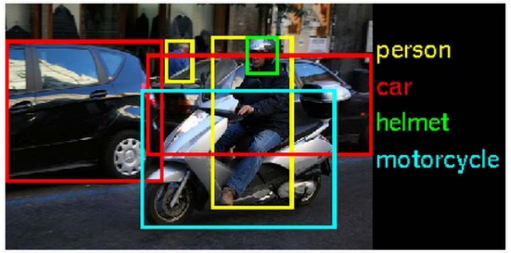
- Deep Learning on Catapult
- Academic Outreach Program

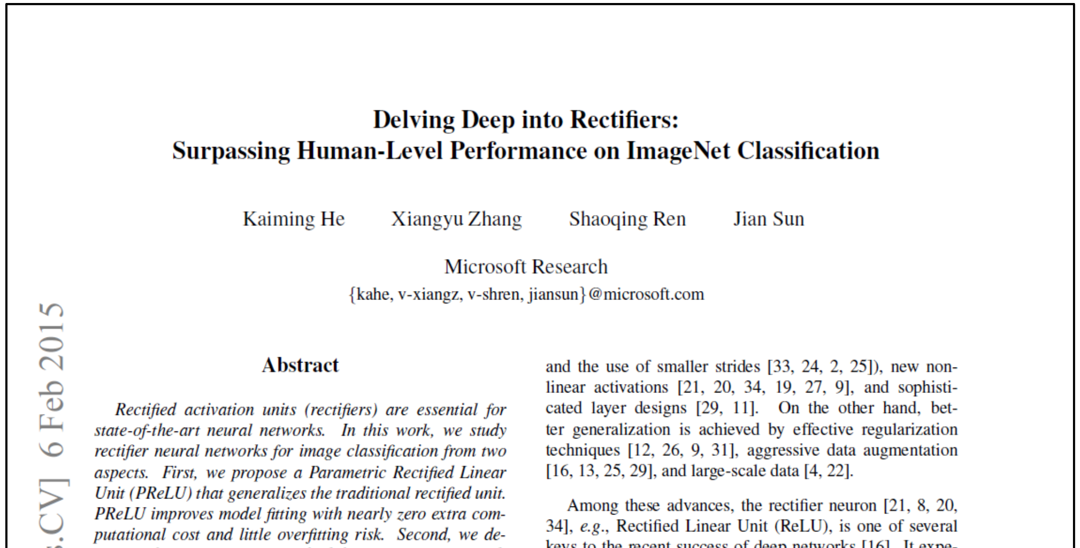
Deep Learning: The "Next Big Thing"?

- Significant advances in
  - Computer vision
  - Speech recognition
  - Natural language processing
  - Intelligent agents
  - Etc.
- State-of-the-art neural nets
  - Convolutional Neural Networks (CNNs)
  - Deep Neural Networks (DNNs)
  - … ?









s.CV] 6 Feb 2015

**Delving Deep into Rectifiers:
Surpassing Human-Level Performance on ImageNet Classification**

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun

Microsoft Research
{kahe, v-xiangz, v-shren, jiansun}@microsoft.com

Abstract

Rectified activation units (rectifiers) are essential for state-of-the-art neural networks. In this work, we study rectifier neural networks for image classification from two aspects. First, we propose a Parametric Rectified Linear Unit (PReLU) that generalizes the traditional rectified unit. PReLU improves model fitting with nearly zero extra computational cost and little overfitting risk. Second, we de-

and the use of smaller strides [33, 24, 2, 25]), new nonlinear activations [21, 20, 34, 19, 27, 9], and sophisticated layer designs [29, 11]. On the other hand, better generalization is achieved by effective regularization techniques [12, 26, 9, 31], aggressive data augmentation [16, 13, 25, 29], and large-scale data [4, 22].

Among these advances, the rectifier neuron [21, 8, 20, 34], *e.g.*, Rectified Linear Unit (ReLU), is one of several keys to the recent success of deep networks [16]. It expe-

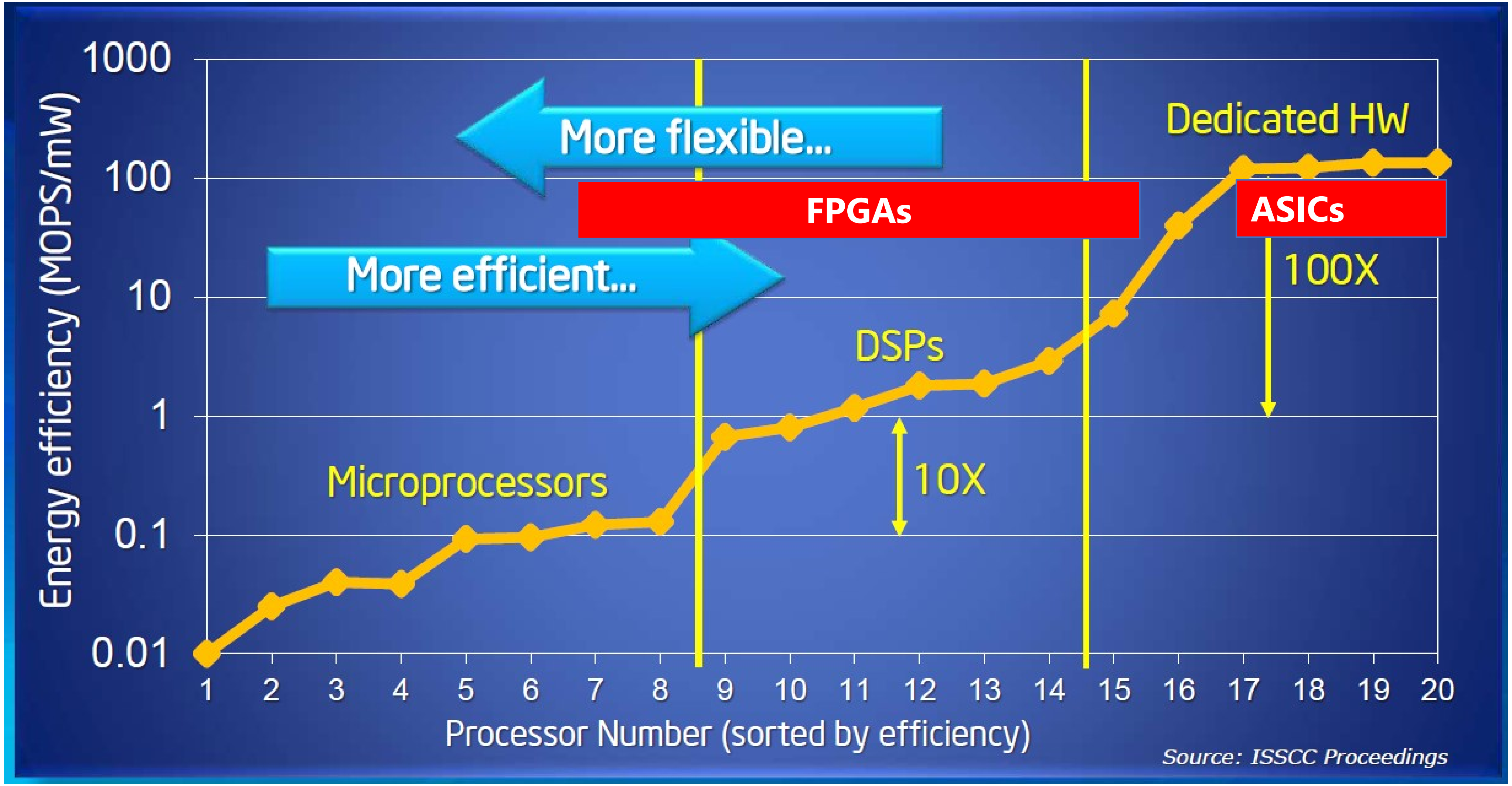
Goal: Deep Learning as a Cloud Service

- ML in the cloud
  - Leverage economies of scale in shared cloud infrastructure
  - Support training of new models
  - Deploy pre-trained models (e.g., classify images in OneDrive)
  - Scale training and deployment up to hundreds of thousands of machines
- Expose through cloud providers
  - Microsoft AzureML
  - Amazon ML-as-a-Service
  - Google Prediction API

Challenges

- Training very slow on conventional CPUs
  - Up to months
  - Yet, most cloud services built on commodity CPUs and components
- Deploying trained models also compute-intensive
- GPUs preferred by many practitioners but
  - Difficult to scale beyond 16-32 nodes
  - Limited in memory capacity (affecting model size and accuracy)
  - Too power-intensive for datacenters
  - Expensive to maintain
  - Have reliability issues

The Efficiency of Specialized Hardware



Source: Bob Broderson, Berkeley Wireless group

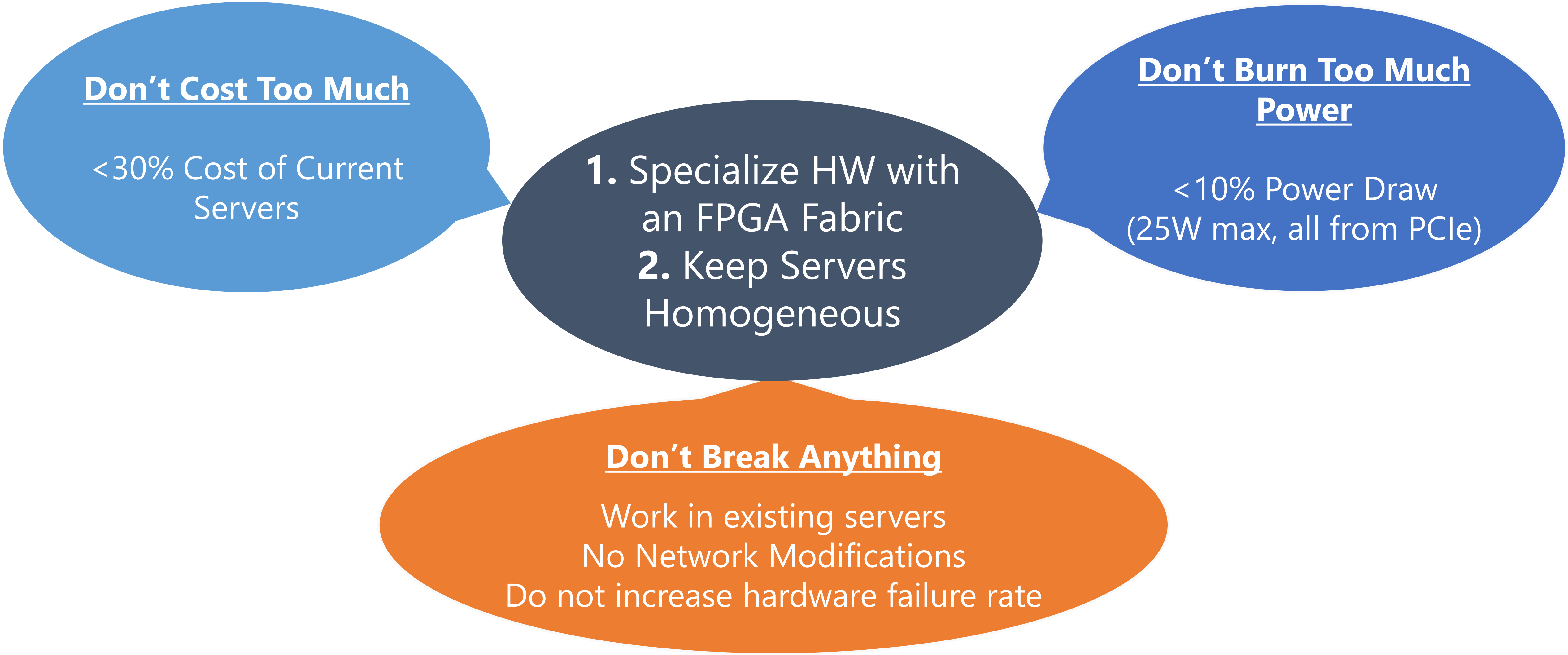
Datacenter Environment

- Software services change monthly
- Machines last 3 years, purchased on a rolling basis
- Machines repurposed ~½ way into lifecycle
- Little/no HW maintenance, no accessibility

- Homogeneity is highly desirable

The paradox: Specialization *and* homogeneity

Our Design Requirements



1. Specialize HW with an FPGA Fabric
2. Keep Servers Homogeneous

Don’t Cost Too Much

<30% Cost of Current Servers

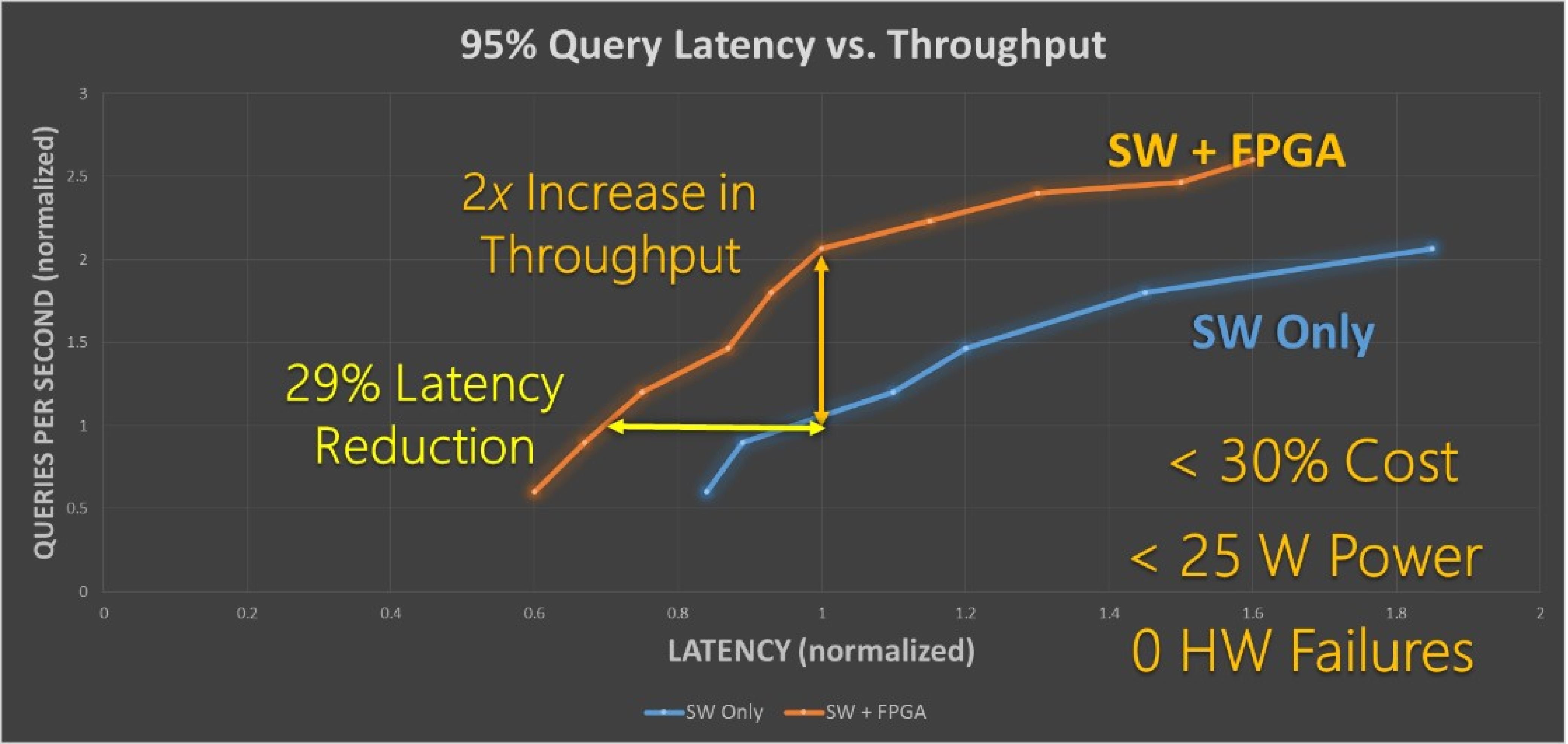
Don’t Burn Too Much Power

<10% Power Draw
(25W max, all from PCIe)

Don’t Break Anything

Work in existing servers
No Network Modifications
Do not increase hardware failure rate

MICROSOFT SUPERCHARGES BING SEARCH WITH PROGRA



95% Query Latency vs. Throughput

2x Increase in Throughput

29% Latency Reduction

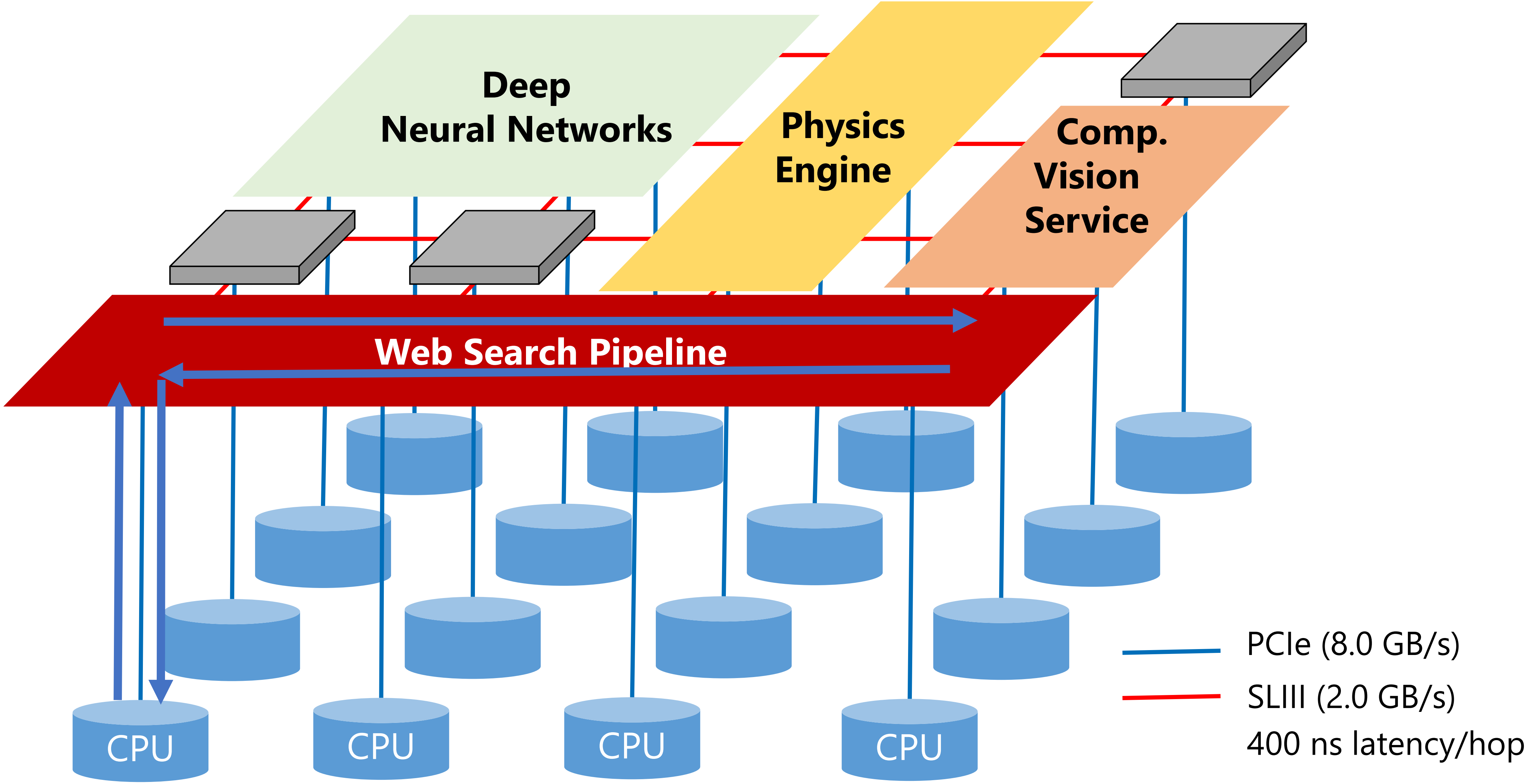
< 30% Cost

< 25 W Power

0 HW Failures

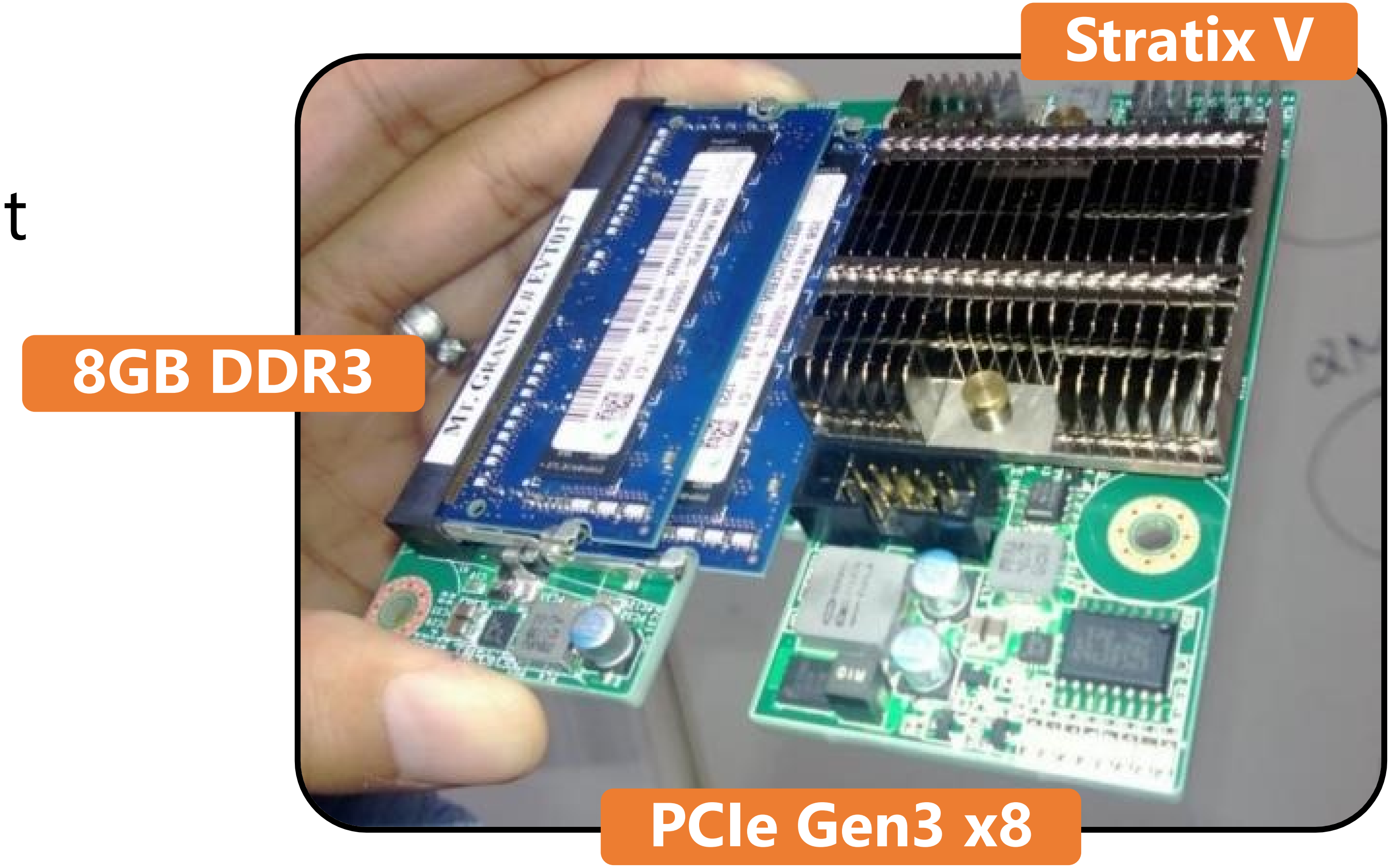
http://www.wired.com/2014/06/microsoft-fpga/

Catapult: An Elastic Reconfigurable Fabric for Datacenters

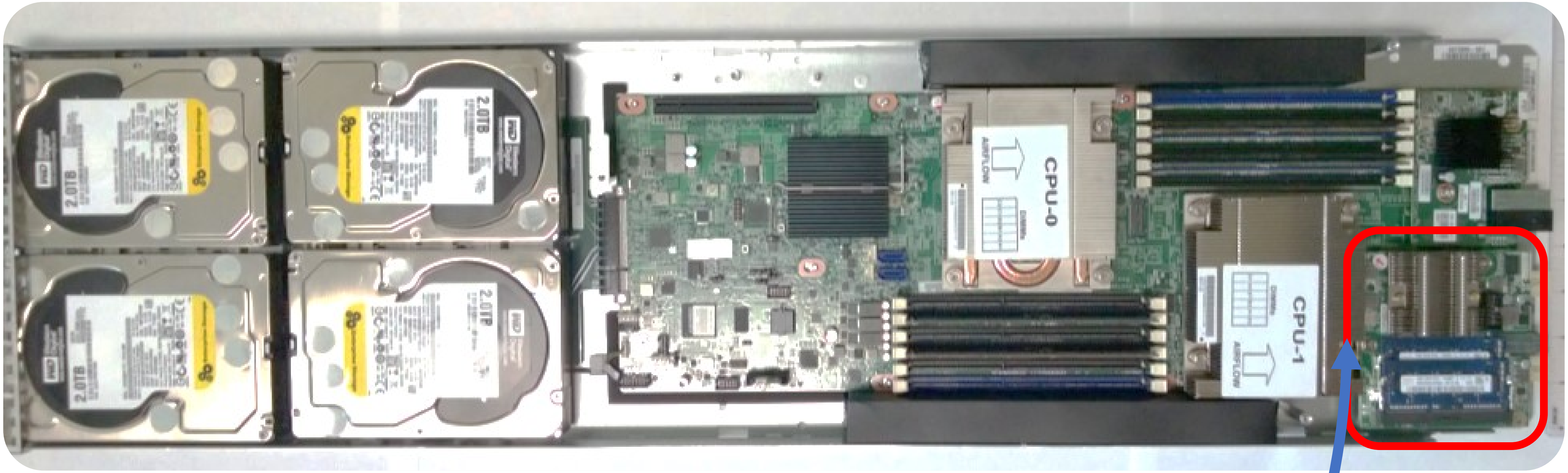


Catapult FPGA Accelerator Card

- Altera Stratix V D5
- 172,600 ALMs, 2,014 M20Ks, 1,590 DSPs
- PCIe Gen 3 x8
- 8GB DDR3-1333
- Powered by PCIe slot
- Torus Network



Microsoft Open Compute Server



- Two 8-core Xeon 2.1 GHz CPUs
- 64 GB DRAM
- 4 HDDs @ 2 TB, 2 SSDs @ 512 GB
- 10 Gb Ethernet
- No cable attachments to server

Air flow

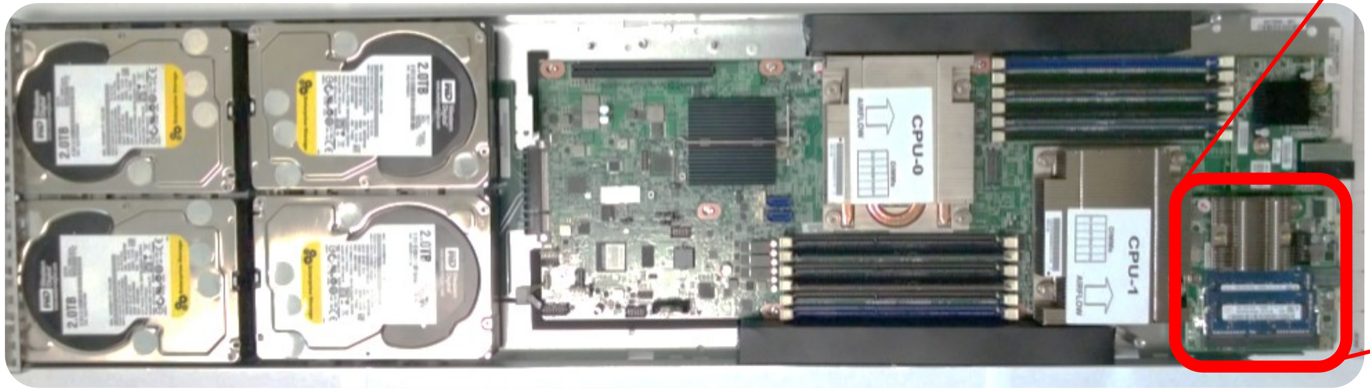
200 LFM

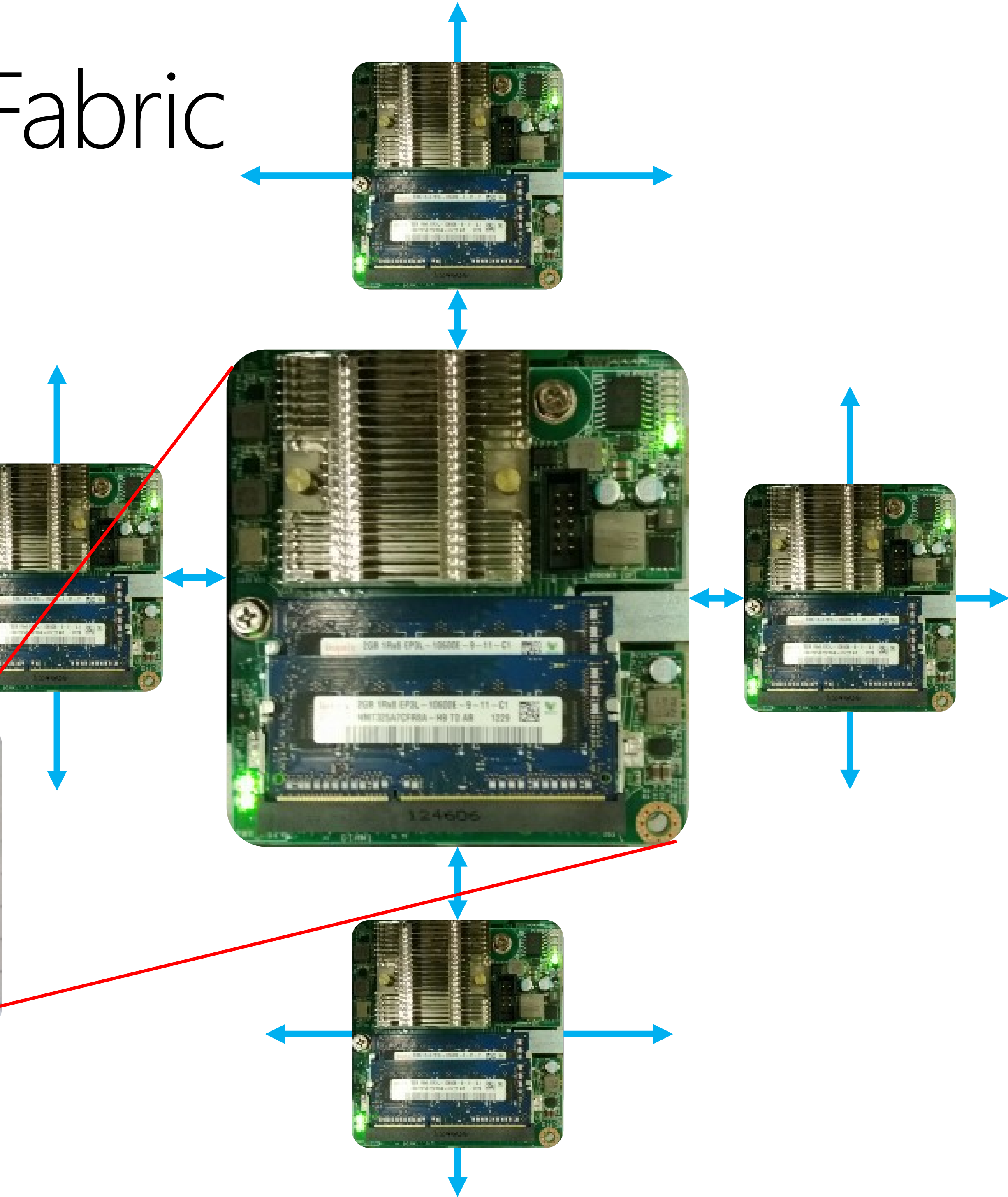
68 ^{0}C Inlet

Scalable Reconfigurable Fabric

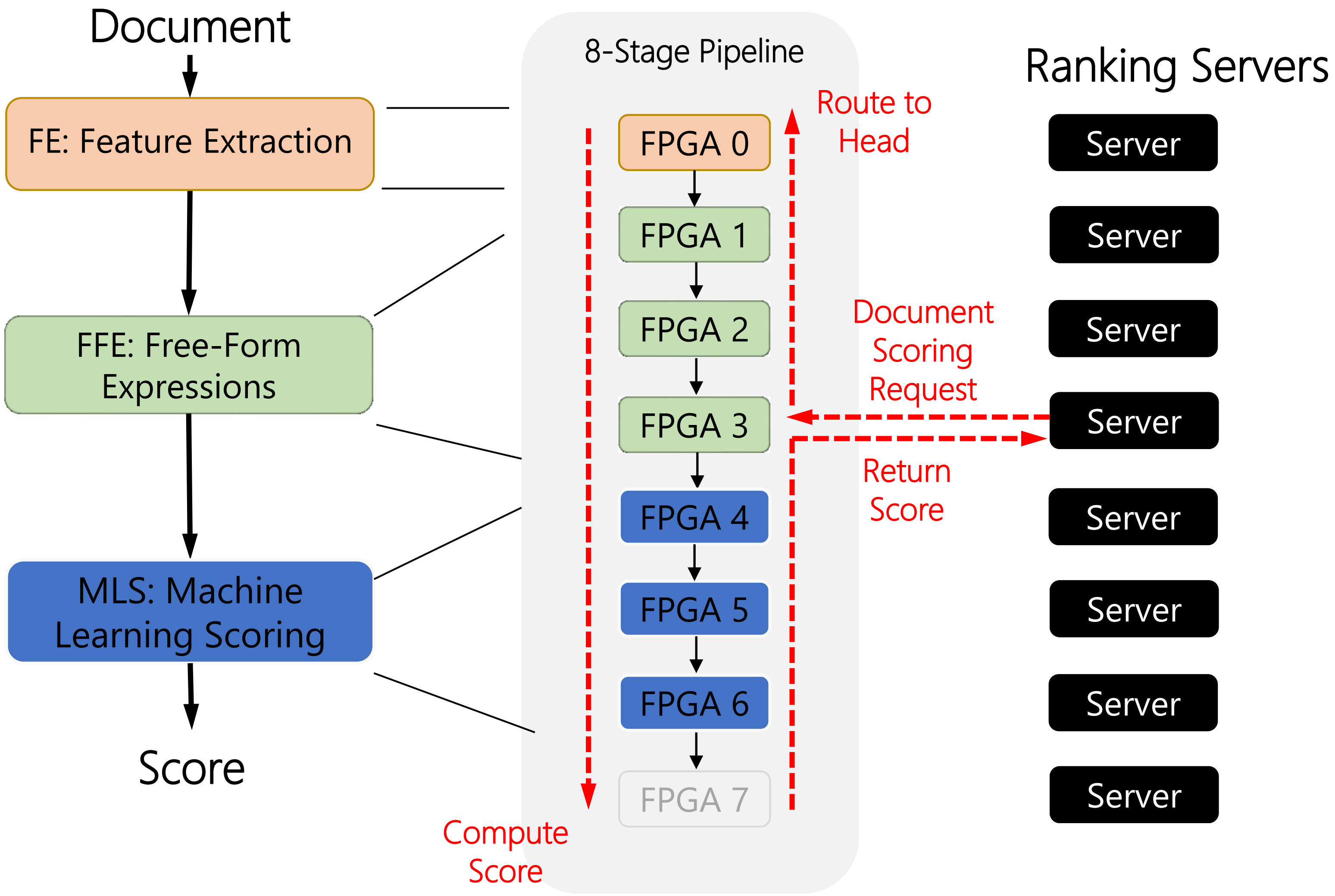
- 1 FPGA board per Server
- 48 Servers per ½ Rack
- 6x8 Torus Network among FPGAs
    - 20 Gb/s over SAS SFF-8088 cables

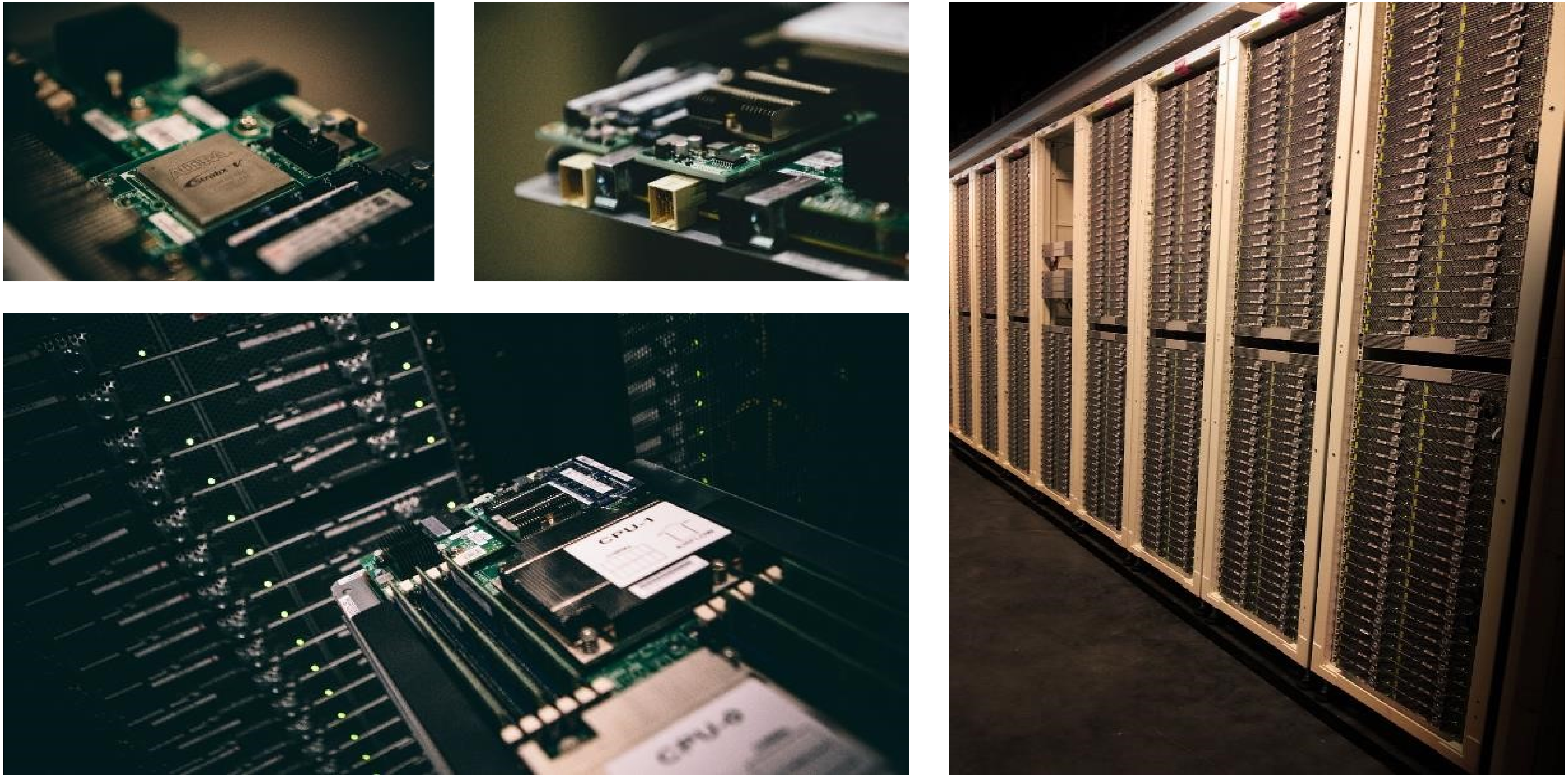
Data Center Server (1U, ½ width)





FPGA Accelerator for Bing Ranking





1,632 Server Pilot Deployed in a Production Datacenter

Scalable Deep Learning on Catapult

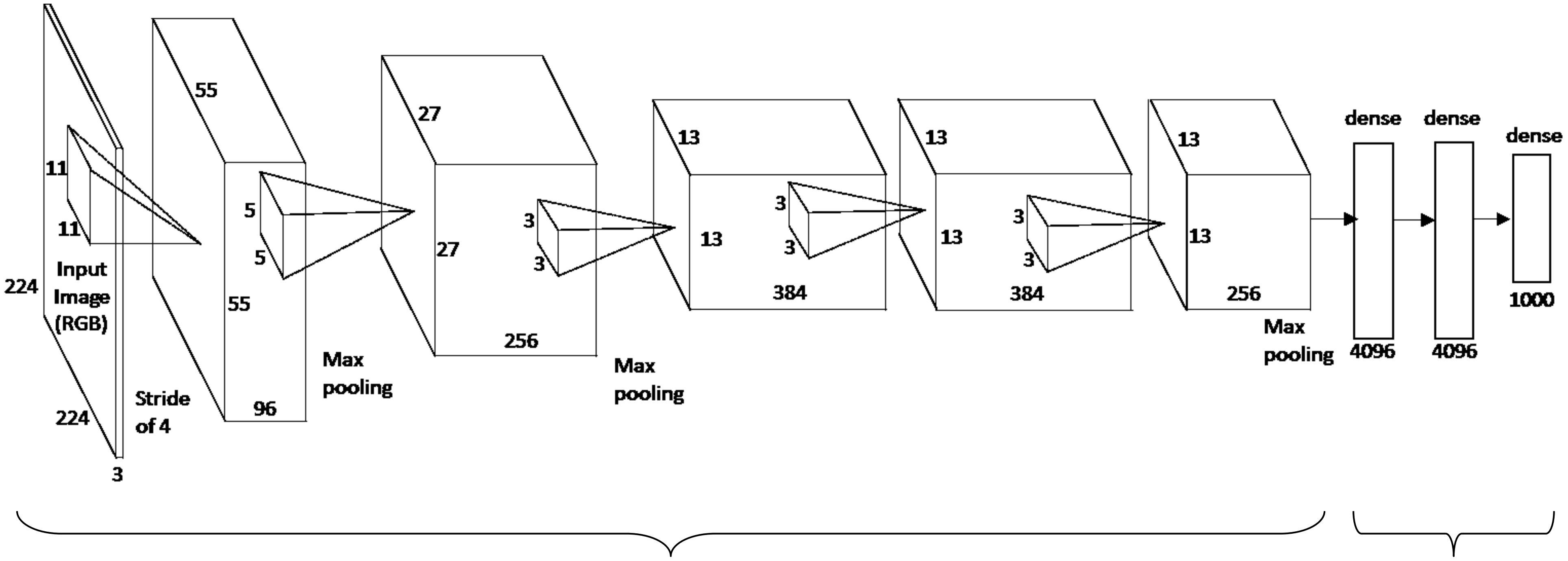
- Provide excellent performance and accuracy at fraction of cost of commodity CPUs

- Leverage abundant FPGA resources in MSFT’s datacenters for scaling up machine learning and model deployment

- Target high-valued kernels and expose to practitioners as composable SW libraries

Image Classification with Deep CNN

INPUT



3-D Convolution and Max Pooling

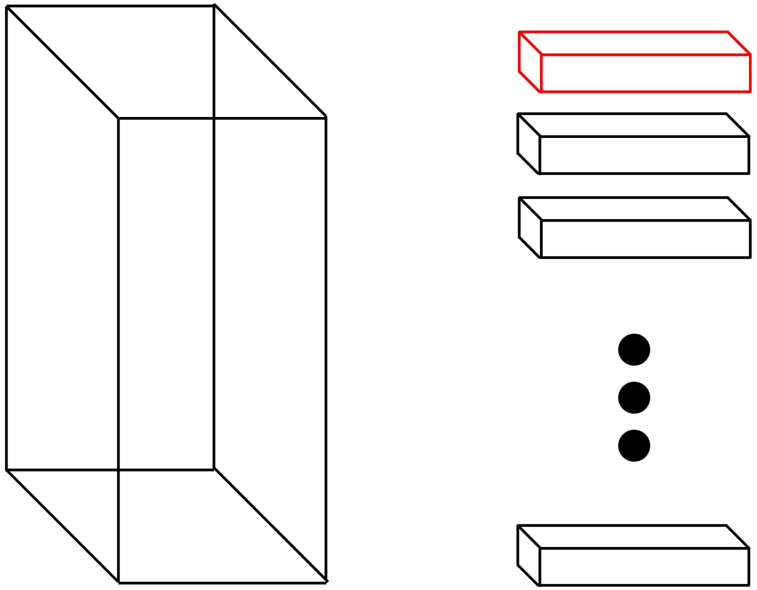
Dense Layers

OUTPUT

"Dog"

* Krizhevsky et al, NIPS'12

3-D Convolution

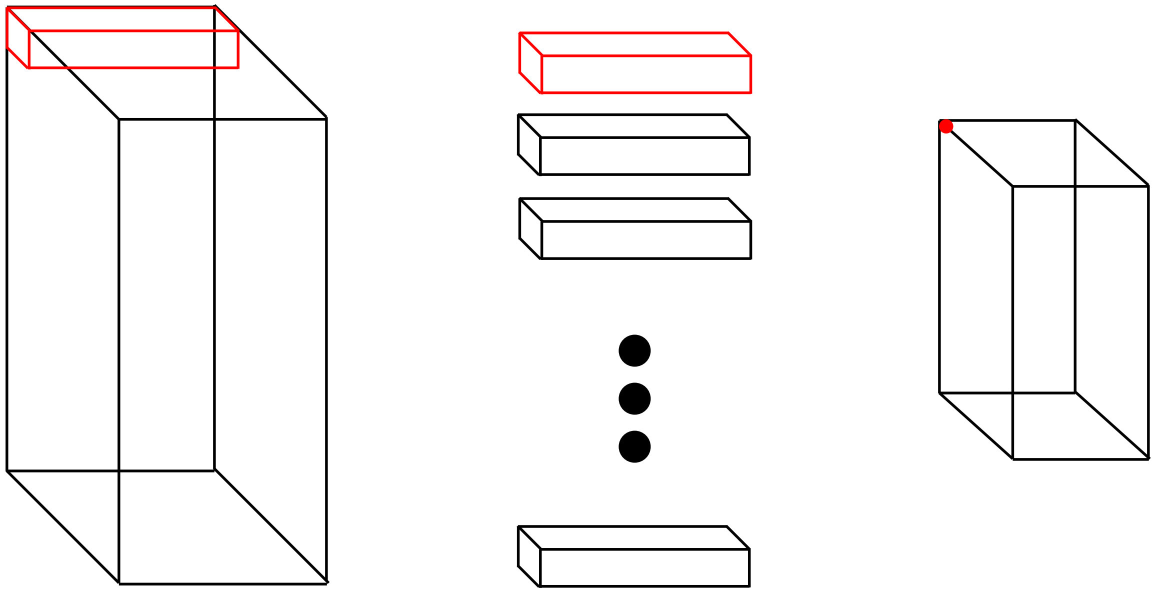


Input

Model Weights

Output

3-D Convolution

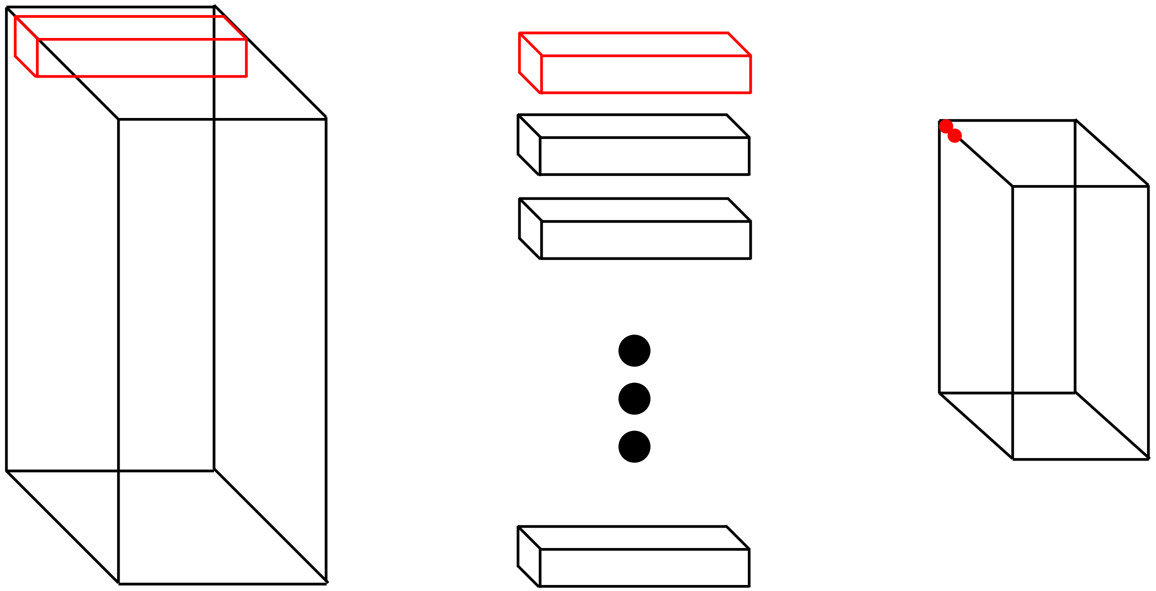


Input

Model Weights

Output

3-D Convolution

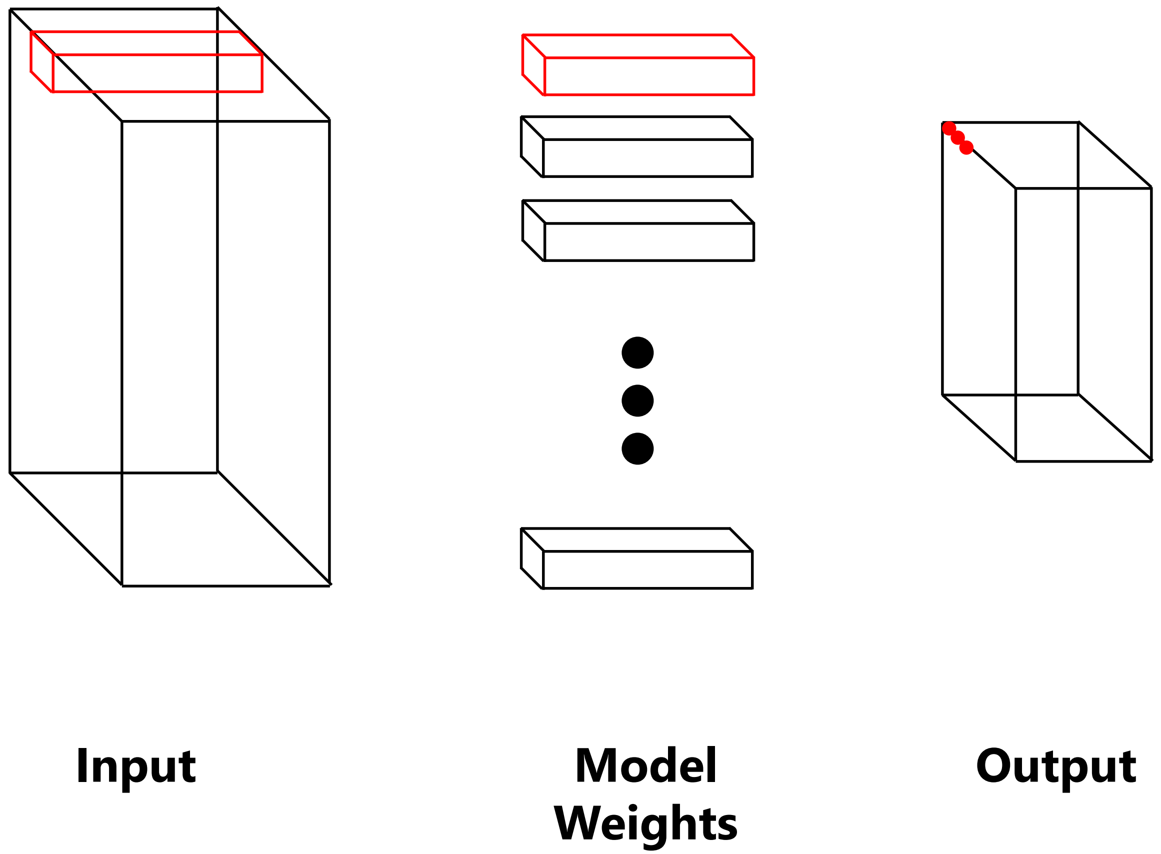


Input

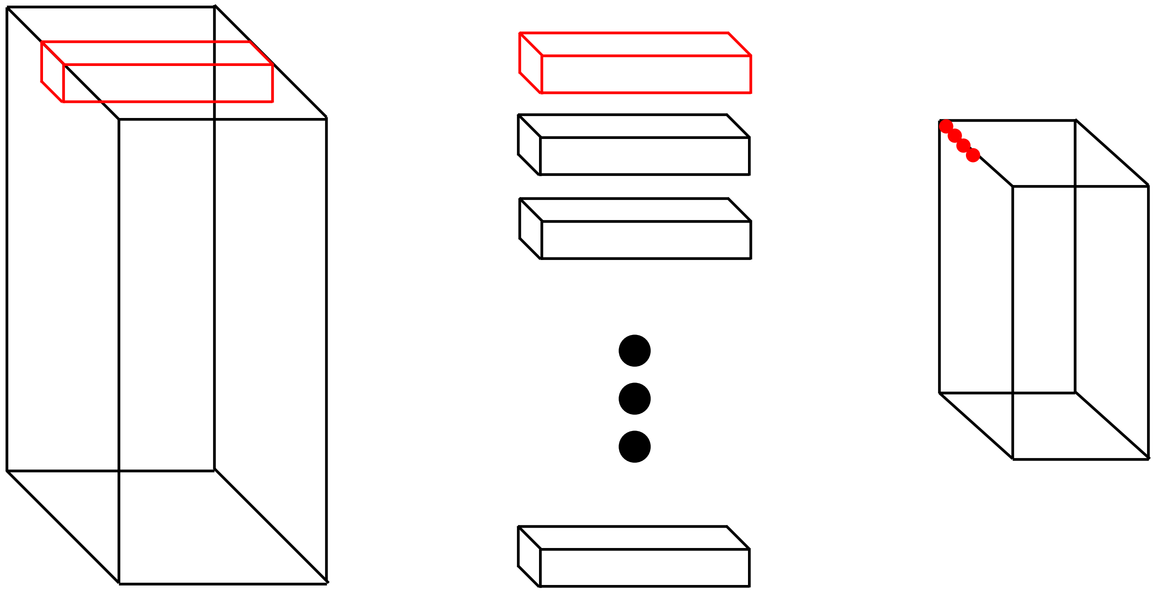
Model Weights

Output

3-D Convolution



3-D Convolution

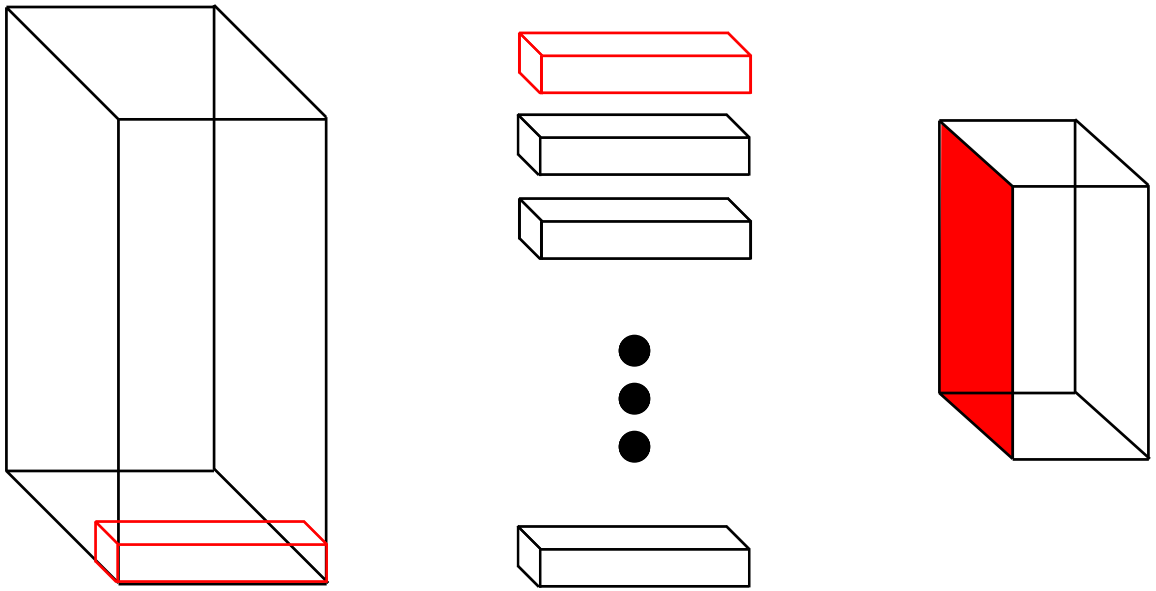


Input

Model Weights

Output

3-D Convolution

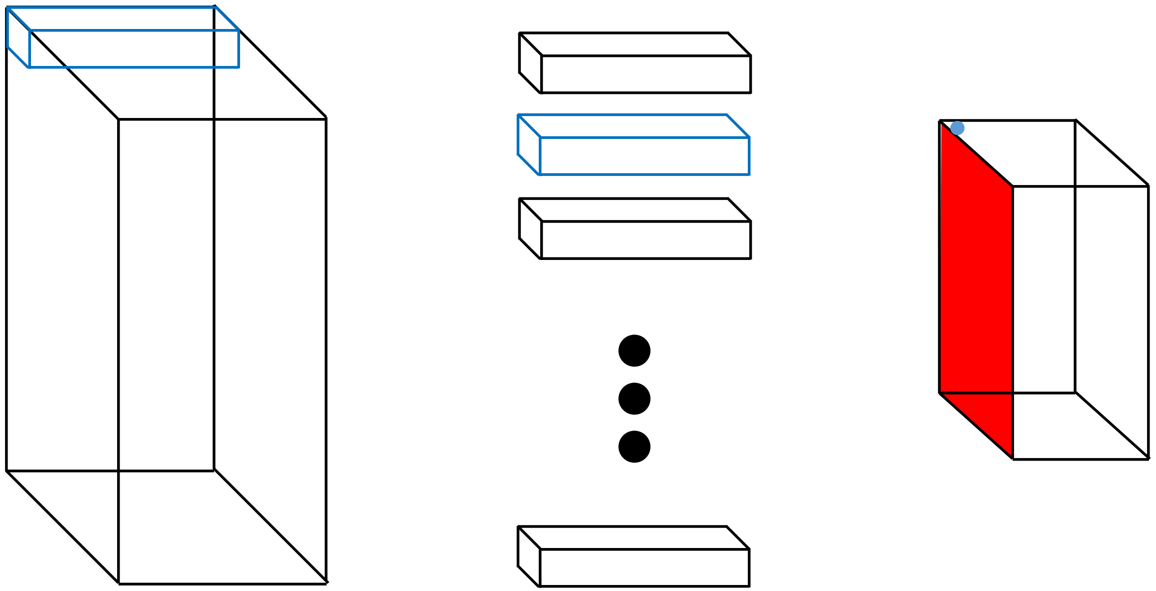


Input

Model Weights

Output

3-D Convolution

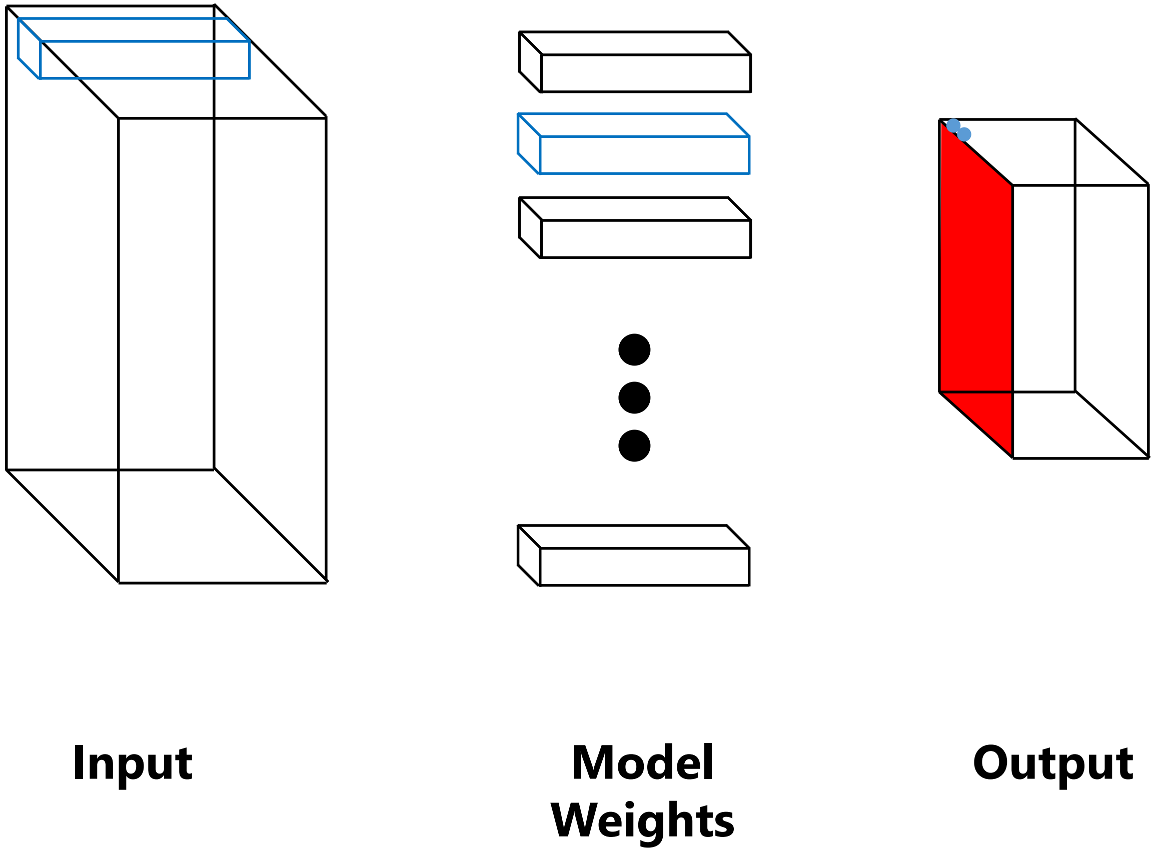


Input

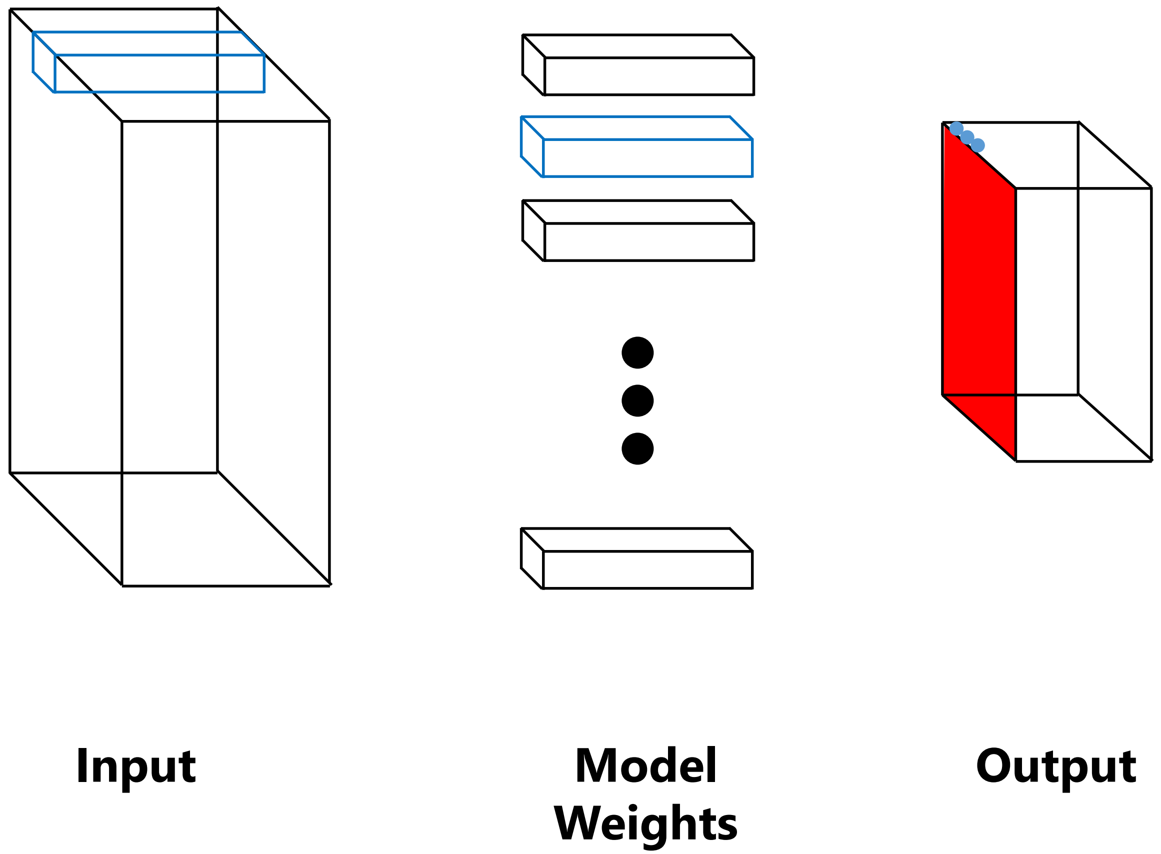
Model Weights

Output

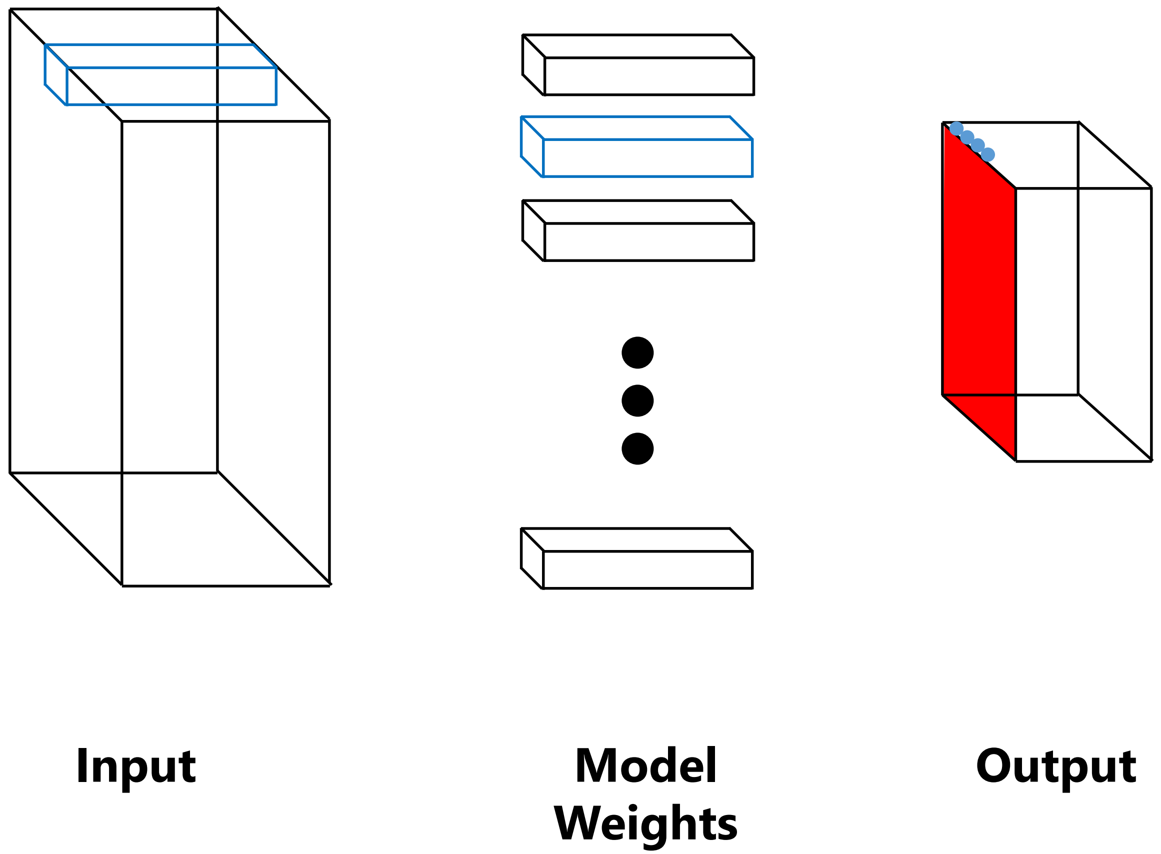
3-D Convolution



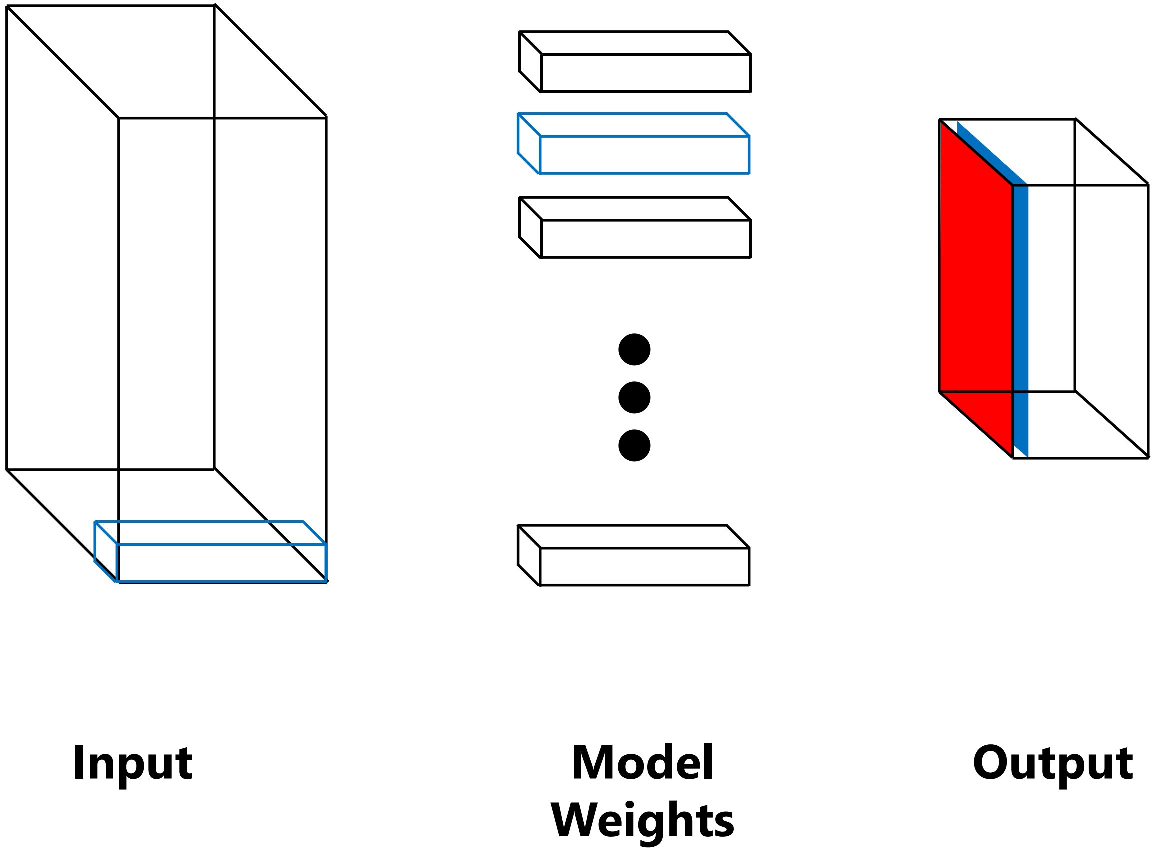
3-D Convolution



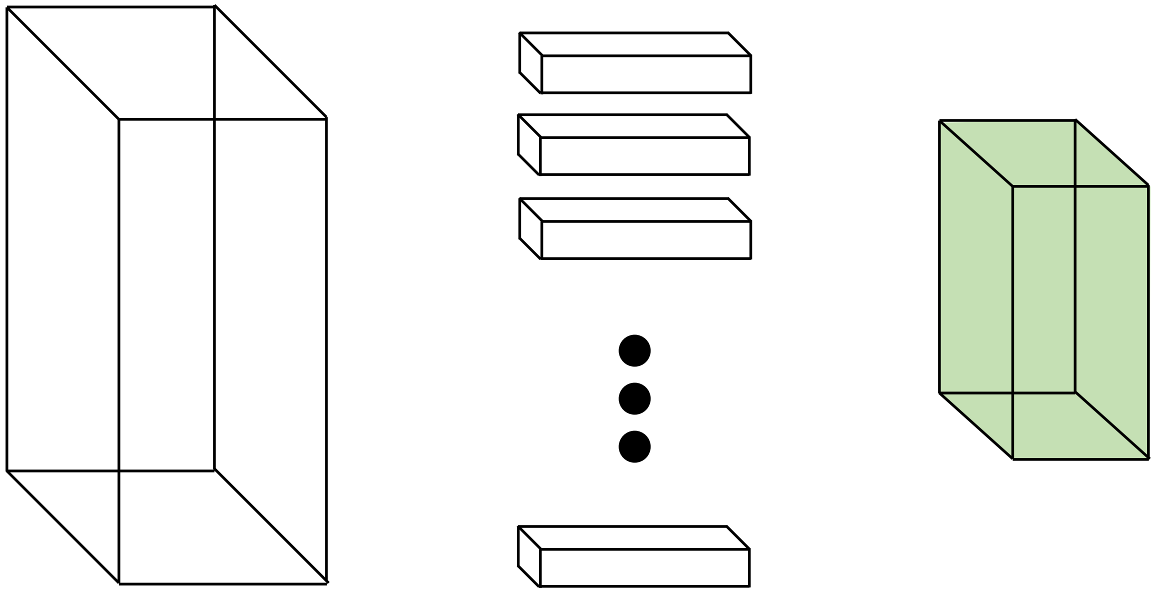
3-D Convolution



3-D Convolution



3-D Convolution

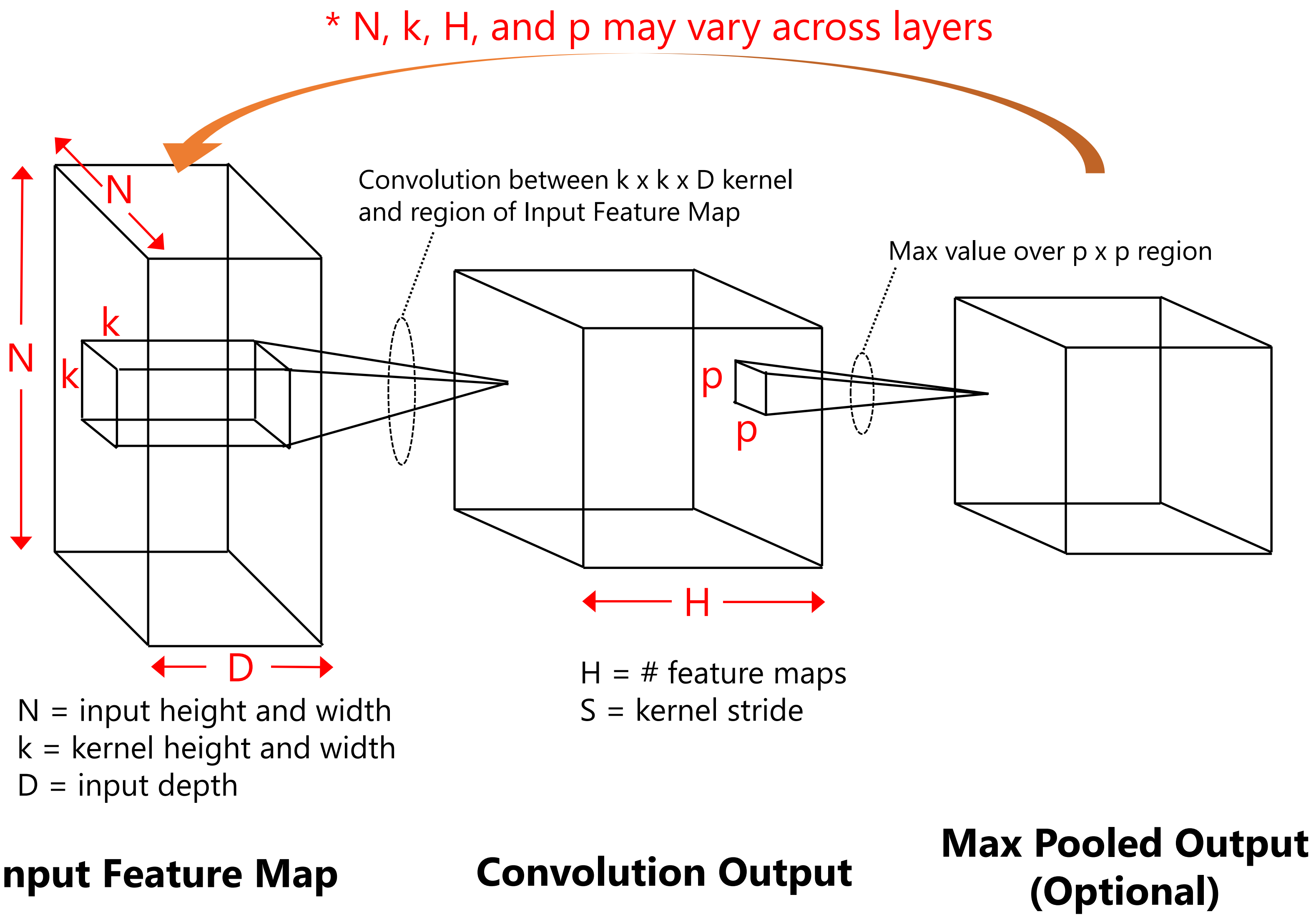


Input

Model Weights

Output

3-D Convolution and Max Pooling



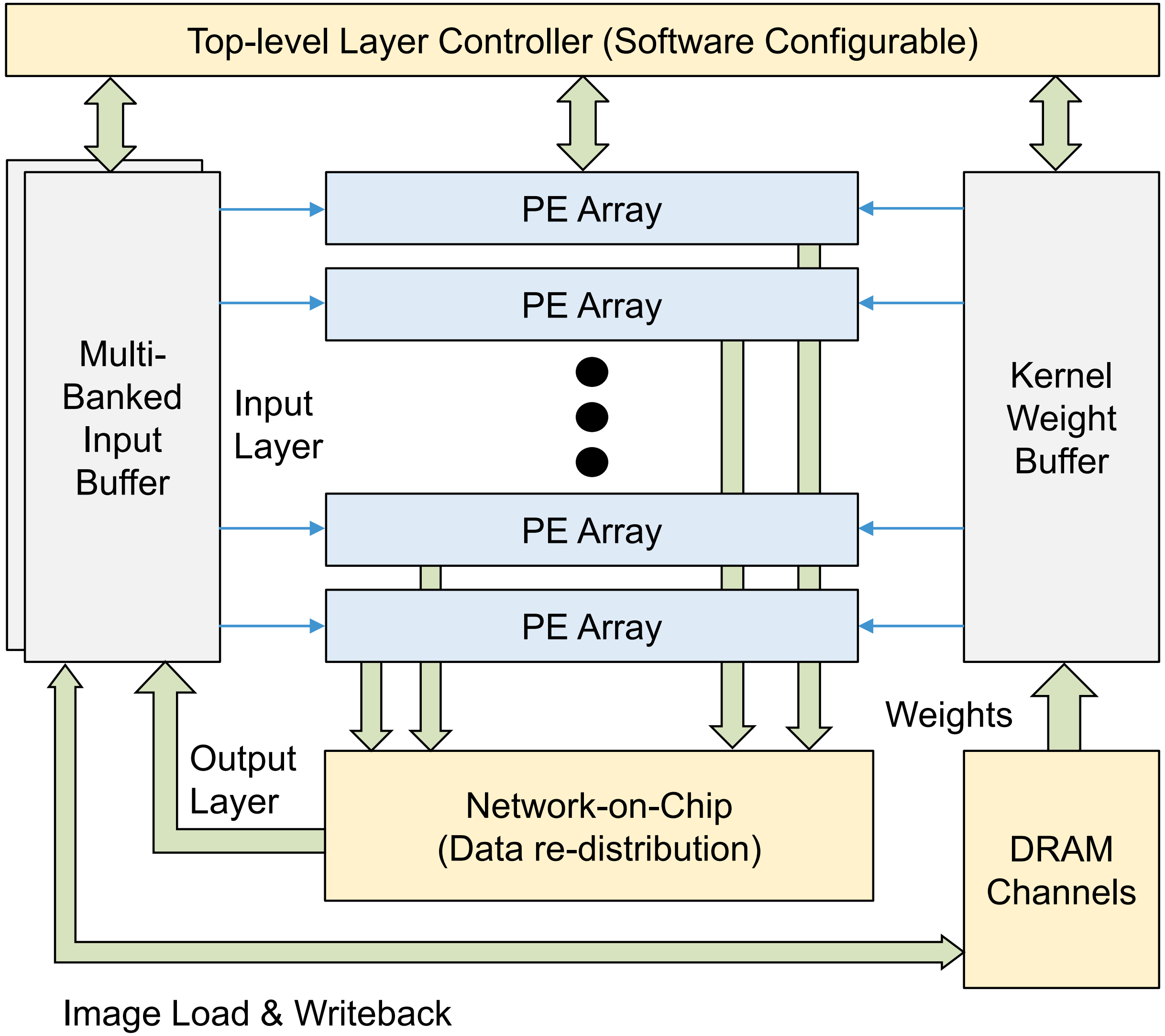
* N, k, H, and p may vary across layers
N
N
k
k
Convolution between k x k x D kernel and region of Input Feature Map
p
p
Max value over p x p region
D
H
Input Feature Map
Convolution Output
Max Pooled Output (Optional)

N = input height and width
k = kernel height and width
D = input depth

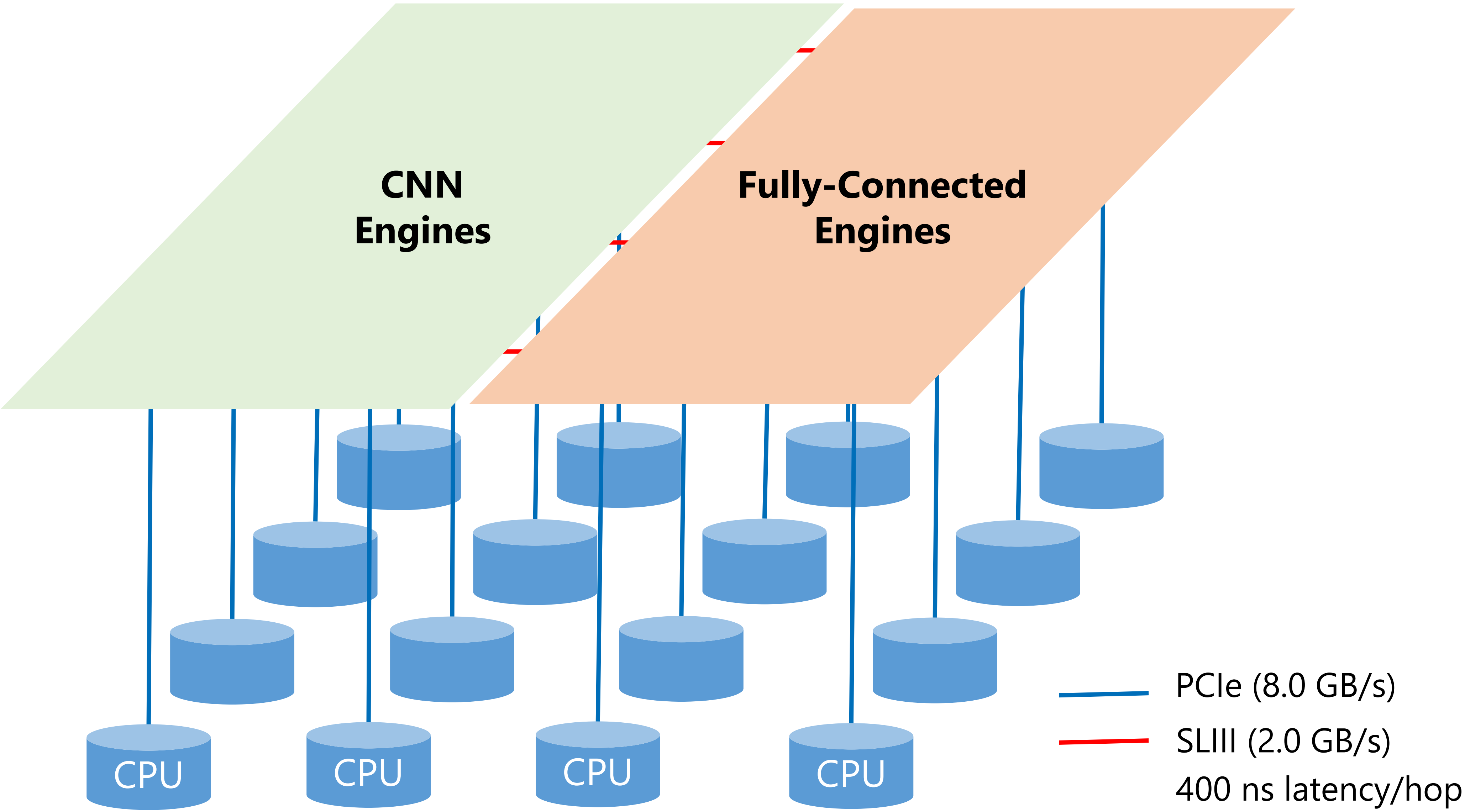
H = # feature maps
S = kernel stride

CNN Accelerator Building Block

- Configurable
  - Numerical precision (static)
  - Number of layers
  - Layer dimensions
  - Stride and pooling
- Scalable
  - Can compose multiple engines together over Catapult network
- Efficient
  - Minimize memory bandwidth via data re-distribution NoC
  - On-chip per-row broadcast



Scalable Deep Learning on Catapult



CNN Engines

Fully-Connected Engines

PCIe (8.0 GB/s)

SLIII (2.0 GB/s)
400 ns latency/hop

CNN Classification Performance

	CIFAR-10	ImageNet 1K	ImageNet 22K	FPGA or GPU Power
Server + Stratix V D5	2318 images/s	134 images/s	91 images/sec	**25W**
Server + Arria 10 GX1150	-	~233 images/s (projected)	~158 images/sec (projected)	**25W**
Best prior CNN on FPGA [FPGA'15]	-	46 images/s	-	**18W**
Caffe+cuDNN on Tesla K20	-	376 images/s	-	**225W**
Caffe+cuDNN on Tesla K40	-	824 images/s	-	**225W**

See whitepaper @ http://research.microsoft.com/apps/pubs/?id=240715

DEMO

Related Work

- ASICs
    - [Holler'90], [Chen'14], [Cavigelli'15], etc.
- FPGAs
    - [LeCun'09], [Farabet'10], [Aysegul'13], [Gokhale'15], [Zhang'15], etc.
- GPUs/Appliances
    - Nvidia DIGITS, Ersatz, etc.
- Existing solutions not cloud-friendly
    - ASICs, GPUs, and appliances difficult to justify at scale in datacenter
    - ASICs lack flexibility
    - Existing FPGA designs target single FPGA

Conclusions

- Specialized HW for ML is promising for the cloud
  - Inter-networked FPGAs provide scalability, homogeneity, and flexibility
  - Offers compelling performance relative to conventional systems
- Future work
  - Multi-FPGA training pipeline
  - Prototyping on Arria 10
  - OpenCL
- Questions?
  - Eric Chung (erchung@microsoft.com)

Agenda

- Deep Learning on Catapult
- Academic Outreach Program

Academic Outreach

- Purpose
  - Create research eco-system around FPGAs in the data center that will include access to our enabling IP (drivers, shell), potentially research funding, and contests.
- Resources
  - Microsoft will provide FPGA boards, tools, and IP to academics
  - Access to full 48-server machines in a shared cloud
- Look for announcements @
  - http://research.microsoft.com/en-us/projects/catapult