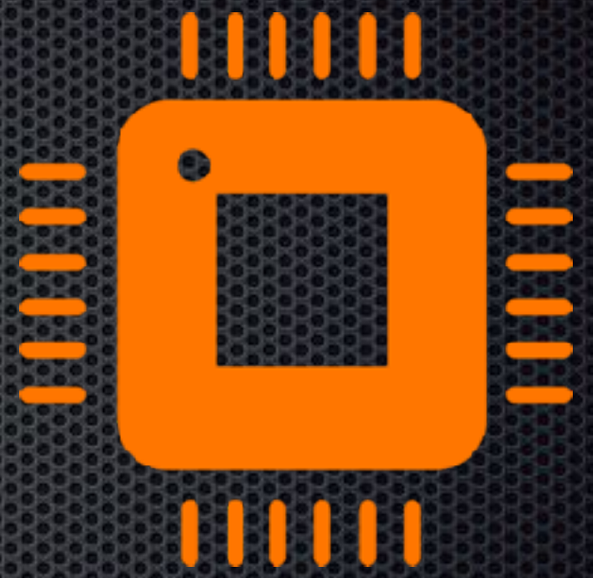
Beyond FPGAs: Compute and Network Accelerators

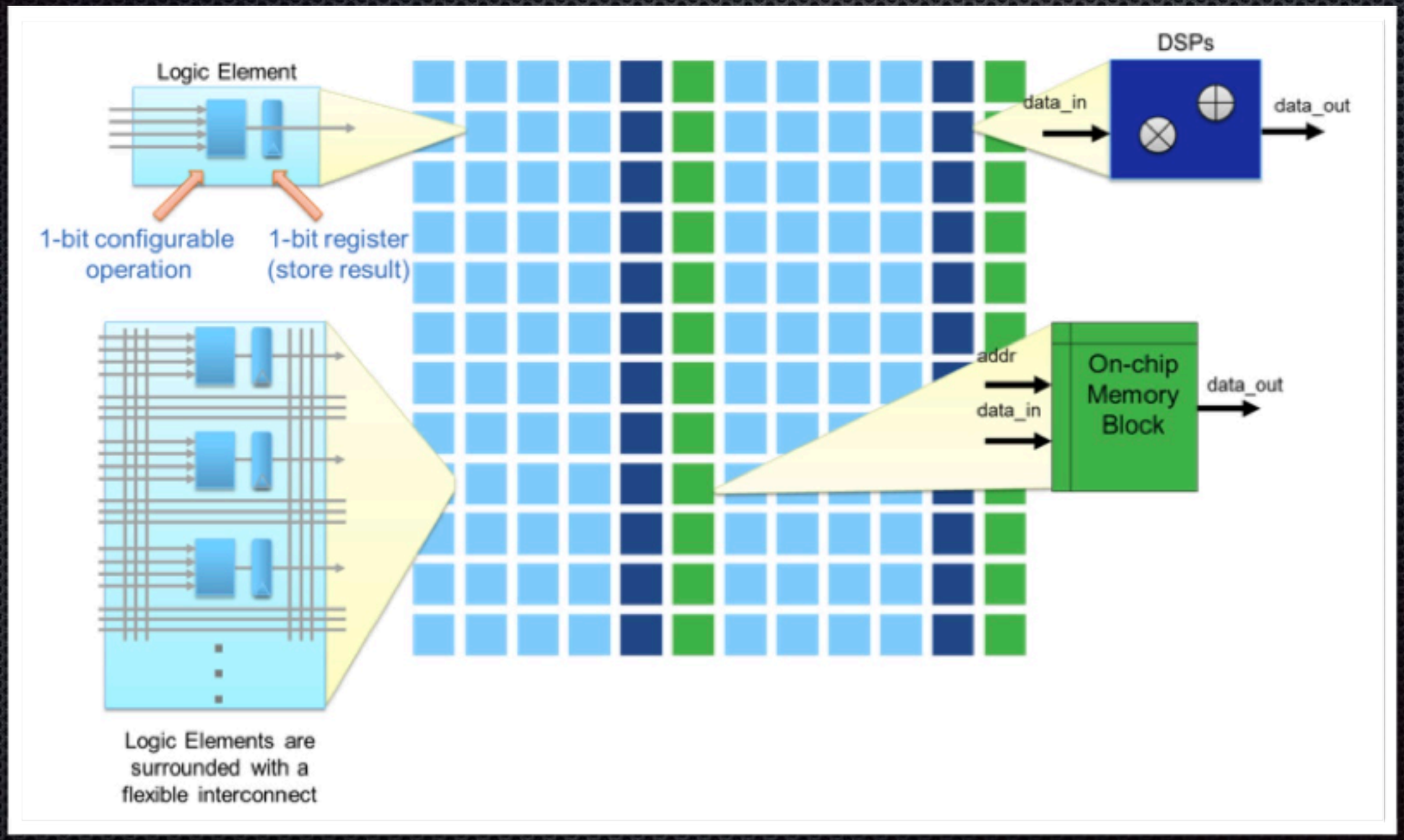
Wenxin Li
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
June 2, 2018

The FPGA



- First commercial product by Xilinx in 1985
- Field Programmable Gate Arrays
- Not a CPU (although you could build one with it)
- 《Lego》 Hardware: logic cells, lookup tables, DSP, I/O
- Small amount of very fast on-chip memory
- Build custom logic to accelerate your SW application

FPGA architecture



https://www.embedded-vision.com/industry-analysis/technical-articles/fpgas-deep-learning-based-vision-processing

I am going to talk about four things

1

FPGAs in the Cloud

2

FPGA for Deep Learning

3

FPGA-based SmartNICs

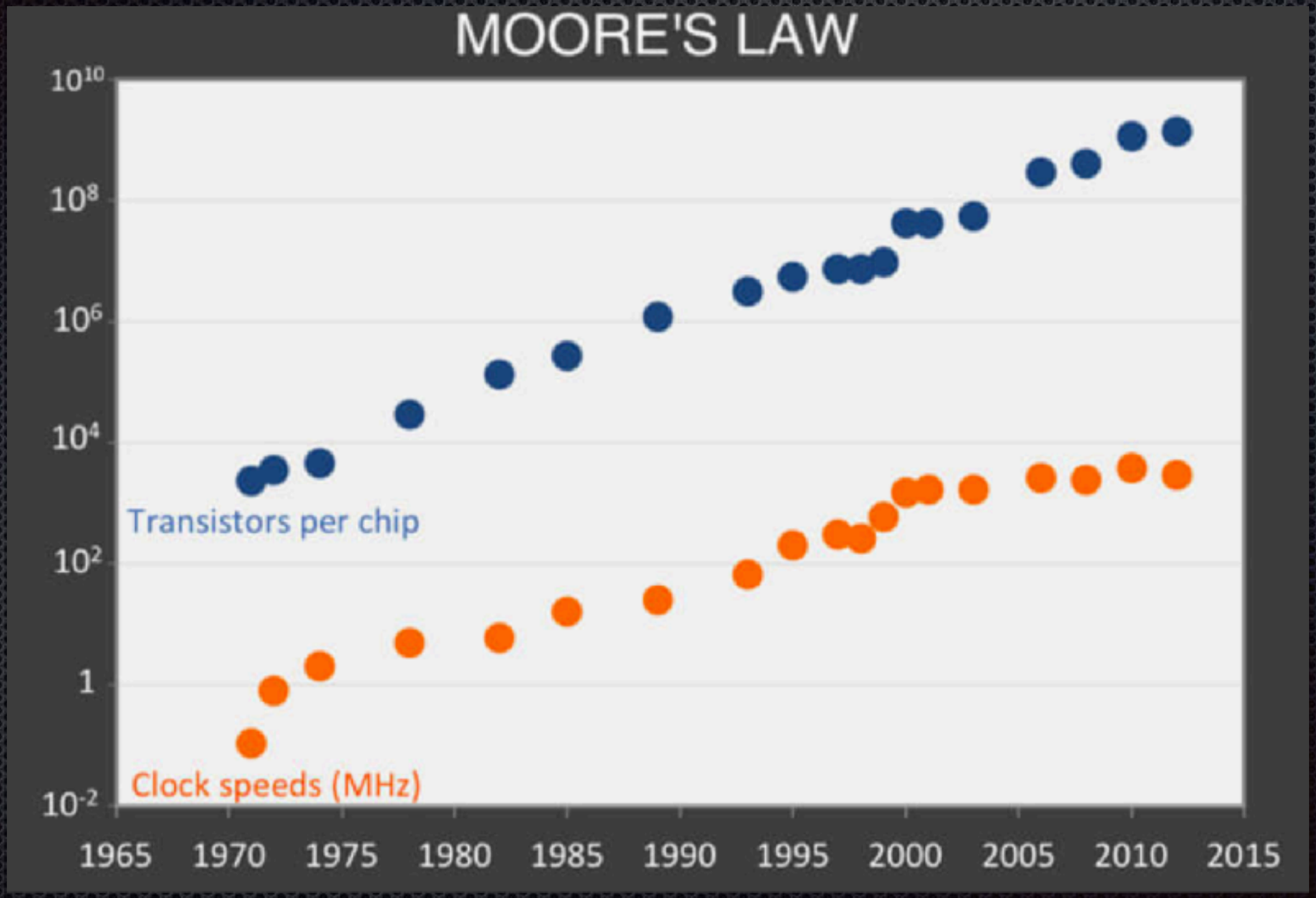
4

Future Research Direction

1

FPGAs in the Cloud

CPU degradation



Source: Intel

GPUs are not optimal for some applications

- Power consumption and efficiency
- Strict latency requirements
- Other requirements
  - Custom data types, irregular parallelism, etc.

- Building your own ASIC may solve this, but
  - It's a huge, costly and risky effort
  - ASICs can't be reconfigured

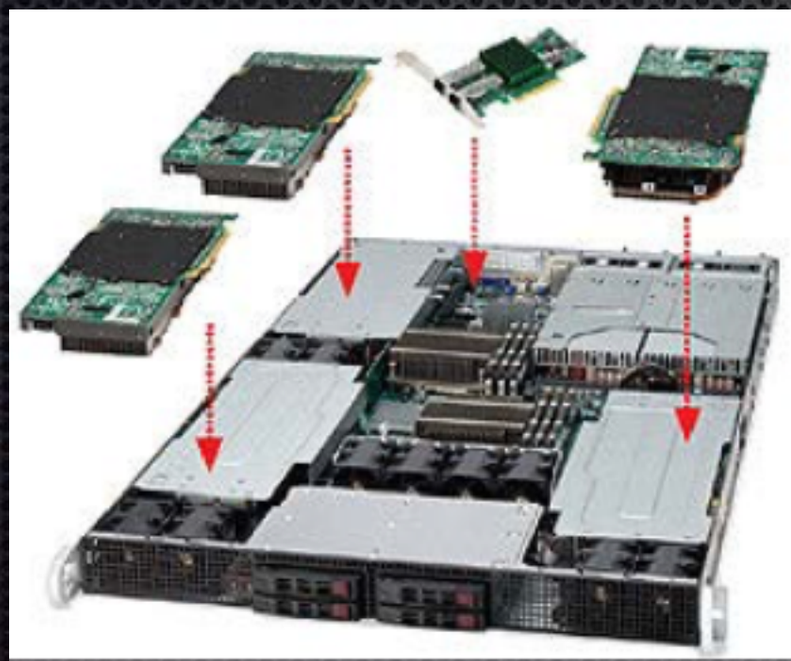
- Time for an FPGA renaissance?

Most Cloud Providers have announced FPGA offerings

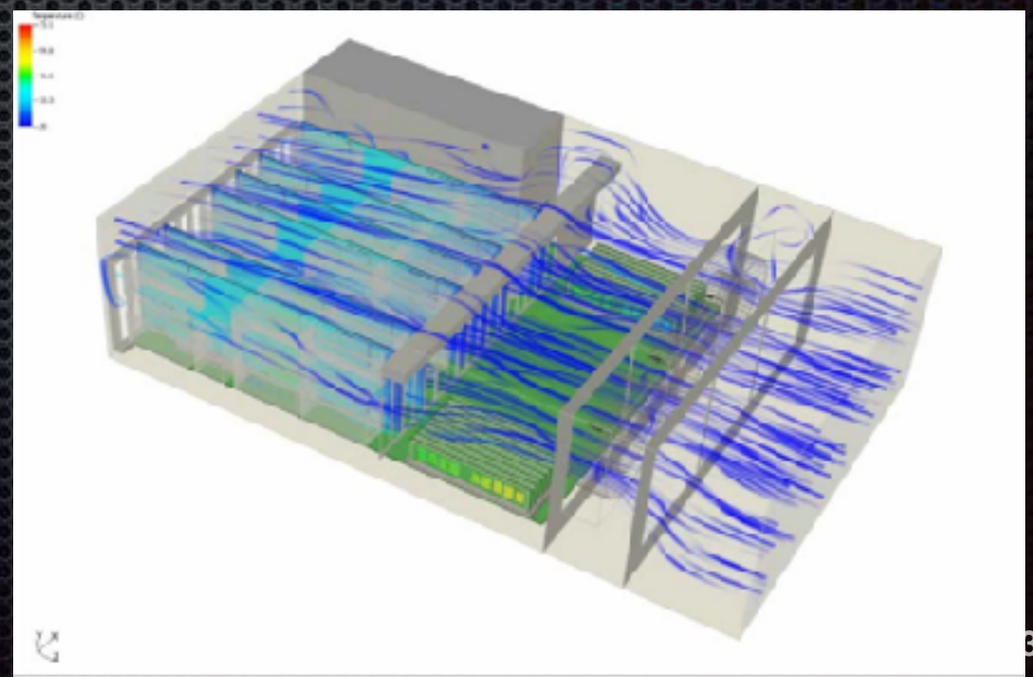
Inside Microsoft's Cloud

Catapult V0: BFB(2011)

- A dedicated rack
- 6 Xilinx Virtex 6 SX315T FPGAs on each card
- 4 cards on each server
- All servers in a rack communicate over 1Gb Ethernet



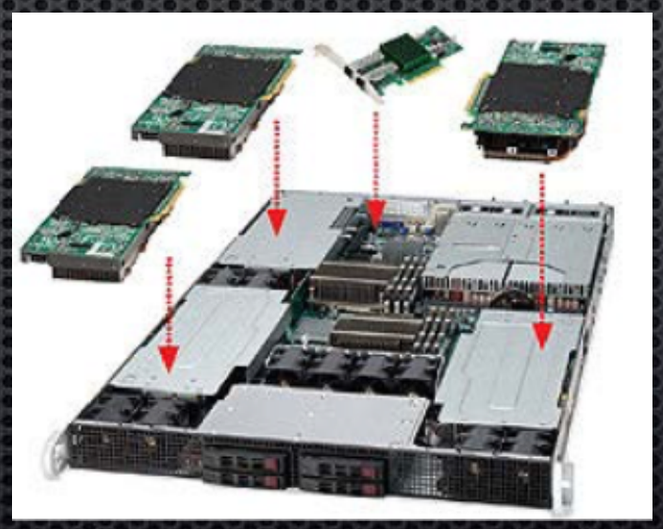
- 1U rack-mounted
- 2 x 10Ge ports
- 3 x16 PCIe slots
- 12 Intel Westmere cores (2 sockets)



Inside Microsoft's Cloud

Several problems with Catapult V0 solution:

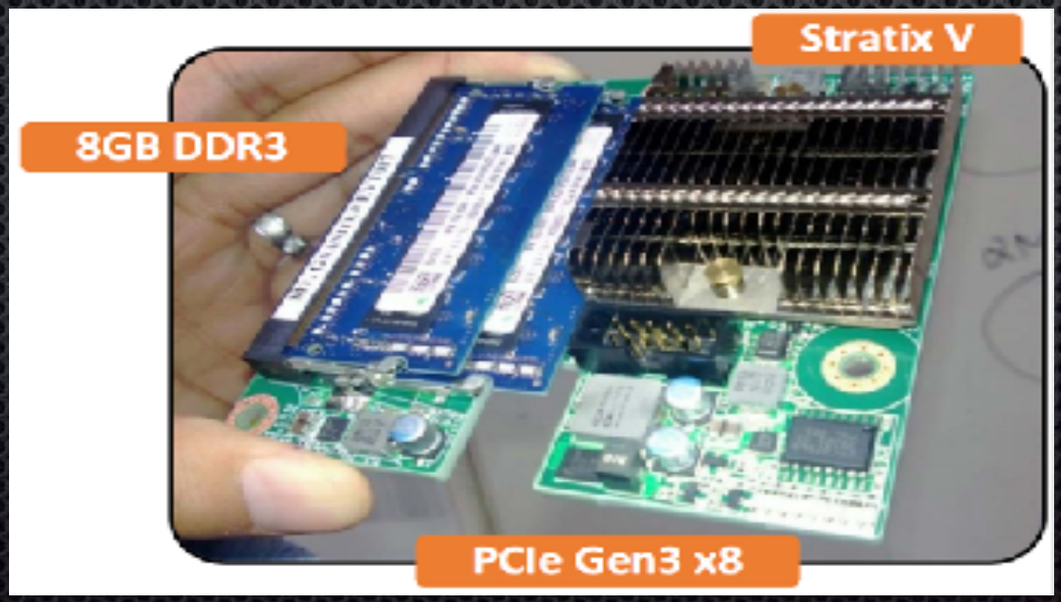
- Inelastic FPGA scaling
- Stranded capacity
- Additional single point failure
- Additional cost on cooling and maintenance
- Too much load on the 1Gb network

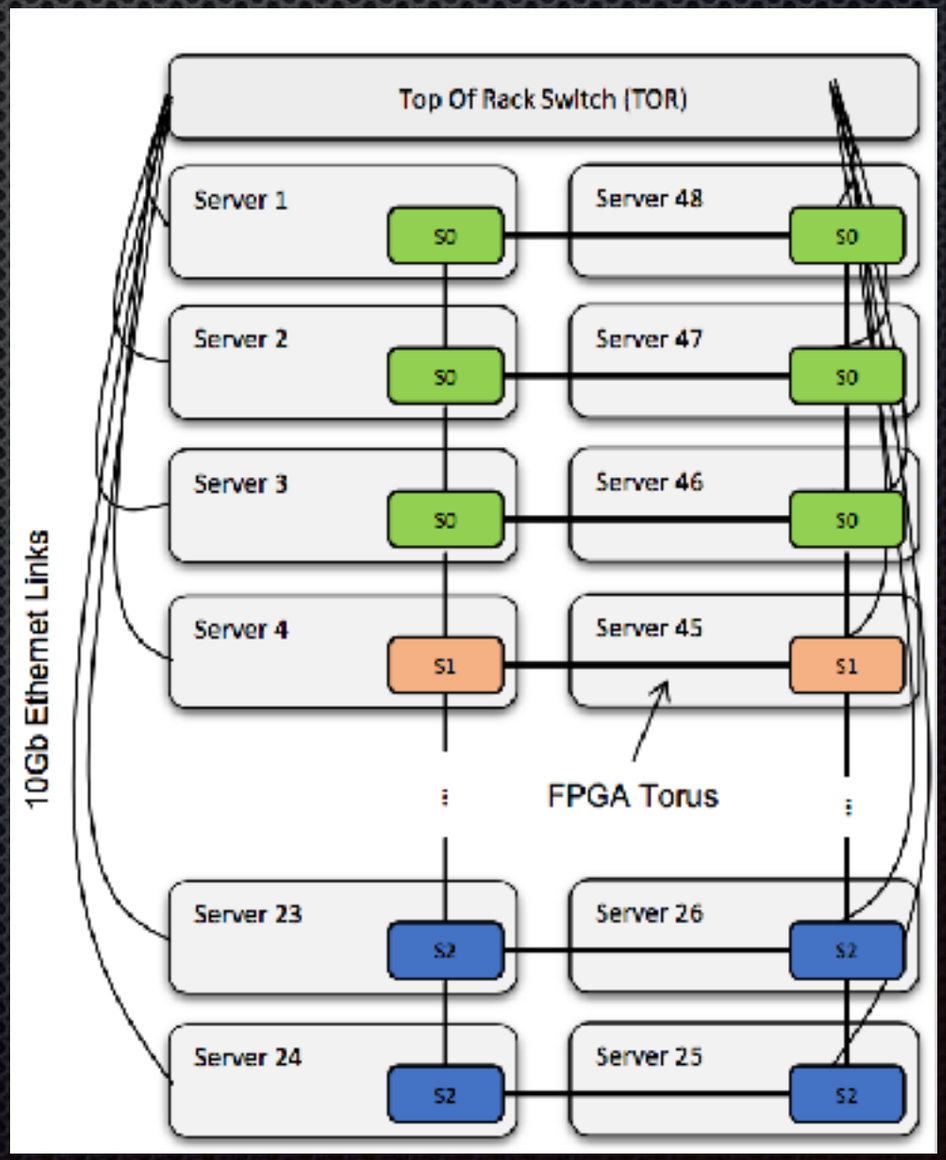


Inside Microsoft's Cloud

Catapult V1 Card (2012-2013): ISCA'14

- Altera Stratix V D5
- 172K ALMs, 457K LEs, 1590 DSPs
- 2014 M20Ks:
  - M20K is a 2.5KB SRAM
- PCIe Gen 2x8, 8GB DDR3
- 20Gb (2x10Gb) links





Inside Microsoft's Cloud

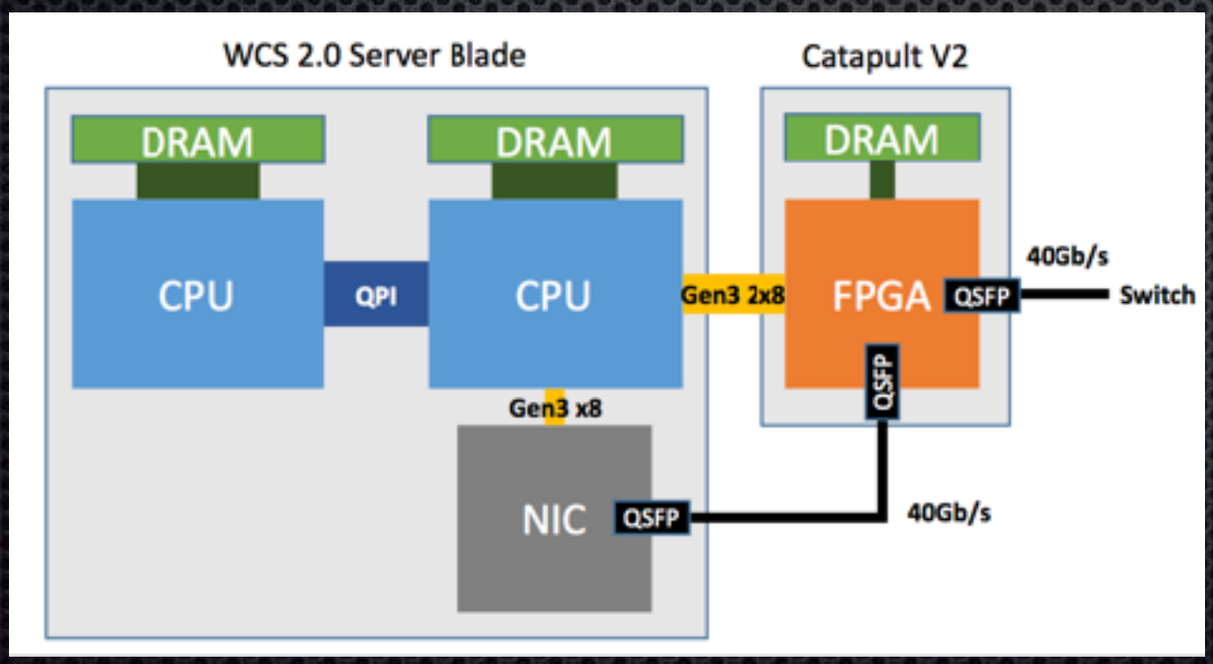
Several limitations in Catapult V1 solution:

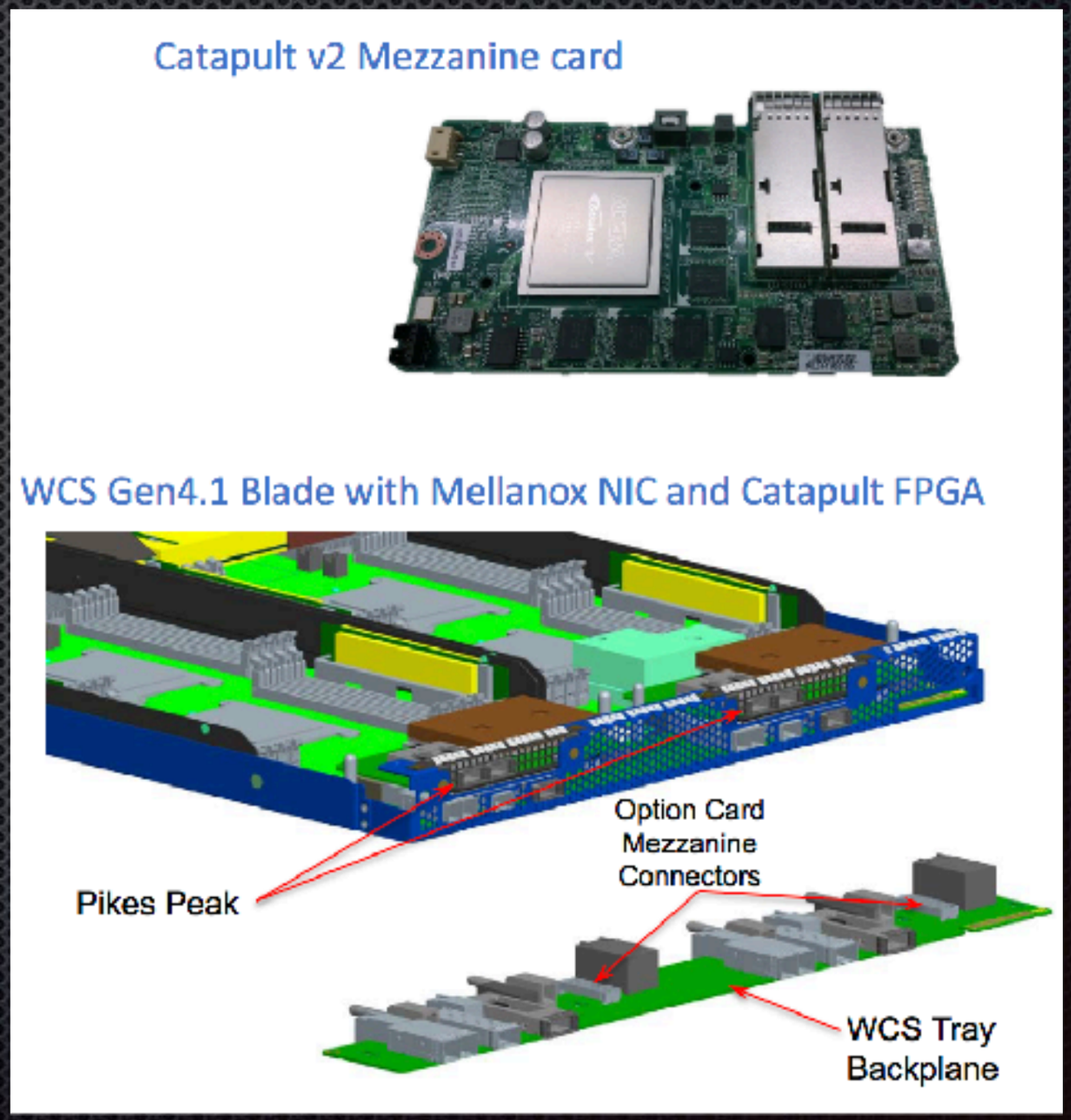
- No one else wanted the secondary network
  - Complex
  - Require awareness of the physical location of machines
- Difficult to handle failures
- Limited scalability
- No killer infrastructure accelerator
  - Application presence is too small

Inside Microsoft's Cloud

Catapult V2: Micro'16

- Can act as a local compute accelerator
- Can act as a network accelerator

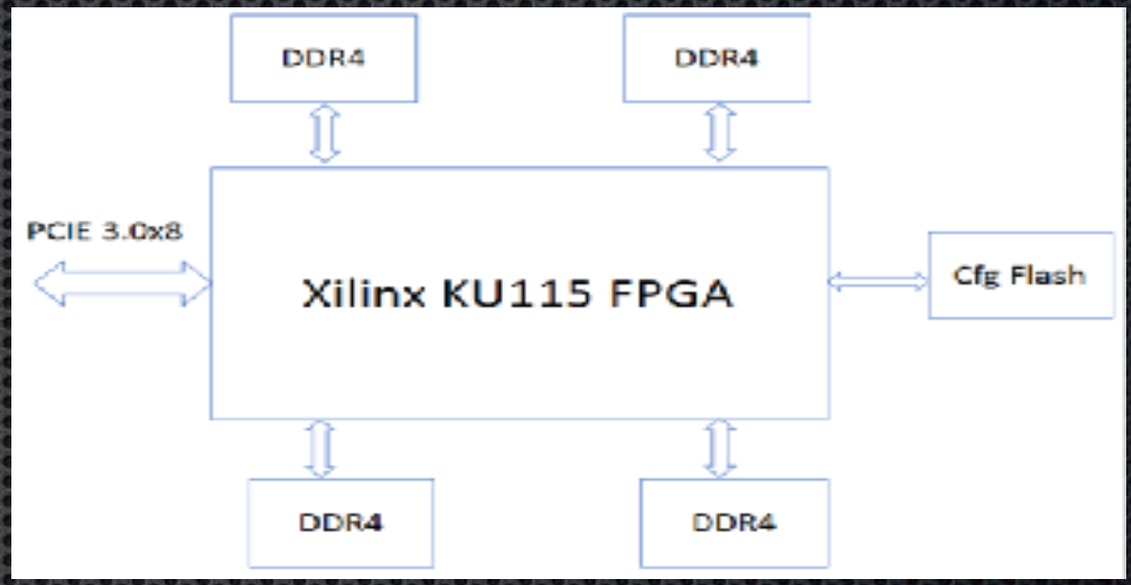




FPGAs in other Clouds

Model	FPGAs	vCPU	Mem (GiB)	SSD Storage (GB)	Networking Performance
f1.2xlarge	1	8	122	470	Up to 10 Gigabit
f1.16xlarge	8	64	976	4 x 940	20 Gigabit

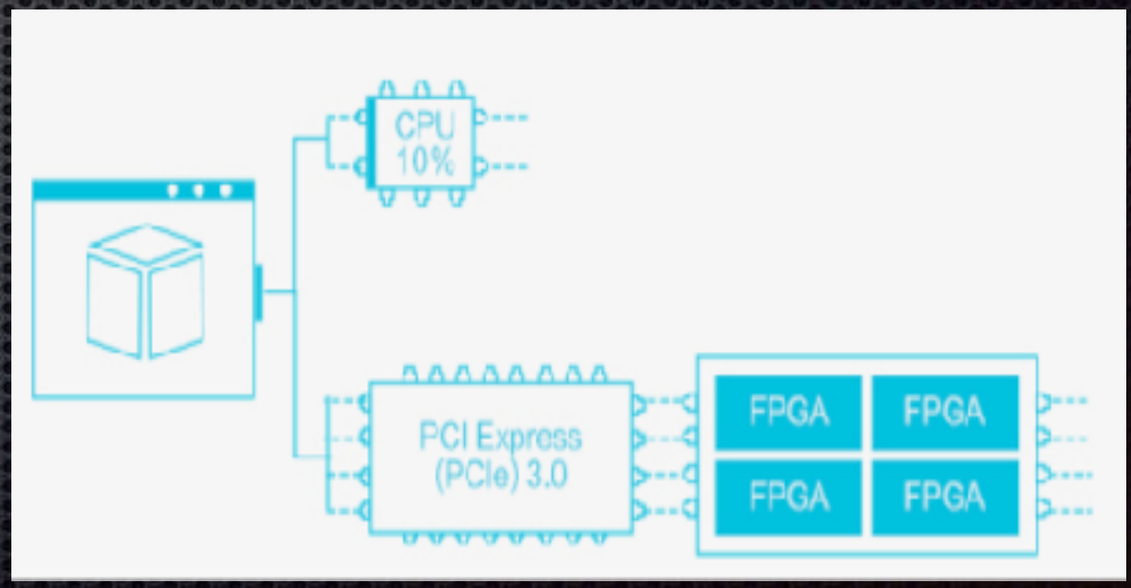
Amazon



Baidu

实例规格	FPGA	vCPU	内存 (GiB)	数据盘	网络
FX2.7xlarge60	1	14核	60	搭载SSD云盘	万兆网络
FX2.14xlarge120	2	28核	120	搭载SSD云盘	万兆网络
FX2.28xlarge240	4	56核	240	搭载SSD云盘	万兆网络

Tencent



Alibaba

2

FPGA for Deep Learning

The rise of deep learning

Unmanned Vehicle

Speech & Audio

Text & Language

Genomics

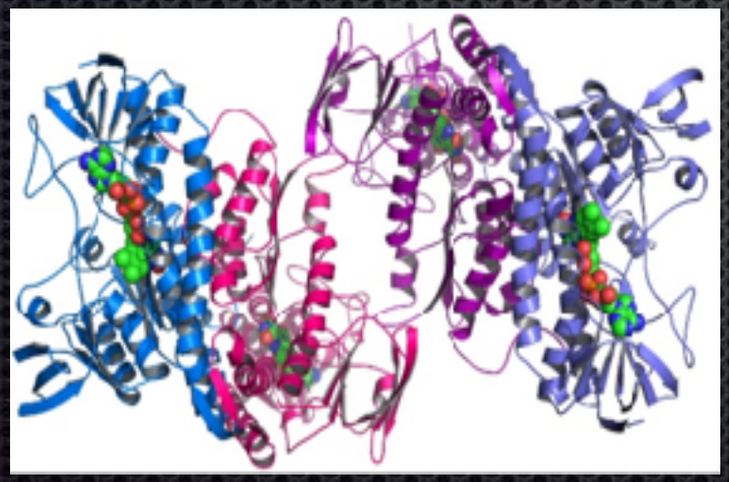


Image & Video

Multi-media

Why deep learning on FPGA

Why deep learning on FPGA

Flexibility

Why deep learning on FPGA

Flexibility

Latency

Why deep learning on FPGA

Flexibility

Latency

Precision

Timeline of important events in FPGA deep learning research

1987 — VHDL for FPGAs is first standardized by IEEE

1992 — GANGLION is first FPGA neural network implementation (Cox et al.)

1994 — First generation behavioral synthesis for FPGAs introduced by Synopsys

1996 — VIP is the first FPGA CNN implementation (Cloutier et al.)

2005 — FPGA market approaches $2 billion

2006 — First implementation of back-propagation to achieve 5 G ops on FPGA (Paul et al.)

2011 — Altera introduces OpenCL support for FPGAs

Emergence of large-scale FPGA-based CNN research (Farabet et al.)

2015 — Emergence of FPGA-based CNN acceleration for datacenter based on Microsoft Catapault project

FUTURE

Deep Learning Processor List

IC Vendors	Intel, Qualcomm, Nvidia, Samsung, AMD, Xilinx, IBM, STMicroelectronics, HiSilicon, Rockchip
Tech Giants & HPC Vendors	Google, Amazon_AWS, Microsoft, Alibaba, Tencent, Baidu, etc.
IP Vendors	ARM, Synopsys, Imagination, CEVA, Cadence, VeriSilicon, Videantis
Startups in China	Cambricon, Horizon Robotics, DeePhi, Bitmain, Chipintelli, Thinkforce
Startups Worldwide	Cerebras, Wave Computing, Graphcore, PEZY, KnuEdge, Tenstorrent, ThinCI, Koniku, Adapteva, Knowm, Mythic, Kalray, BrainChip, AImotive, DeepScale, Leepmind, Krtkl, NovuMind, REM, etc.

Source: https://basicmi.github.io/Deep-Learning-Processor-List/

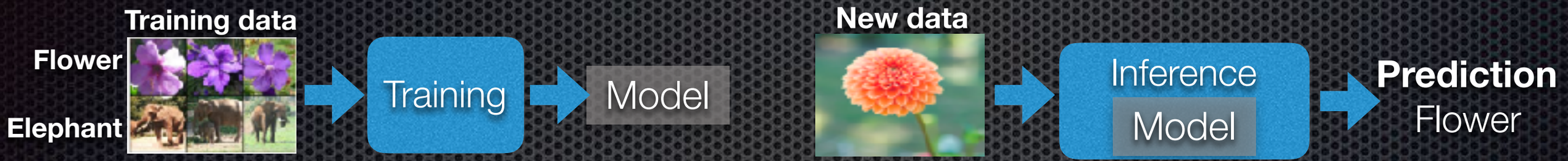
How FPGAs accelerate deep learning

Deep Neural Networks (DNN)

Popular Machine Learning approach for data analytics

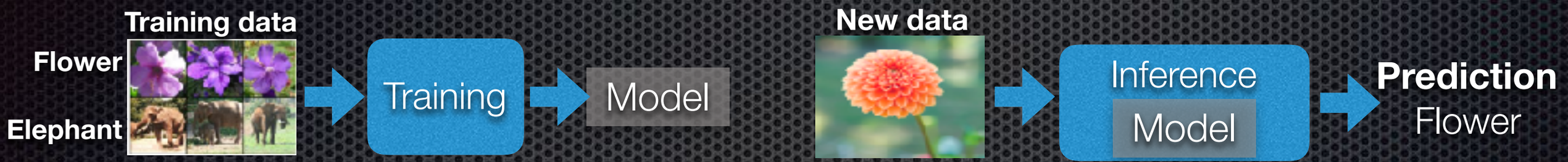
Deep Neural Networks (DNN)

Popular Machine Learning approach for data analytics

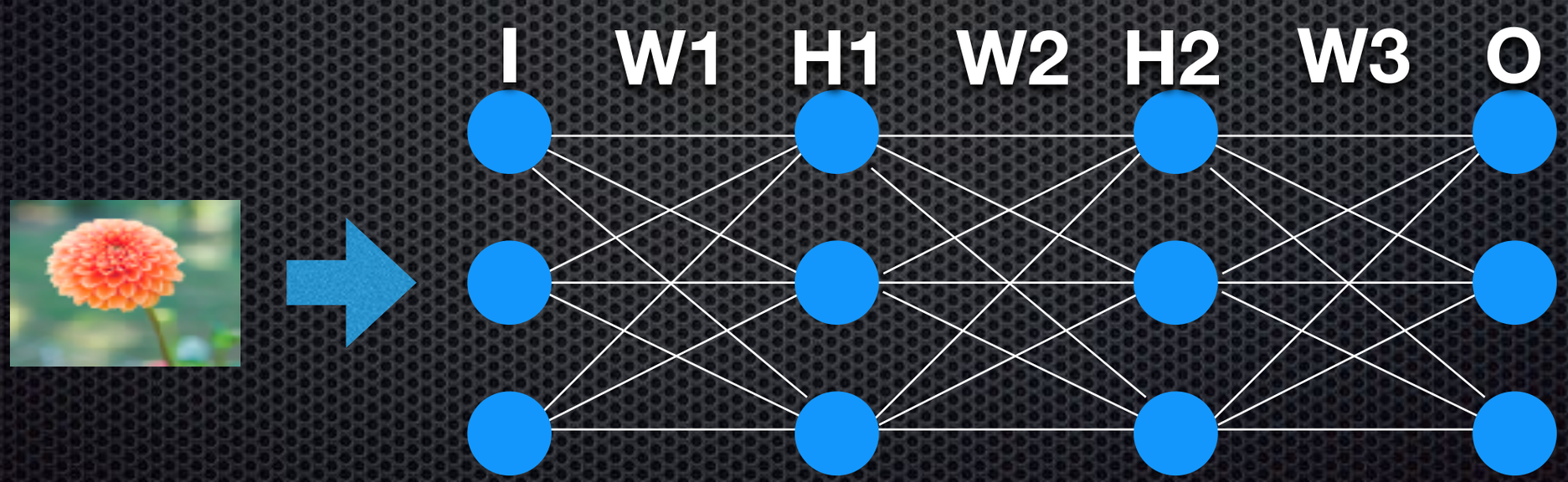


Deep Neural Networks (DNN)

Popular Machine Learning approach for data analytics

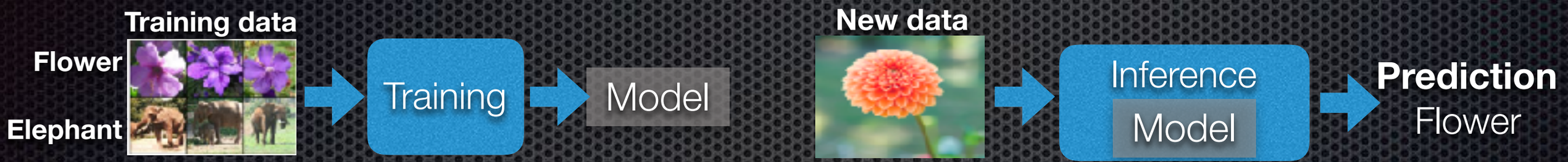


Consists of layers of neurons connected via weighted edges

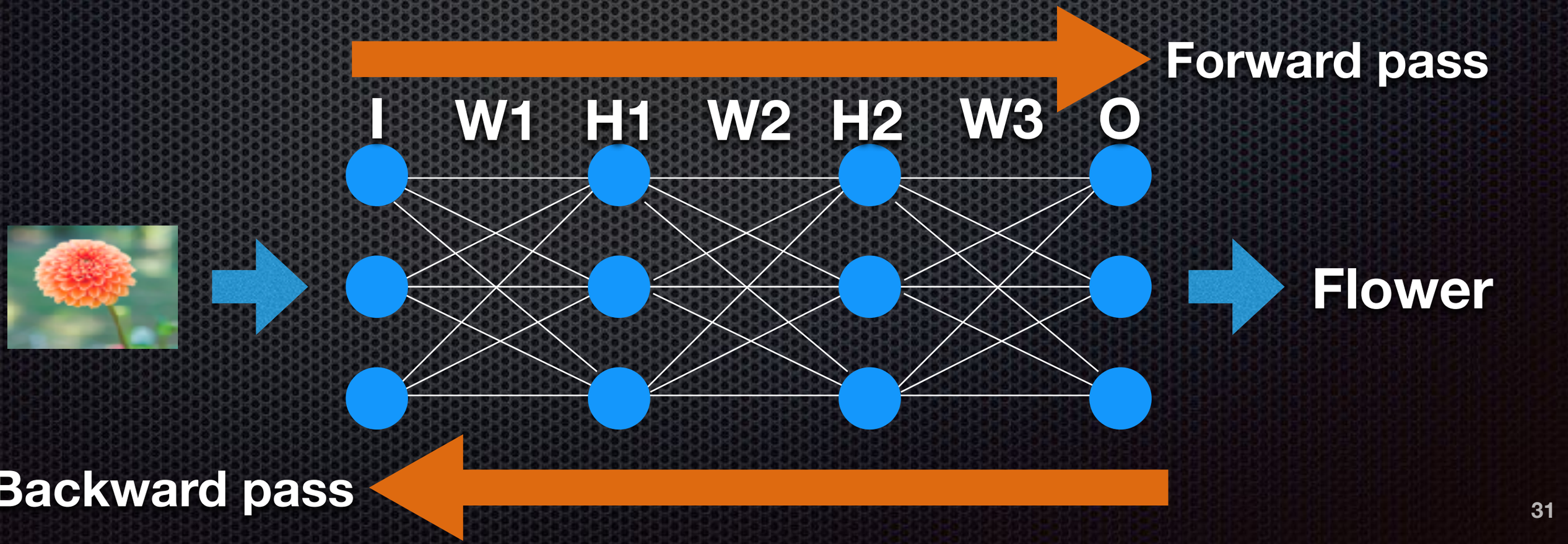


Deep Neural Networks (DNN)

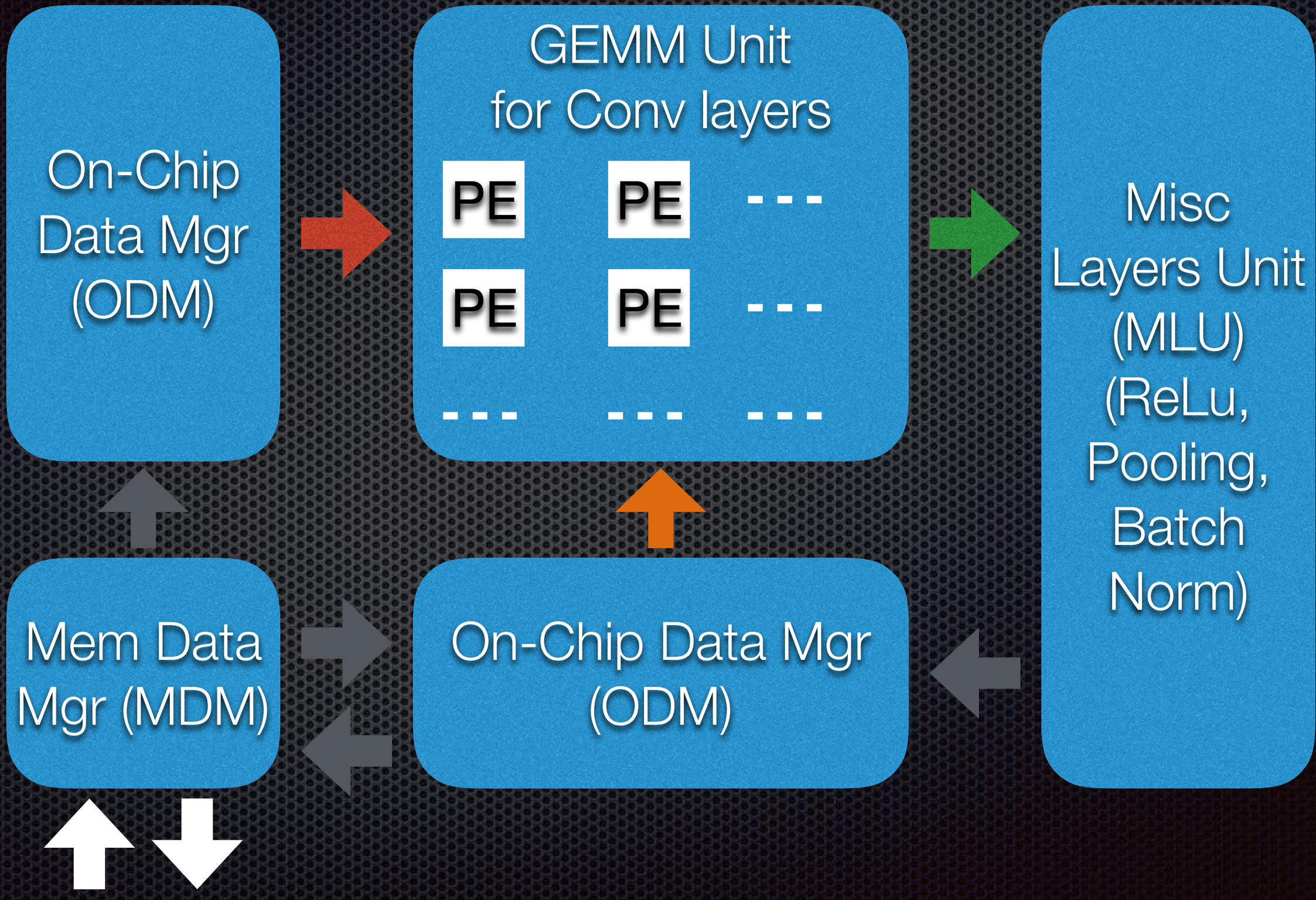
Popular Machine Learning approach for data analytics



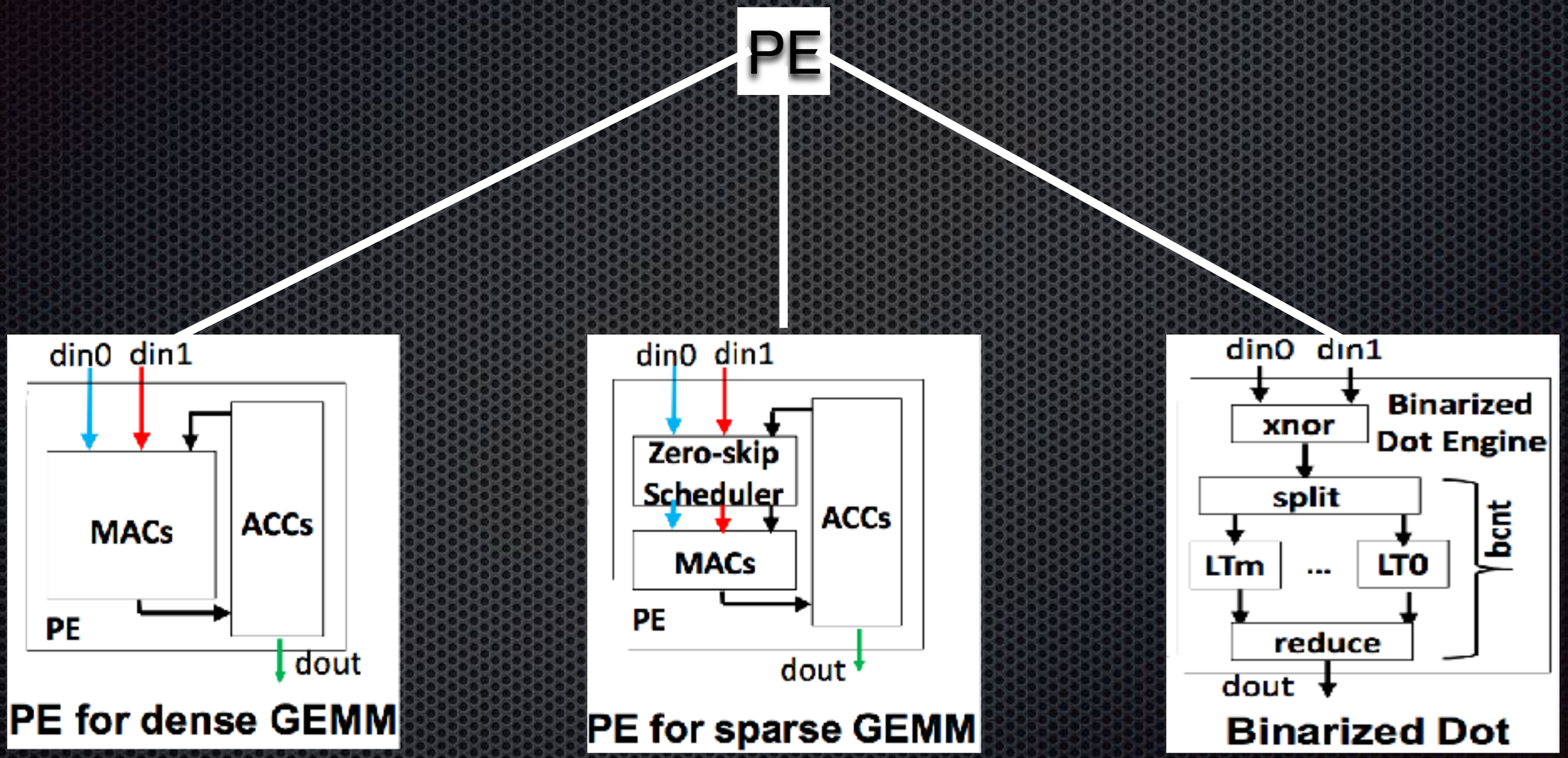
Consists of layers of neurons connected via weighted edges



Typical FPGA design for DNN



Typical FPGA design for DNN

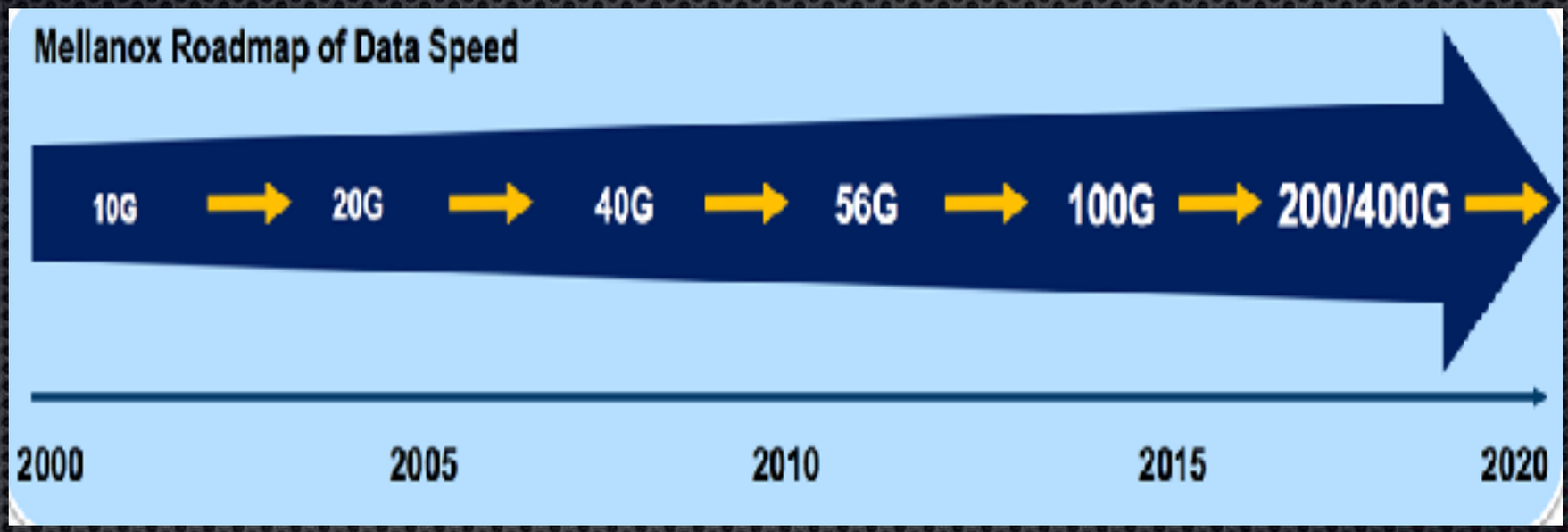


3

FPGA-based SmartNICs

Incommensurate Scaling

- Server network bandwidth is growing rapidly



- The compute capacity headroom for processing network I/O is shrinking
  - CPU-bound

The Need for Line Rate Processing

SmartNICs: old idea

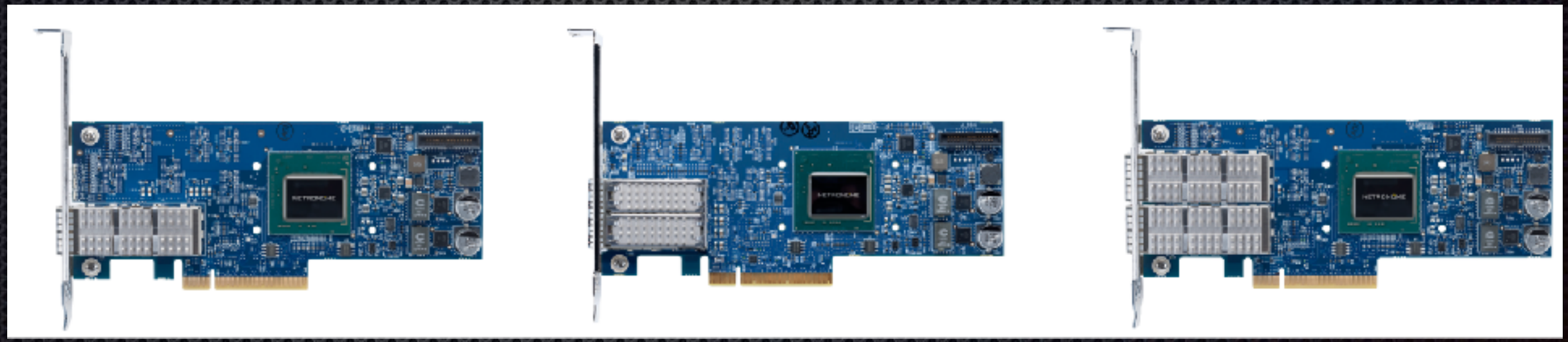
The idea of adding programmable hardware to the NIC was actively researched in the early 2000s with the introduction of Network Processors

Modern SmartNICs

- SoC-based
  - Mellanox BlueField
  - Cavium LiquidIO
  - Netronome Agilio-CX
- FPGA-based
  - Mellanox InnovaFex
  - Azure SmartNIC

SoC-based SmartNICs

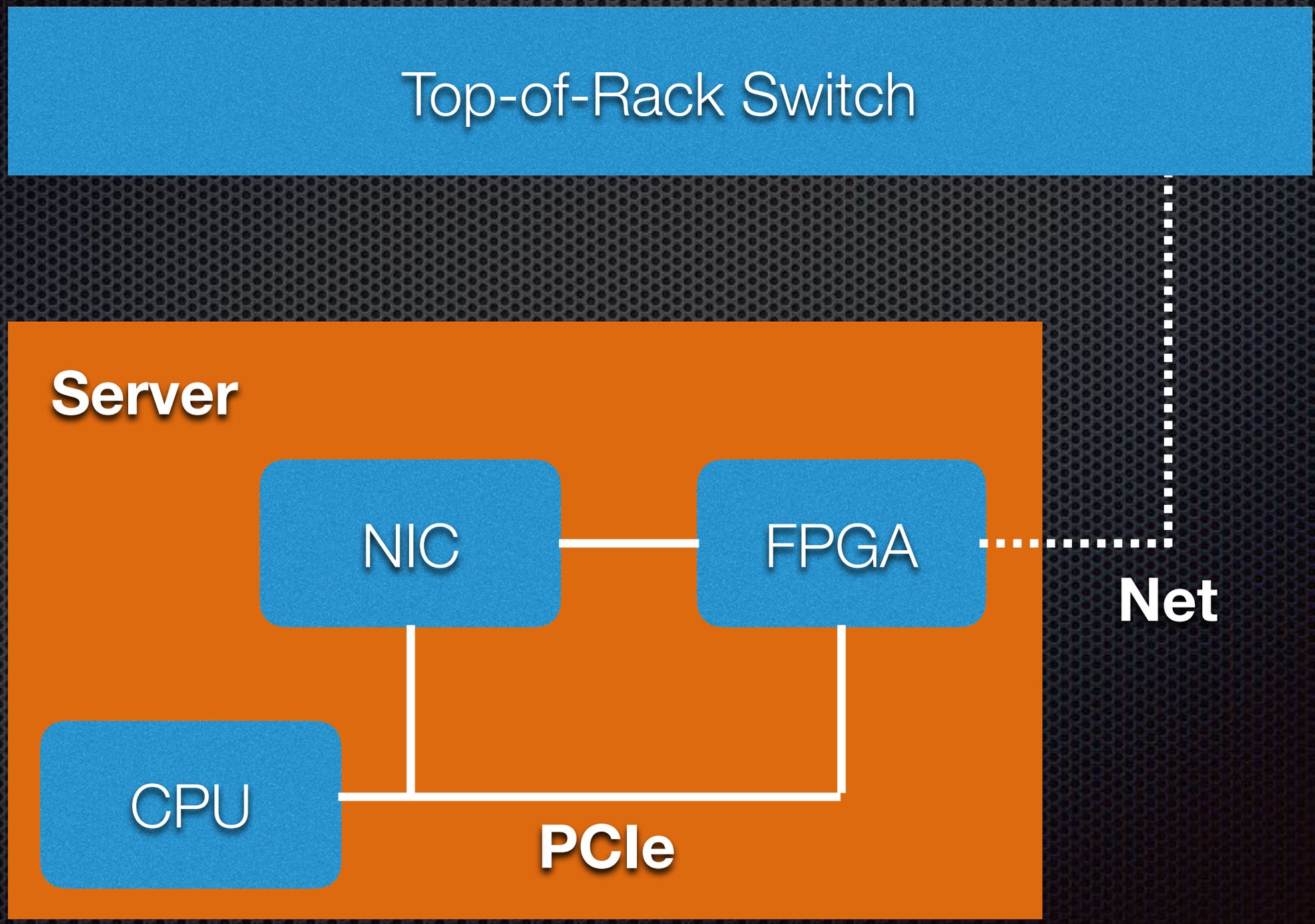
- Advantage
  - Easier to Programm
- Disadvantage
  - High latency
  - Poor scalability



Agilio-CX SmartNICs

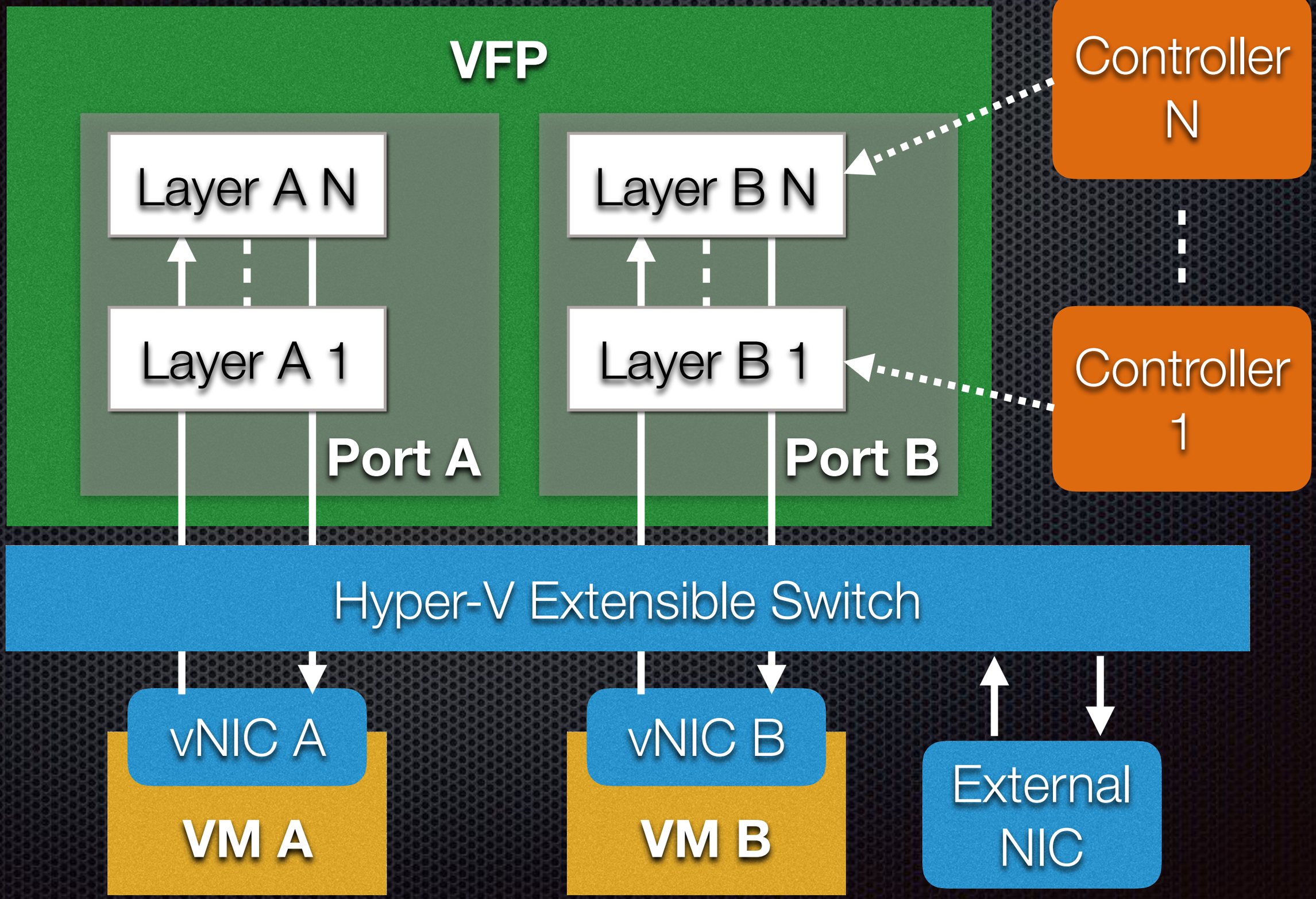
10/40/100GbE solutions

FPGA-based SmartNICs

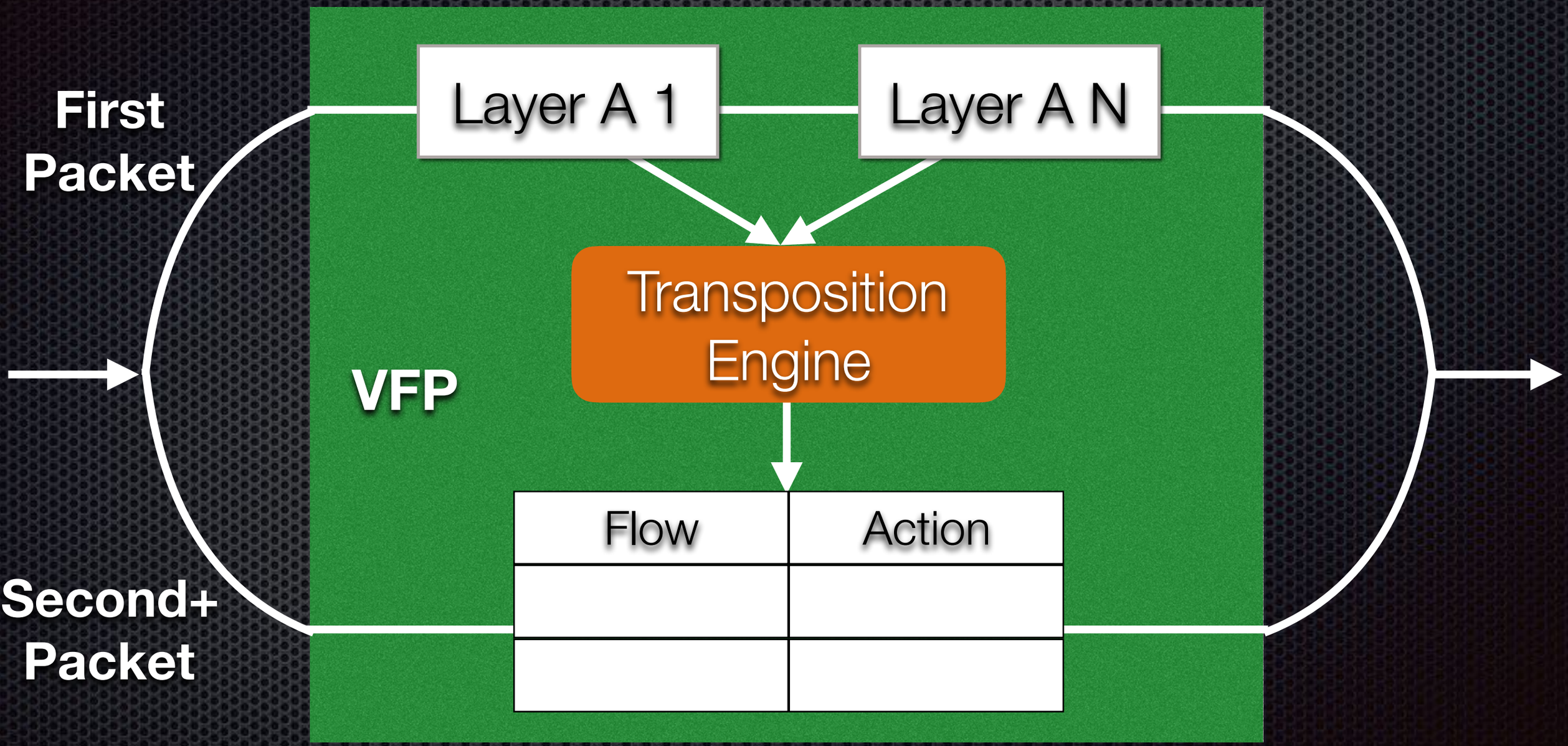


How FPGA-based SmartNICs work?

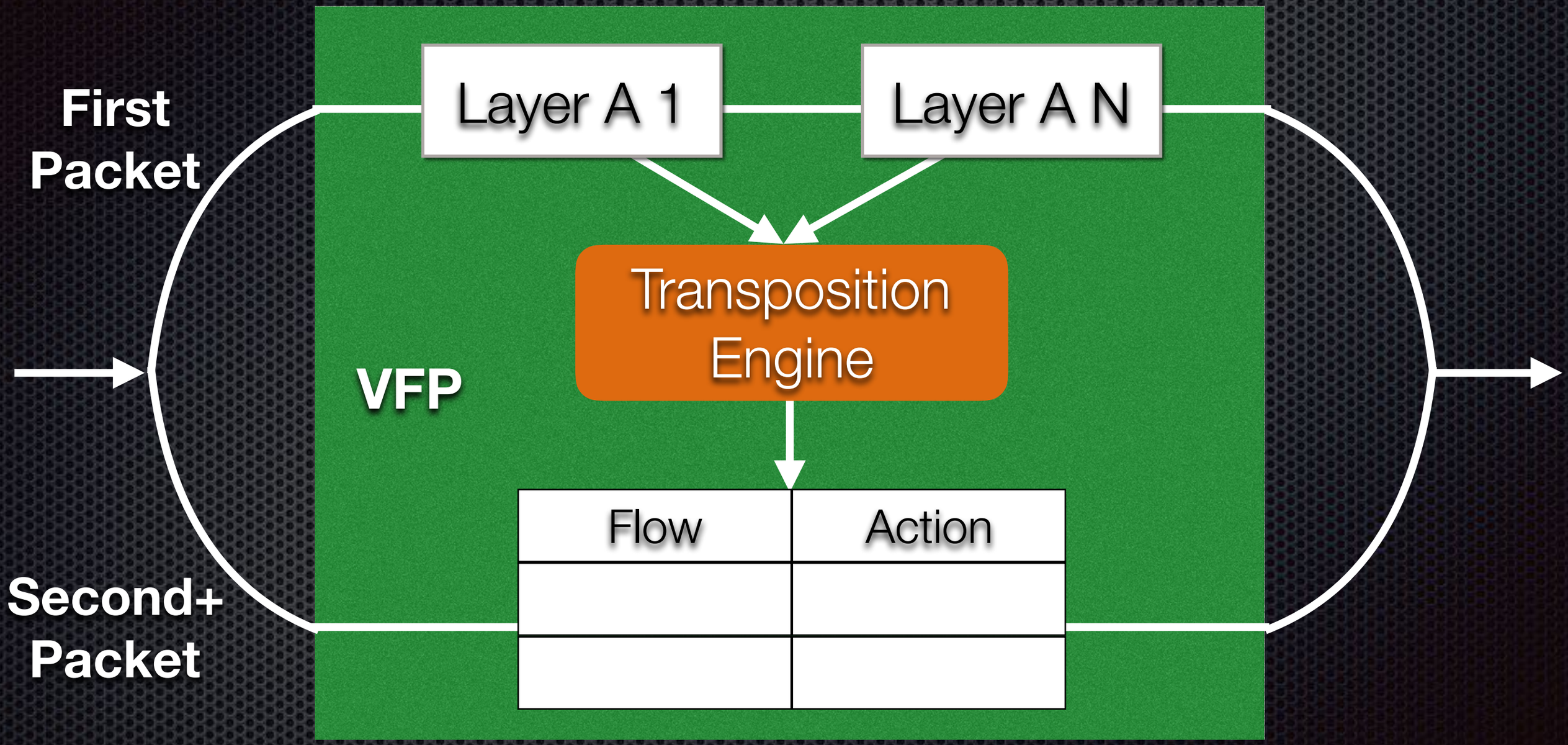
Microsoft VFP



Microsoft Unified Flow Table

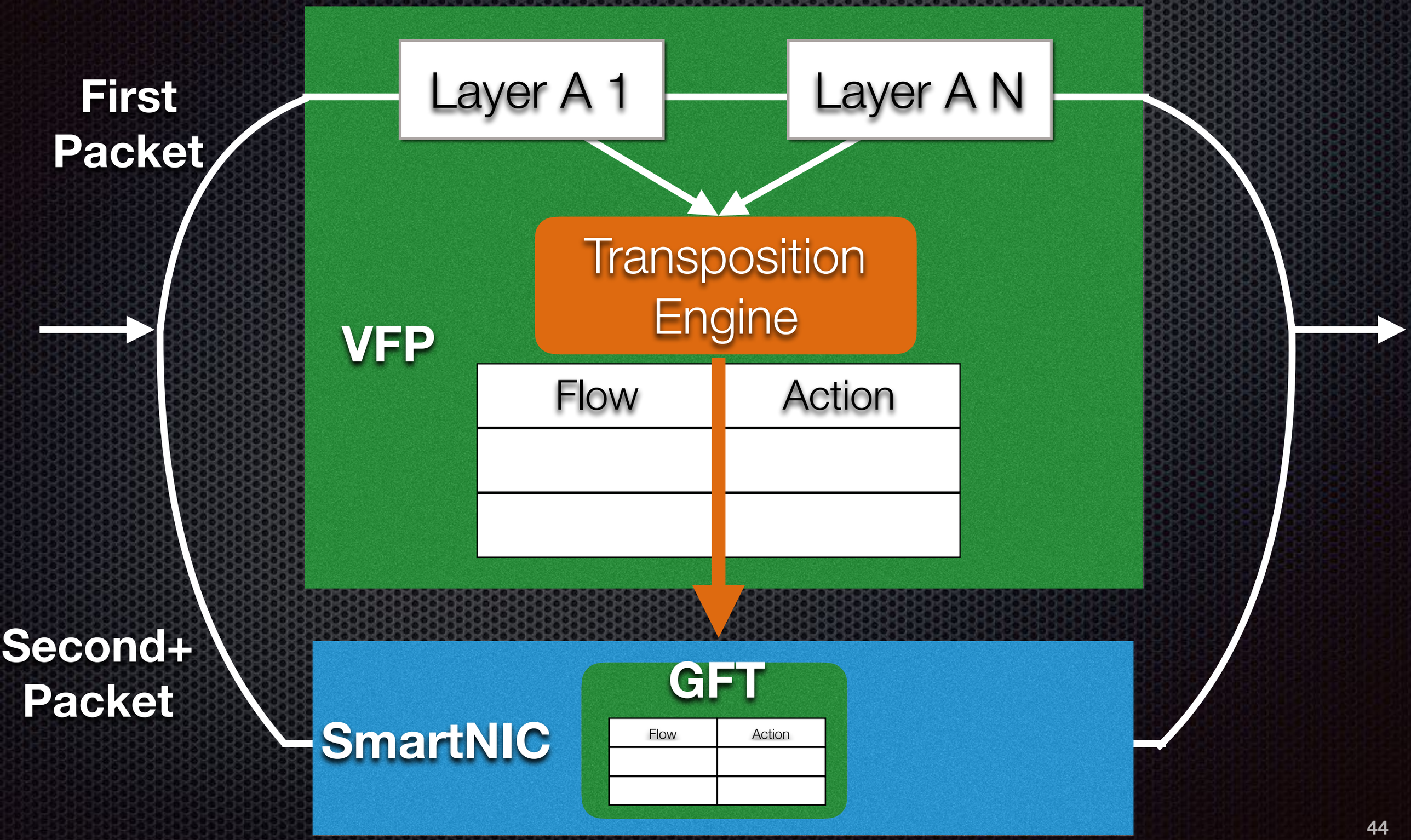


Microsoft Unified Flow Table



Host SDN worked well at 1GbE, ok at 10GbE… what about 40GbE+?

Offloading with SmartNIC



4

Future Research Direction

Programming Abstractions

TensorFlow | CNTK | Caffe — **Applications**

Matrix Multiply | Max pooling | ReLu | Batch Norm — **Building Blocks**

CPU | GPU | FPGA — **Accelerators**

Offload more

Congestion Control

Flow scheduling

Rate limiting

packet segmentation

scatter-gather I/O

checksum computation

Thank you!