Microsoft

Azure

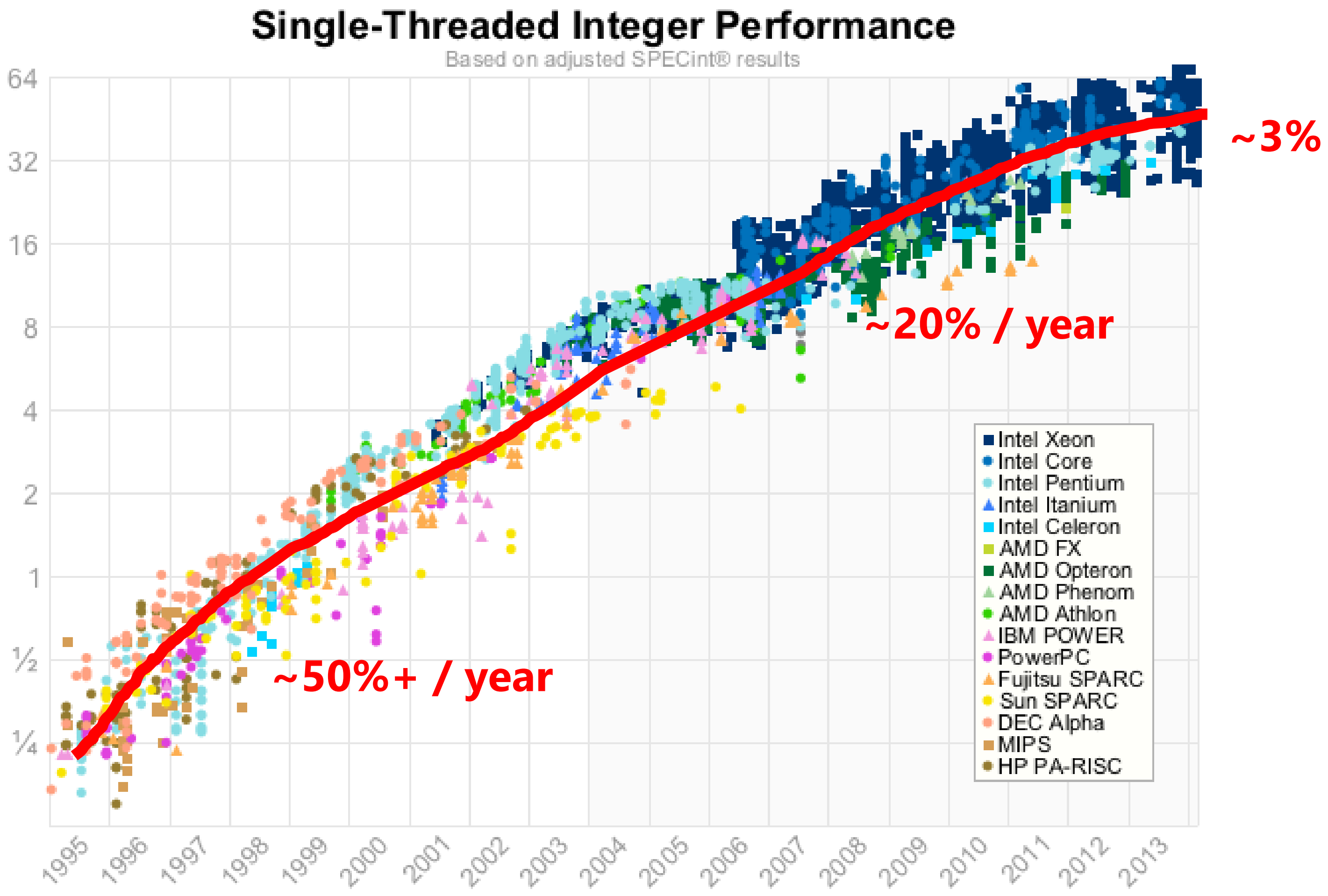
# Heterogeneous Computing @ Microsoft

**Andrew Putnam – Azure Networking / Microsoft Research**
**Kalin Ovtcharov – AI & Architectures**

September 11, 2019

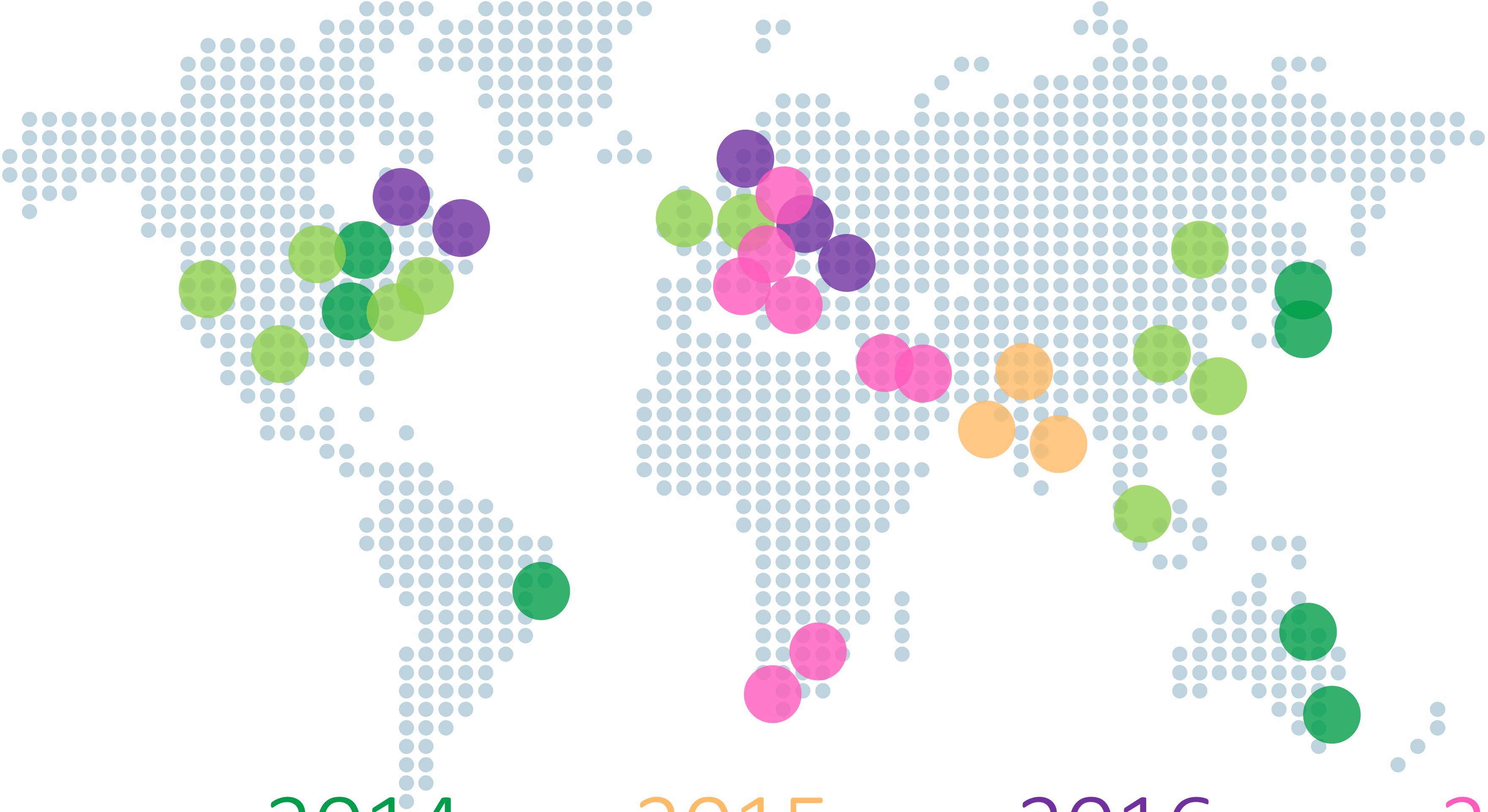
# Technology Scaling

- Interactive Cloud apps rely on single threaded performance
- CPUs aren’t getting much faster. We just get a few more
- 2x users requires ~2x the number of servers



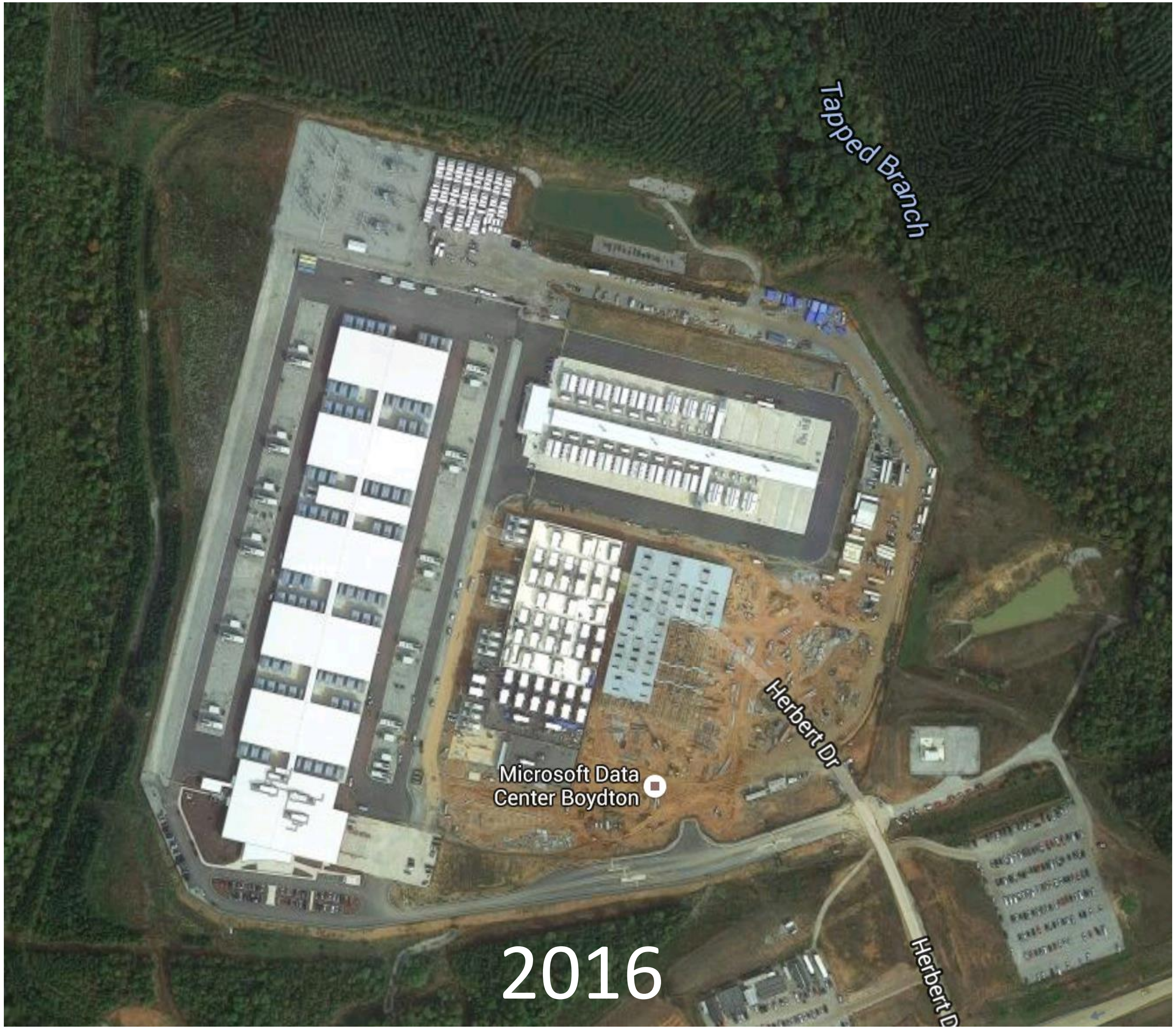
Jeff Preshing, Henk Poley, http://preshing.com/20120208/a-look-back-at-single-threaded-cpu-performance/

# Datacenter Scaling

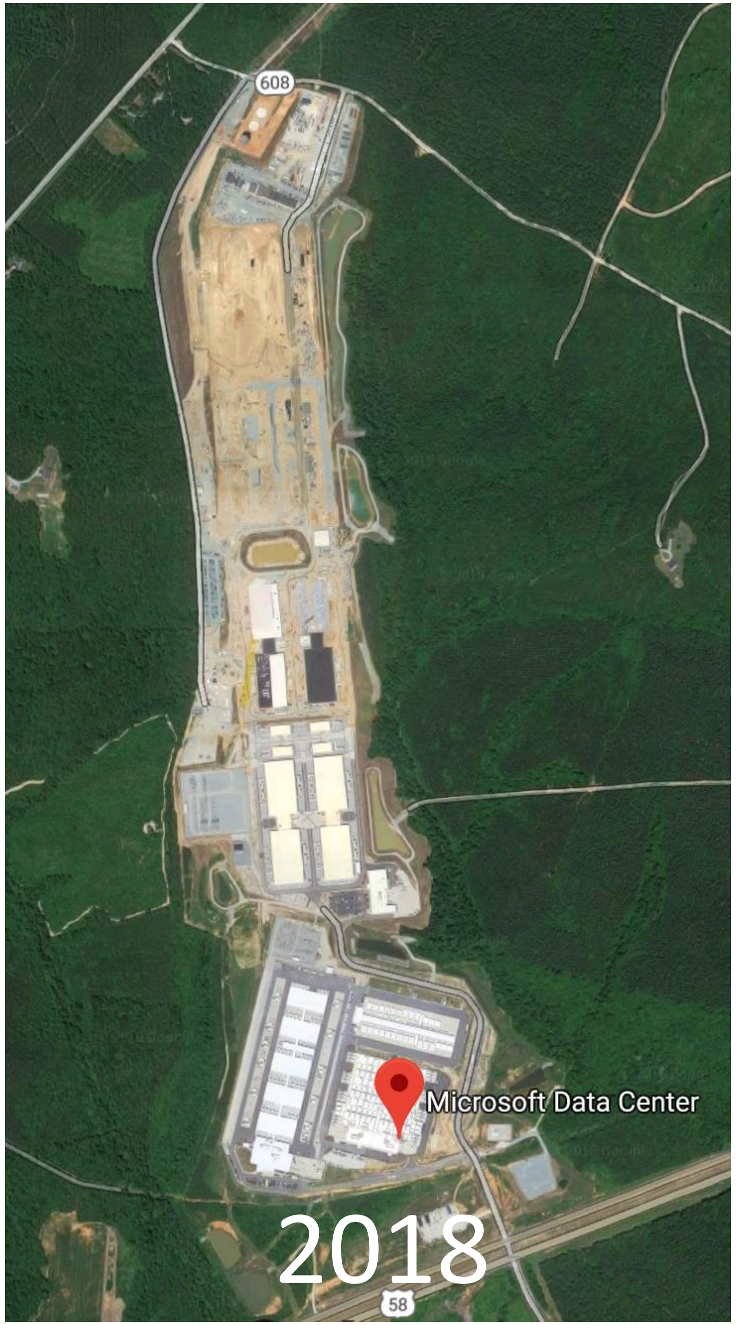


2013 2014 2015 2016 2017

~100%+ Growth for the past 5 years



Tapped Branch
Herbert Dr
Microsoft Data Center Boydton
Herbert Dr
2016



608
Microsoft Data Center
2018
58

# Cloud Server Changes

| | 2012 | 2018 | Ratio |
|---|---|---|---|
| CPU Cores | 16 | 36 | 2.25x |
| Storage | 4 TB HDD | 7 TB SDD (M.2)<br>(120TB HDD*) | 1.75x<br>30x |
| Network | 1Gb | 50Gb | 50x |

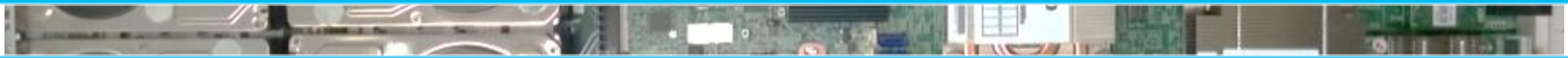
* - Bing SKU

# Cloud Server Changes

| | 2012 | 2018 | Ratio |
|---|---|---|---|
| CPU Cores | 16 | 36 | 2.25x |
| Storage | 4 TB HDD | 7 TB SDD (M.2)<br>(120TB HDD*) | 1.75x<br>30x |
| Network | 1Gb | 50Gb | 50x |

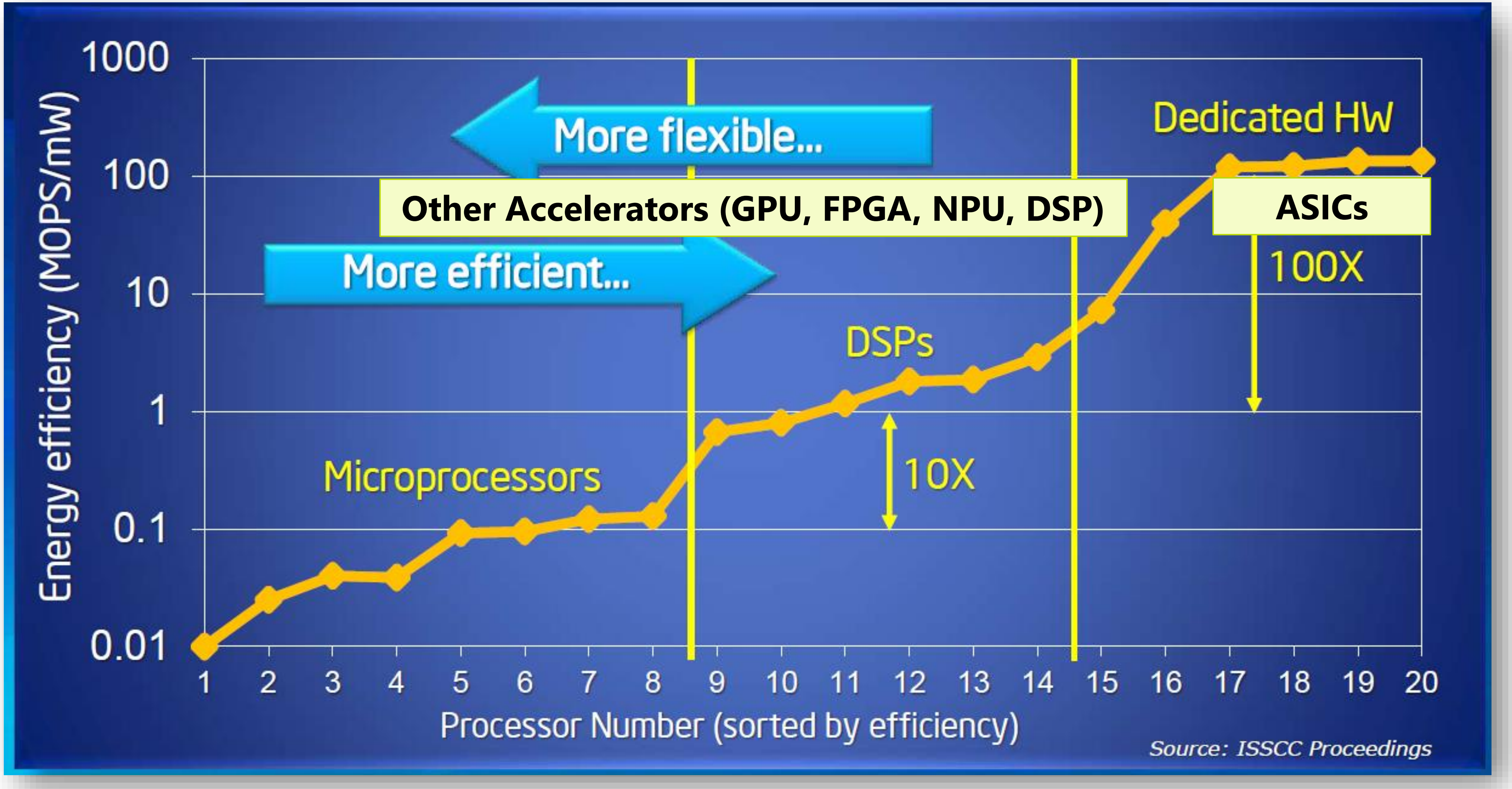
* - Bing SKU

Increasing Server Lifetimes



More Data, Coming in Faster… and CPUs aren’t keeping up

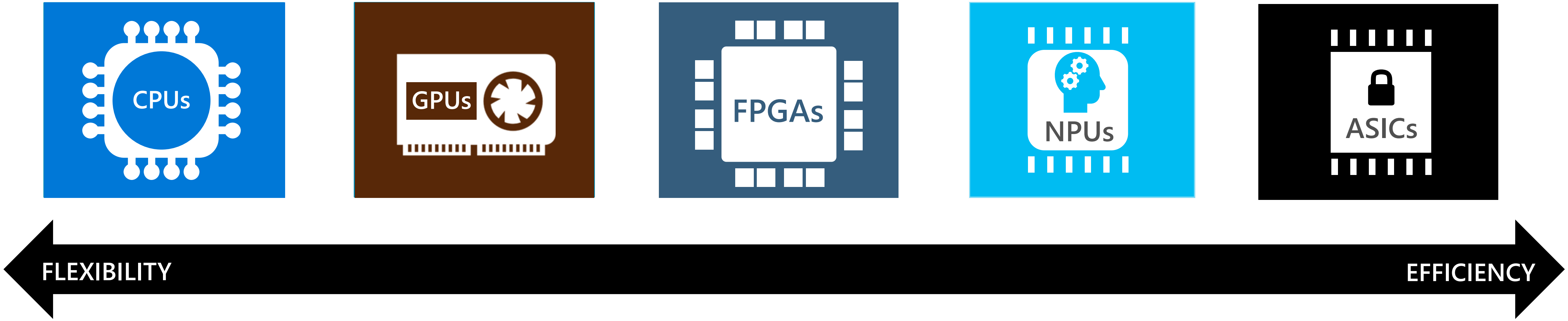
# Performance & Efficiency via Specialization



*Source: ISSCC Proceedings*

Source: Bob Broderson, Berkeley Wireless group

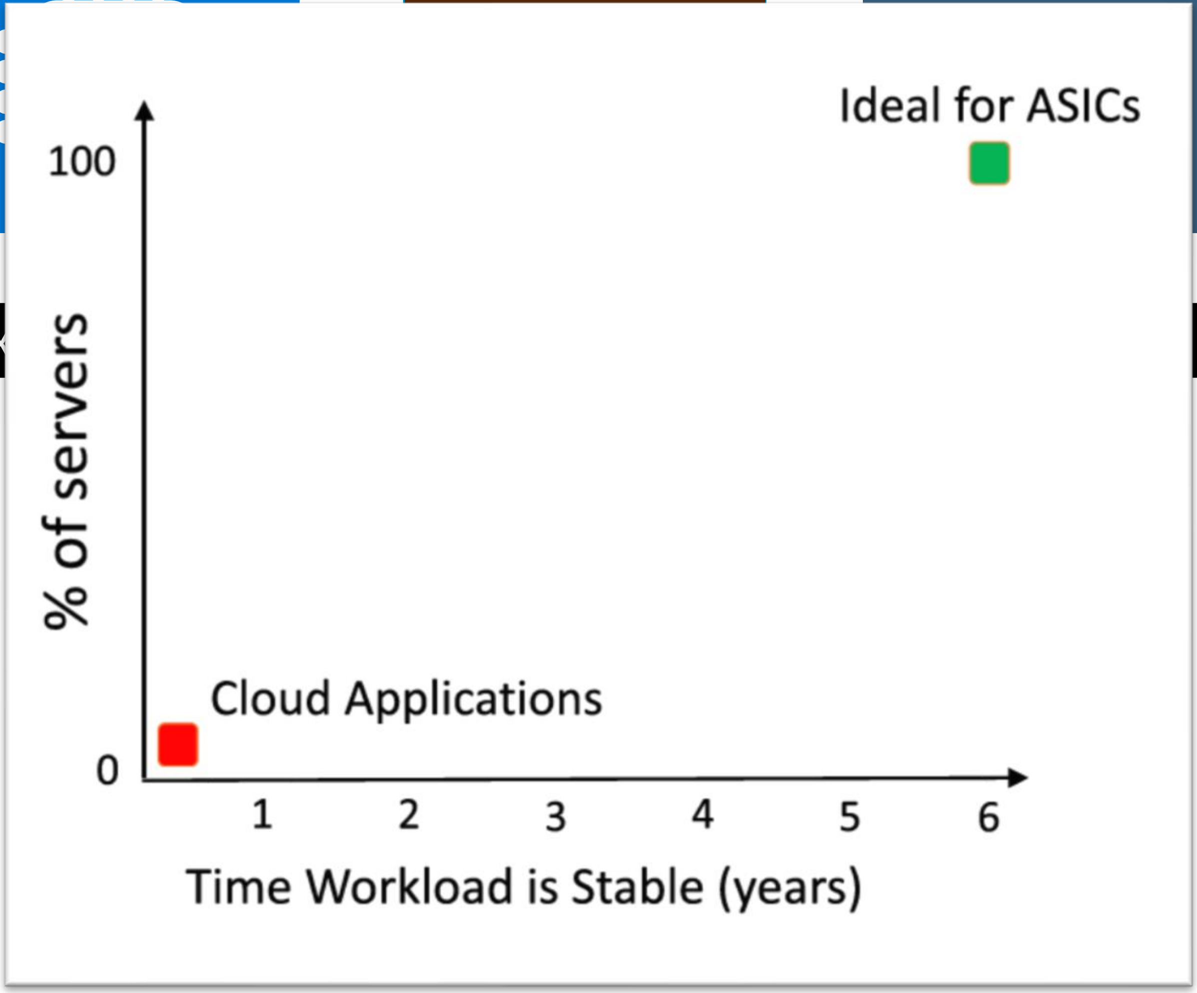
# Microsoft's Heterogeneous Cloud

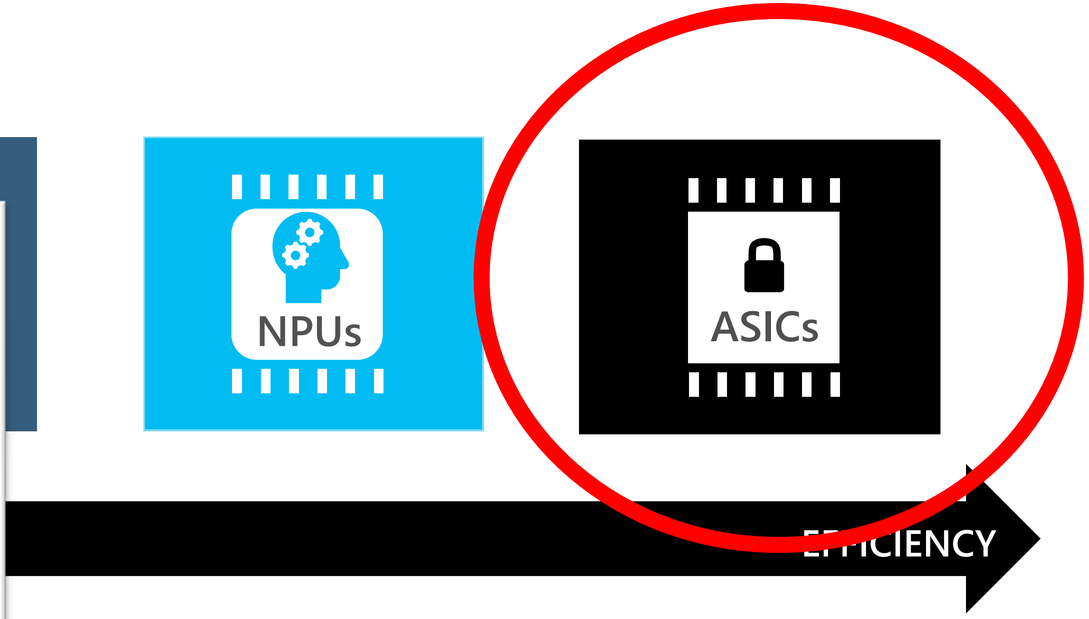


CPUs GPUs FPGAs NPUs ASICs

FLEXIBILITY EFFICIENCY

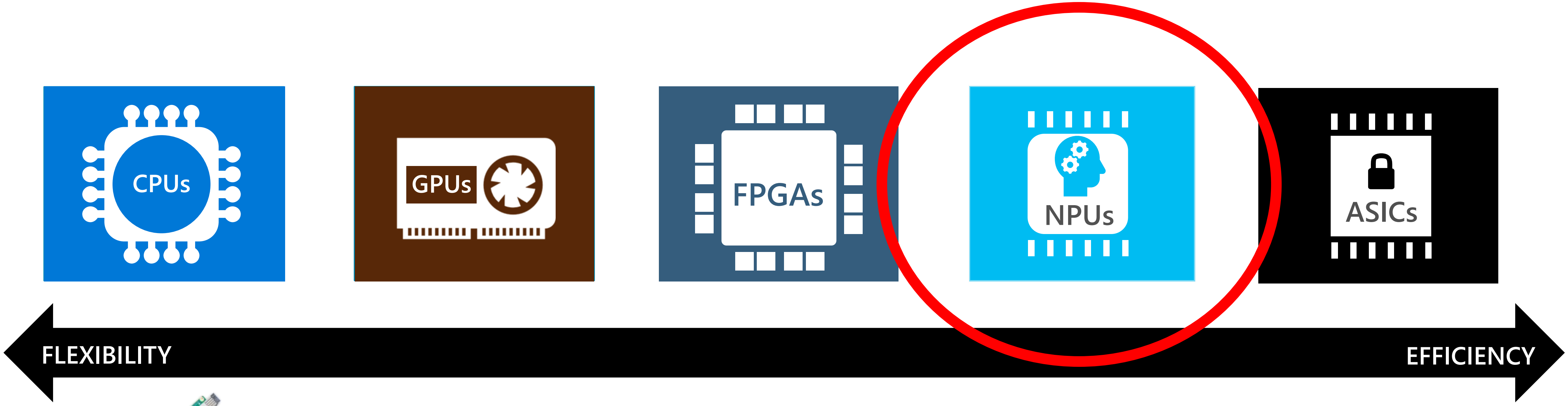
# Microsoft's Heterogeneous Cloud





- Slow time to market
- Unlikely to last 6+ years
- ficult

# Microsoft's Heterogeneous Cloud



CPUs

GPUs

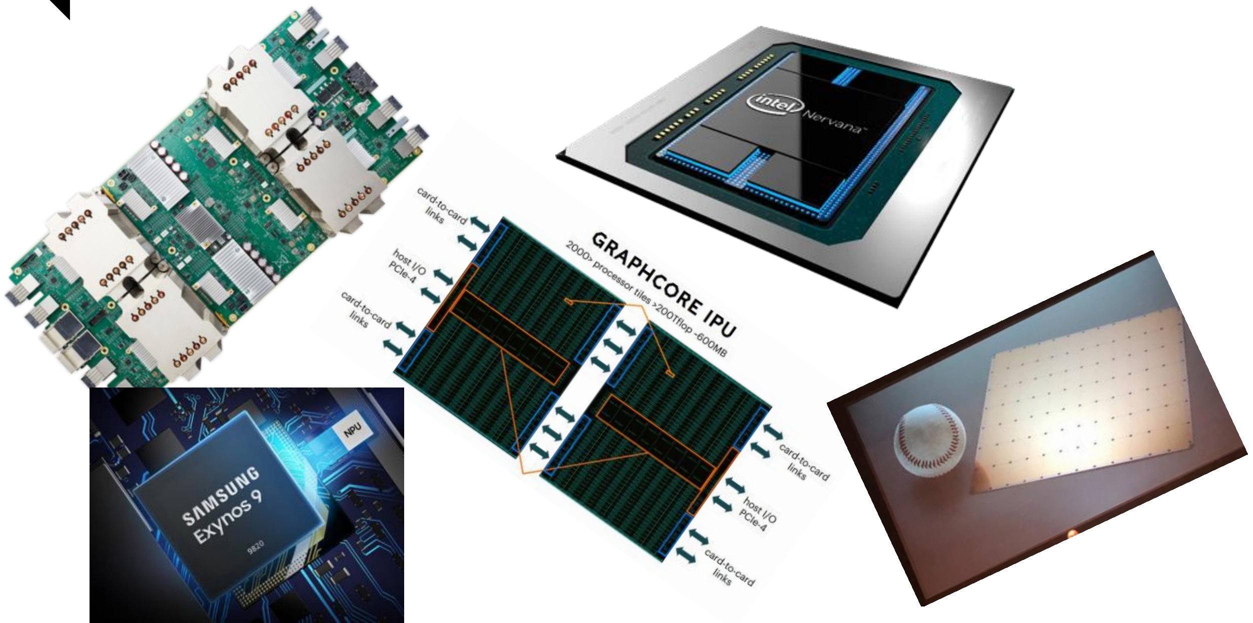
FPGAs

NPUs

ASICs

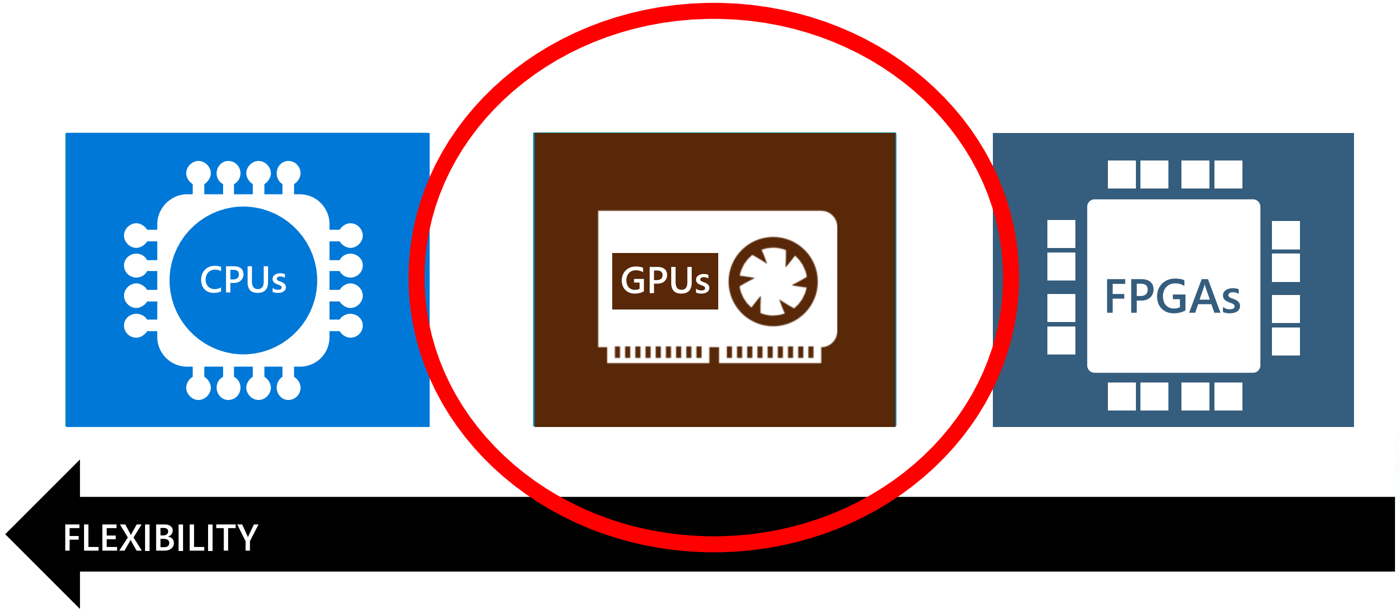
FLEXIBILITY

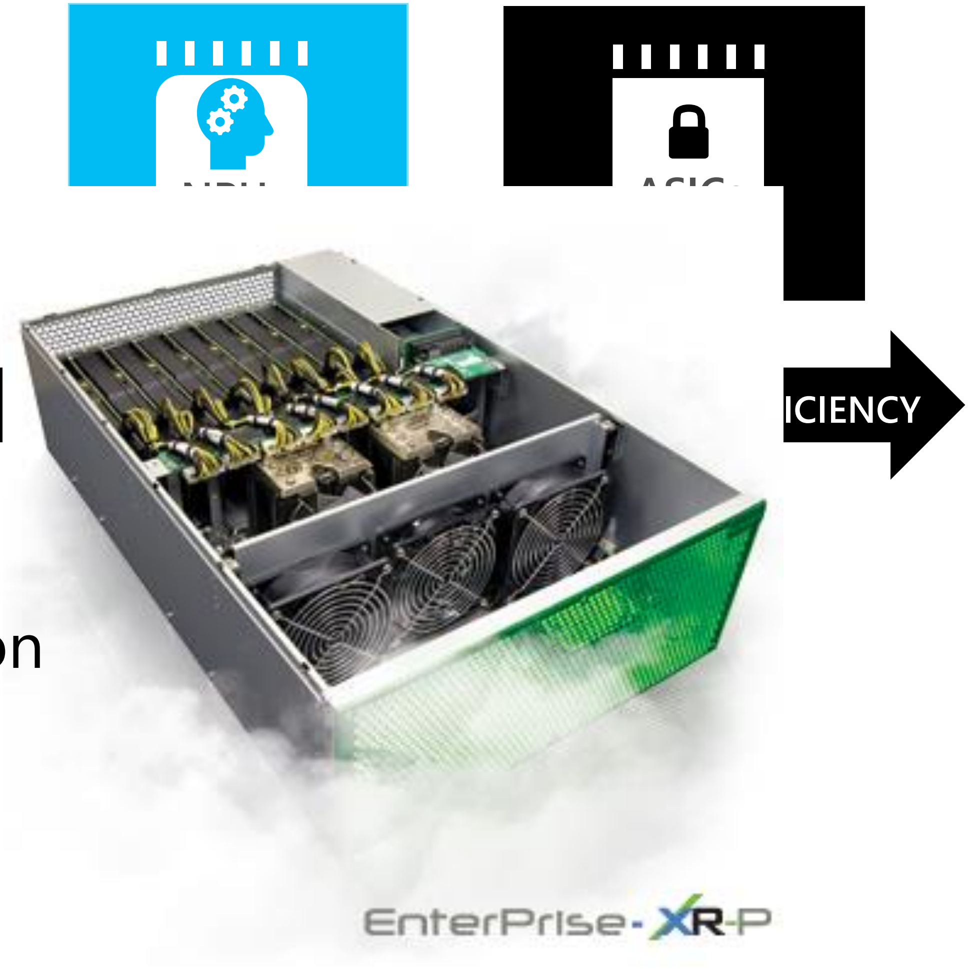
EFFICIENCY



- Very fast moving space
- Immature software ecosystem
- No clear path for development across generations

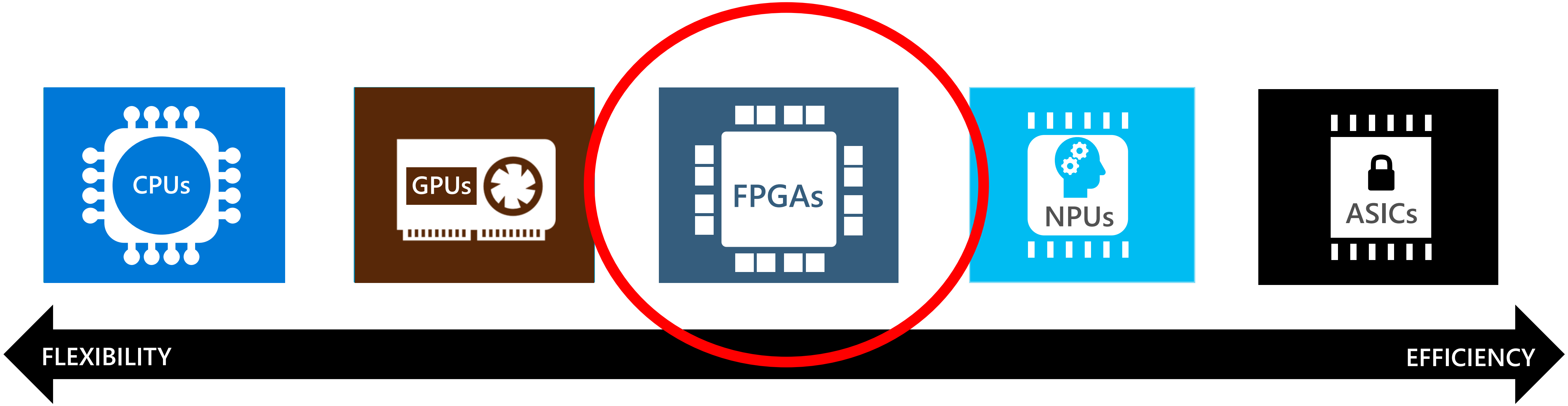
# Microsoft's Heterogeneous Cloud





- Great for HPC on batch data
- Stable software ecosystem allows easy adoption
- Power consumption limits deployment size
- Not ideal for latency-sensitive workloads

# Microsoft’s Heterogeneous Cloud



# What are FPGAs?

**Field Programmable Gate Array**

**FPGAs** are a sea of generic logic and interconnect

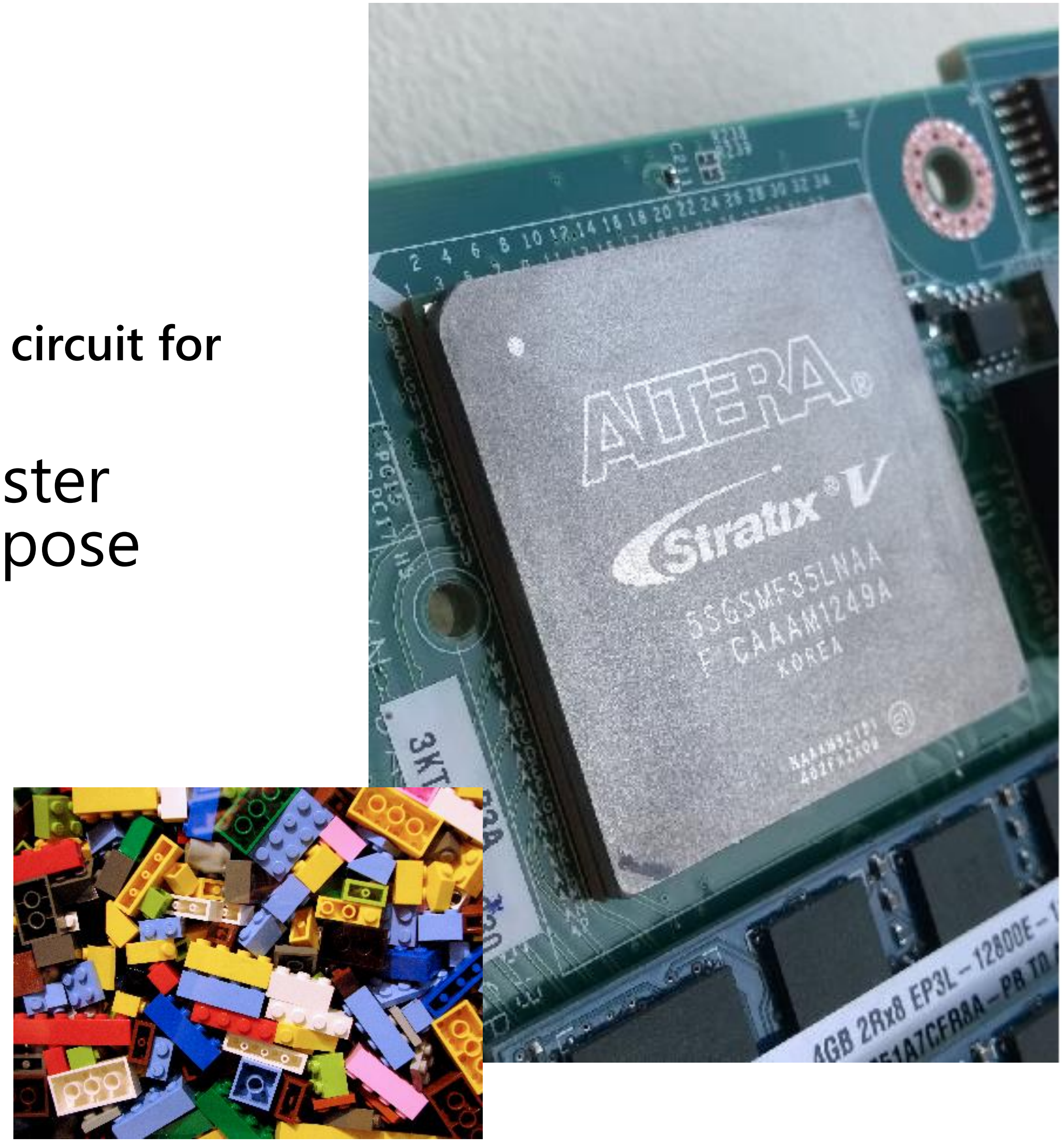
**“Silicon Legos” – build them into exactly the right circuit for each task**

Special-purpose hardware (FPGAs) is faster and more efficient than general-purpose hardware (CPUs)

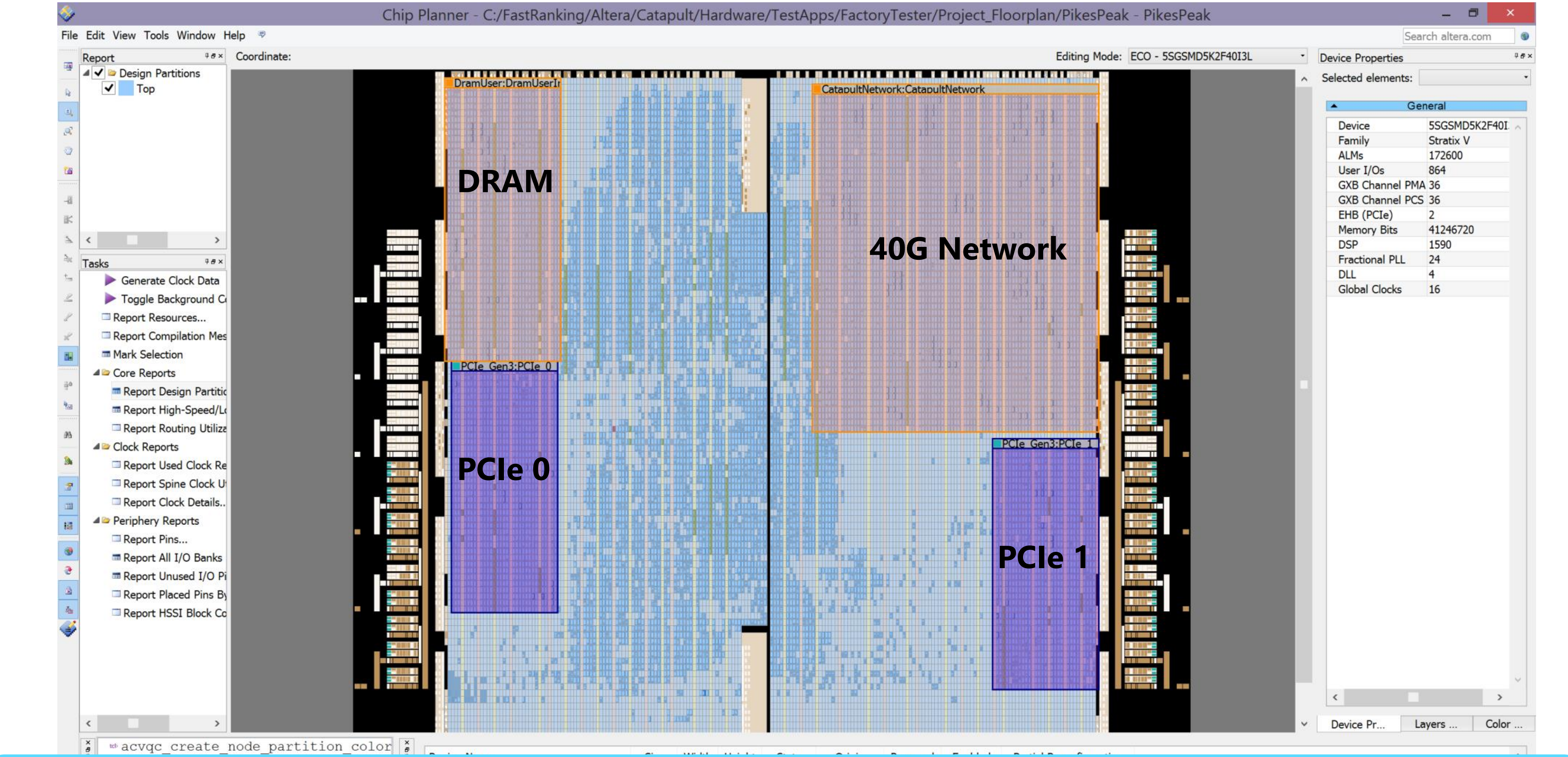
Low power compared to CPU/GPU

**Change the hardware anytime!**

100 ms to 1 second reconfiguration time



# FPGA Physical Layout



Chip Planner - C:/FastRanking/Altera/Catapult/Hardware/TestApps/FactoryTester/Project_Floorplan/PikesPeak - PikesPeak

Editing Mode: ECO - 5SGSMD5K2F40I3L

Device Properties

Selected elements:

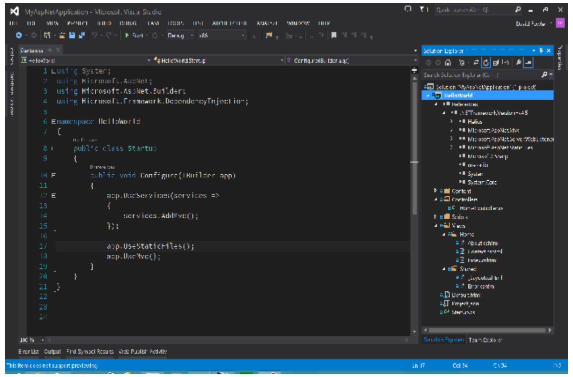
General

| | |
|---|---|
| Device | 5SGSMD5K2F40I |
| Family | Stratix V |
| ALMs | 172600 |
| User I/Os | 864 |
| GXB Channel PMA | 36 |
| GXB Channel PCS | 36 |
| EHB (PCIe) | 2 |
| Memory Bits | 41246720 |
| DSP | 1590 |
| Fractional PLL | 24 |
| DLL | 4 |
| Global Clocks | 16 |

```
acvqc_create_node_partition_color
```

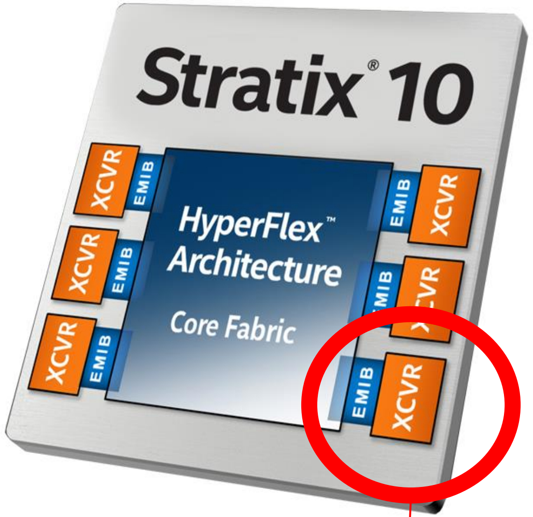
**Customize both the processing logic *and* the I/O**

# Gradual Migration to ASIC

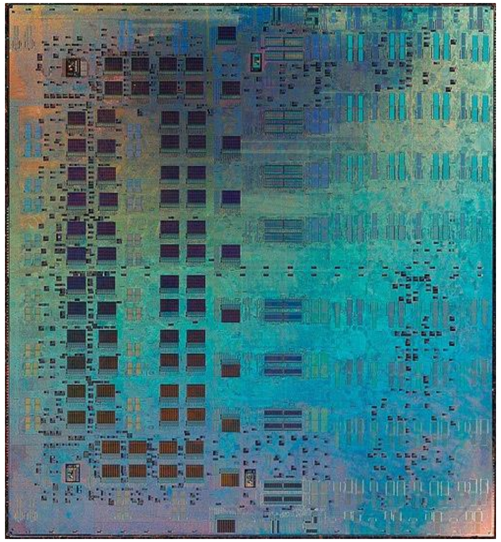


```
using System;
using Microsoft.AspNet;
using Microsoft.AspNet.Builder;
using Microsoft.Framework.DependencyInjection;

namespace HelloWorld
{
    public class Startup
    {
        public void Configure(IBuilder app)
        {
            app.UseServices(services =>
            {
                services.AddMvc();
            });

            app.UseStaticFiles();
            app.UseMvc();
        }
    }
}
```

Software



FPGA

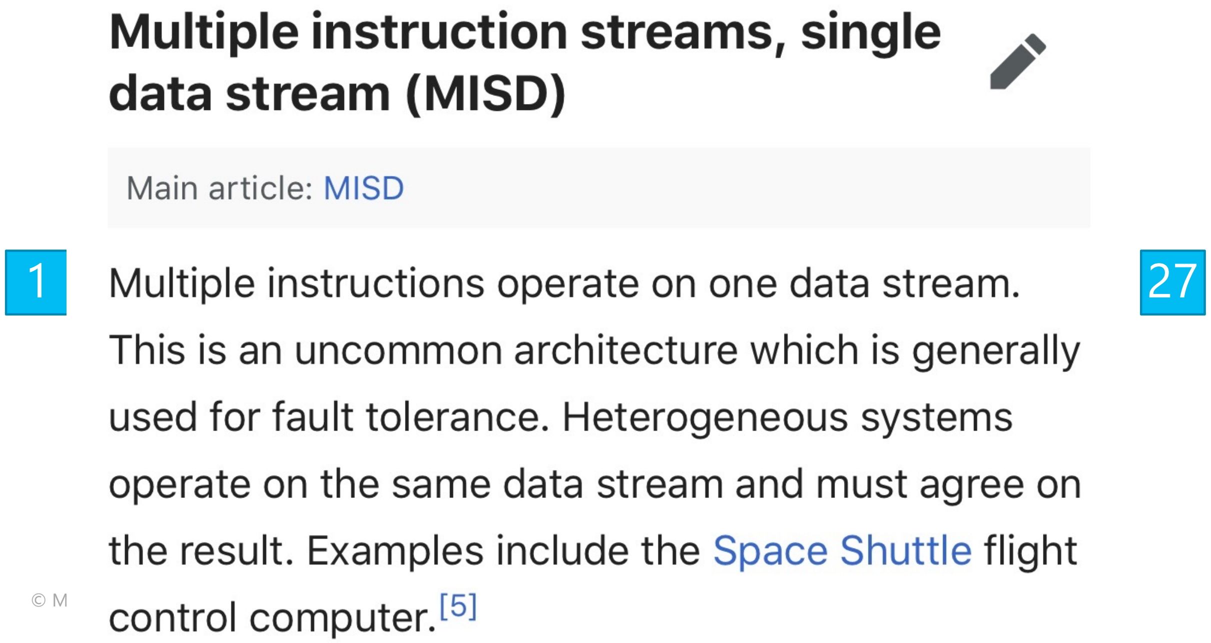


ASIC

ASIC Integration via Chiplets

# Why is the FPGA a good choice as an accelerator?

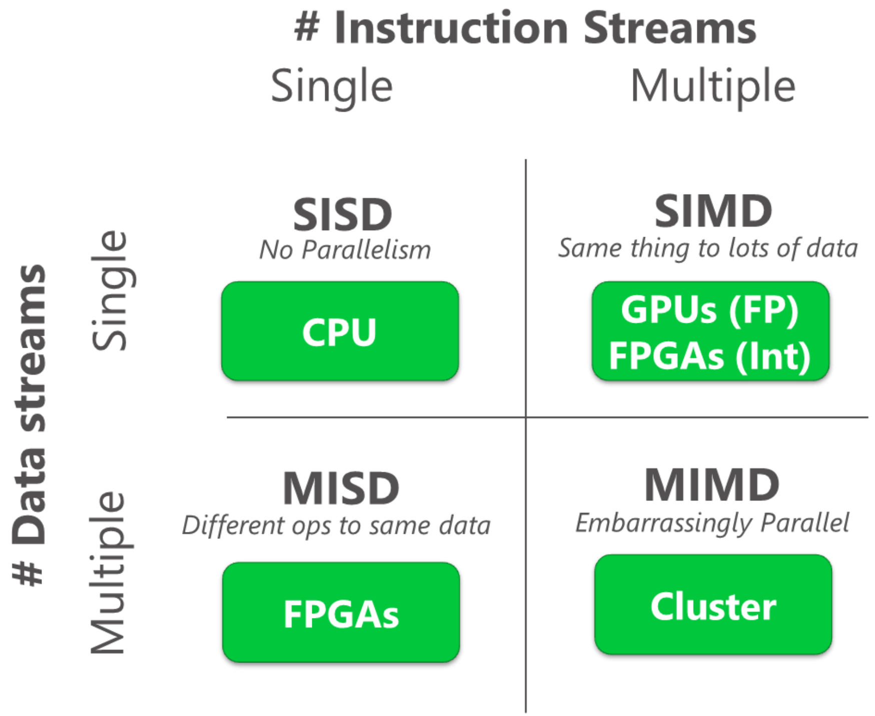
- **Greater Performance and Efficiency than CPU, more general purpose than ASIC**
- **Many applications aren’t about throughput or double-precision floating point**
  - AI/ML, Bioinformatics, text processing, financial services…
- **Exploits different forms of parallelism than other accelerators**



## Multiple instruction streams, single data stream (MISD)

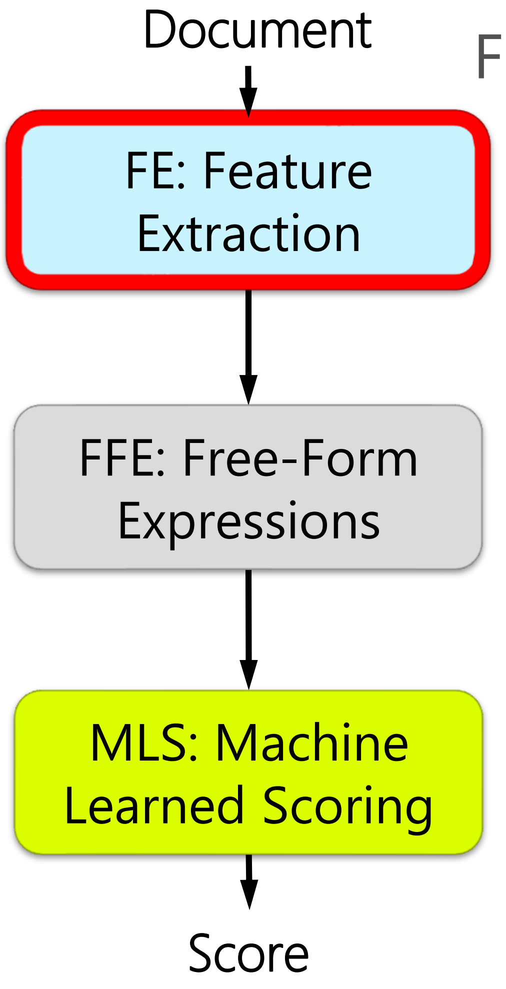
Main article: MISD

1 Multiple instructions operate on one data stream. This is an uncommon architecture which is generally used for fault tolerance. Heterogeneous systems operate on the same data stream and must agree on the result. Examples include the Space Shuttle flight control computer.[5] 27



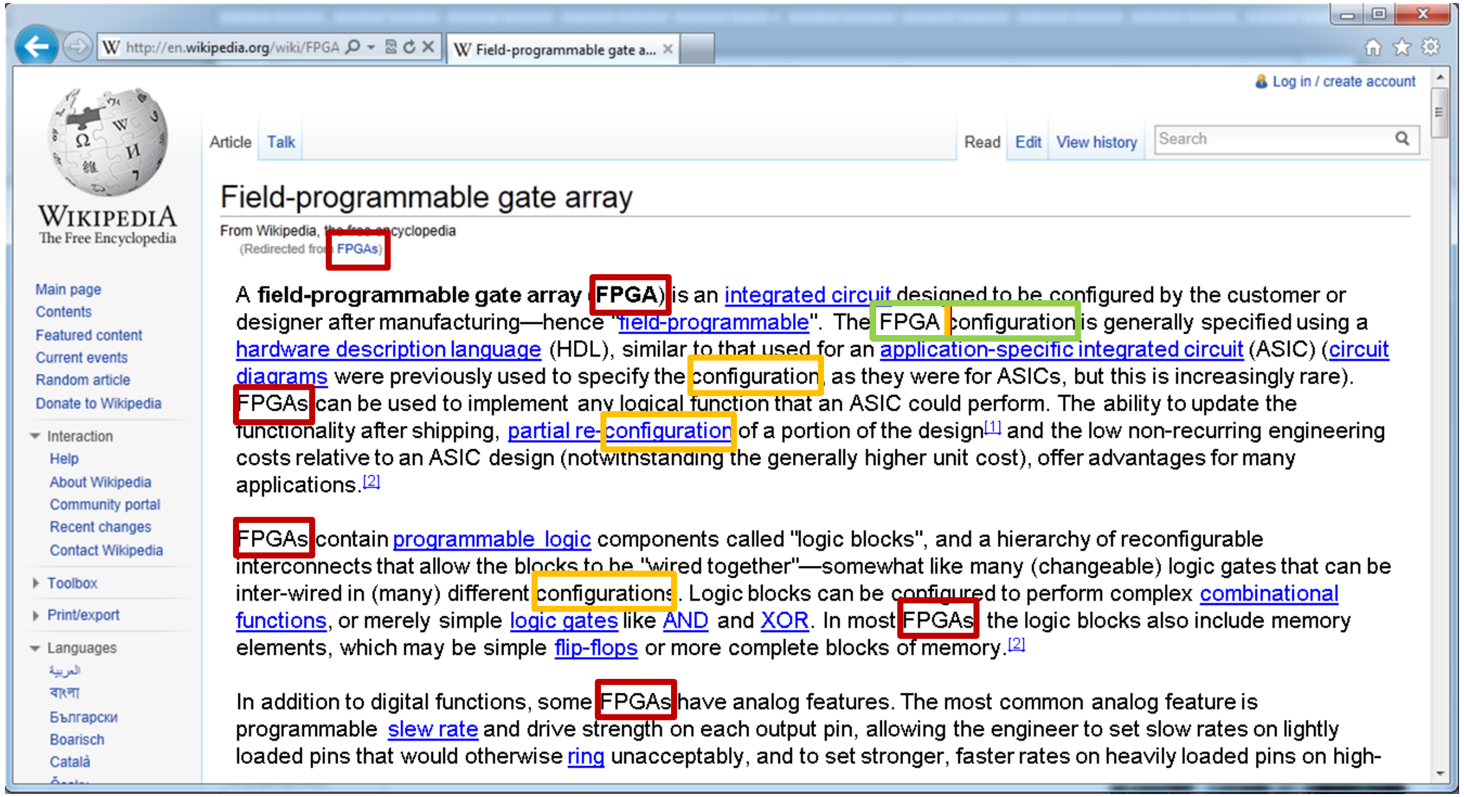
| # Data streams | # Instruction Streams: Single | # Instruction Streams: Multiple |
|---|---|---|
| Single | **SISD**<br>*No Parallelism*<br>CPU | **SIMD**<br>*Same thing to lots of data*<br>GPUs (FP)<br>FPGAs (Int) |
| Multiple | **MISD**<br>*Different ops to same data*<br>FPGAs | **MIMD**<br>*Embarrassingly Parallel*<br>Cluster |

# FE: Feature Extraction



**Query: “FPGA Configuration”**

Features: **NumberOfOccurrences_0 = 7** **NumberOfOccurrences_1 = 4** **NumberOfTuples_0_1 = 1**



## Field-programmable gate array

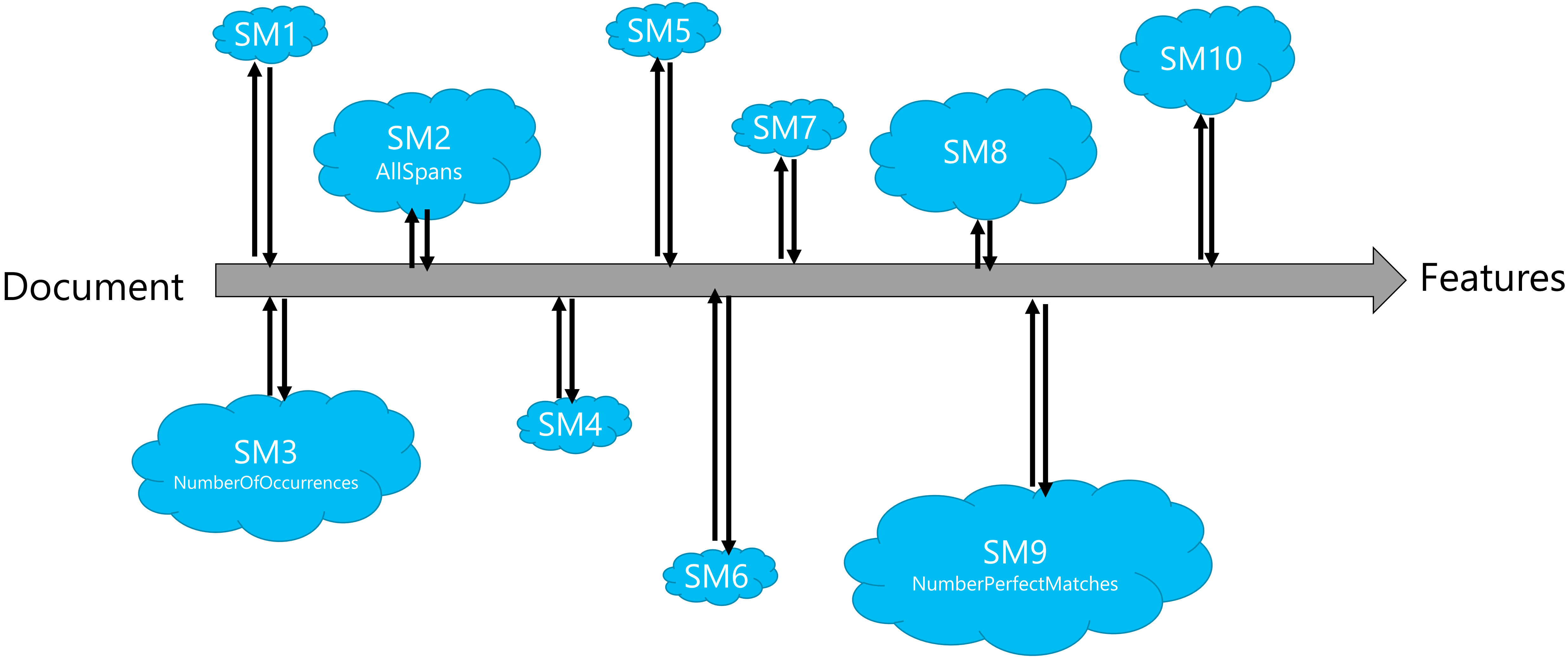
From Wikipedia, the free encyclopedia
(Redirected from FPGAs)

A **field-programmable gate array** (**FPGA**) is an integrated circuit designed to be configured by the customer or designer after manufacturing—hence "field-programmable". The FPGA configuration is generally specified using a hardware description language (HDL), similar to that used for an application-specific integrated circuit (ASIC) (circuit diagrams were previously used to specify the configuration, as they were for ASICs, but this is increasingly rare). FPGAs can be used to implement any logical function that an ASIC could perform. The ability to update the functionality after shipping, partial re-configuration of a portion of the design[1] and the low non-recurring engineering costs relative to an ASIC design (notwithstanding the generally higher unit cost), offer advantages for many applications.[2]

FPGAs contain programmable logic components called "logic blocks", and a hierarchy of reconfigurable interconnects that allow the blocks to be "wired together"—somewhat like many (changeable) logic gates that can be inter-wired in (many) different configurations. Logic blocks can be configured to perform complex combinational functions, or merely simple logic gates like AND and XOR. In most FPGAs the logic blocks also include memory elements, which may be simple flip-flops or more complete blocks of memory.[2]

In addition to digital functions, some FPGAs have analog features. The most common analog feature is programmable slew rate and drive strength on each output pin, allowing the engineer to set slow rates on lightly loaded pins that would otherwise ring unacceptably, and to set stronger, faster rates on heavily loaded pins on high-

# Feature Extraction Accelerator



# FPGAs in Physics Applications

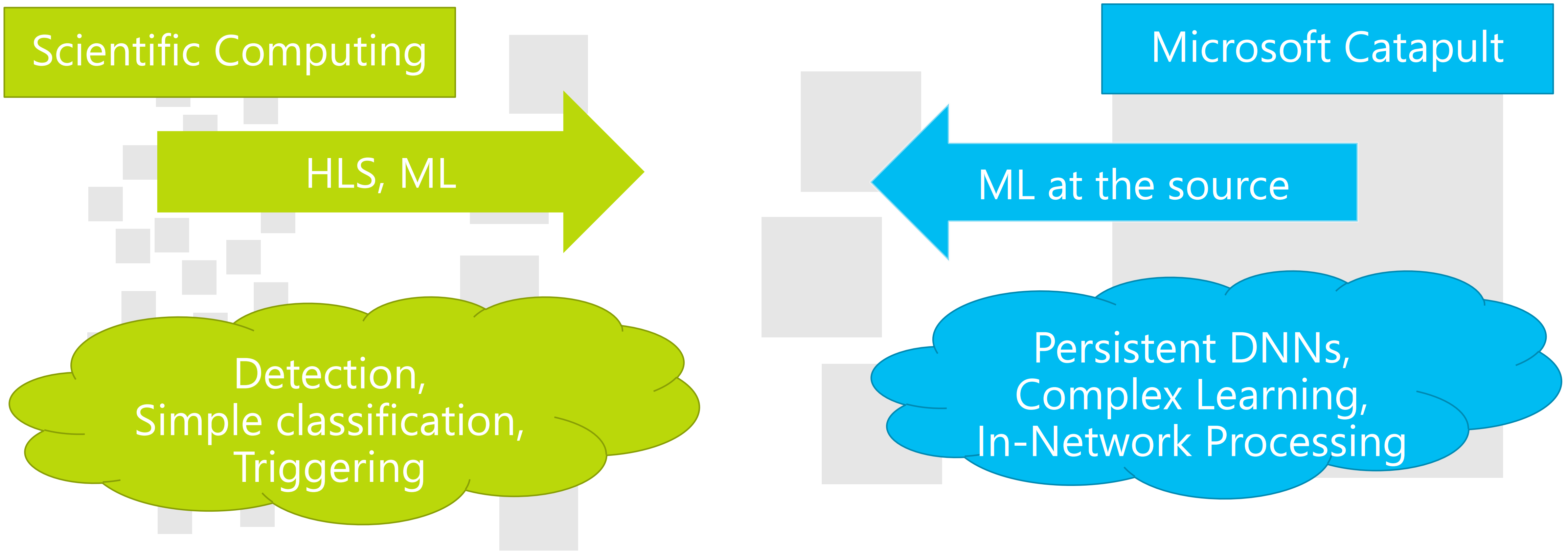
SETI

# FPGAs in Cosmology

EOR Science can be done with a paperclip and a supercomputer
-- Don C. Backer

Cosmologists often refer to their telescopes as "software telescopes"

# Processing Pipeline



Scientific Computing

HLS, ML

Detection,
Simple classification,
Triggering

Microsoft Catapult

ML at the source

Persistent DNNs,
Complex Learning,
In-Network Processing

| | | | | |
|---|---|---|---|---|
| Cloud: | IoT | Edge | Fog | Cloud |
| HPC: | Sensors | Detectors | Clusters | Supercomputer |

# Catapult: Long, Fruitful FPGA Investment

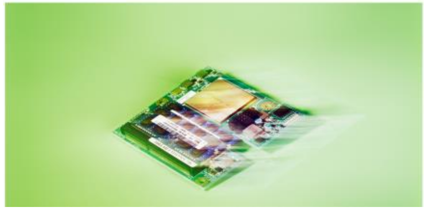
**Pre-History:**
May 2009: Bing Launched
Feb 2010: Azure Launched
Dec 2010: Catapult concept

**Catapult v1: Mt Granite**
Distributed solution
Integrated with WCS (OCP) 1.0

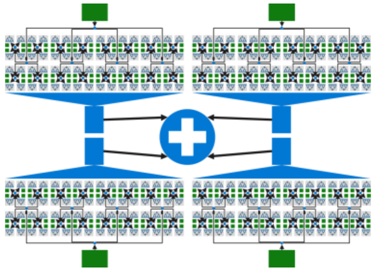
**v2: Pikes Peak**
Integrated Bing + Azure design
Bump-in-the-wire introduced

**Azure AccelNet Unveiled**
Azure production launch
AI Supercomputer demo

MICROSOFT BETS ITS FUTURE ON A REPROGRAMMABLE COMPUTER CHIP



**Project BrainWave / Storm Peak**
Real-Time AI
First 3rd Party FPGA Service



… 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 …

**v0: Research POC**
Built v0 board w/6 Xilinx FPGAs
30k lines of Bing code on FPGA

**v1: Scale Pilot**
1632 servers deployed
Bing IndexServe accelerated

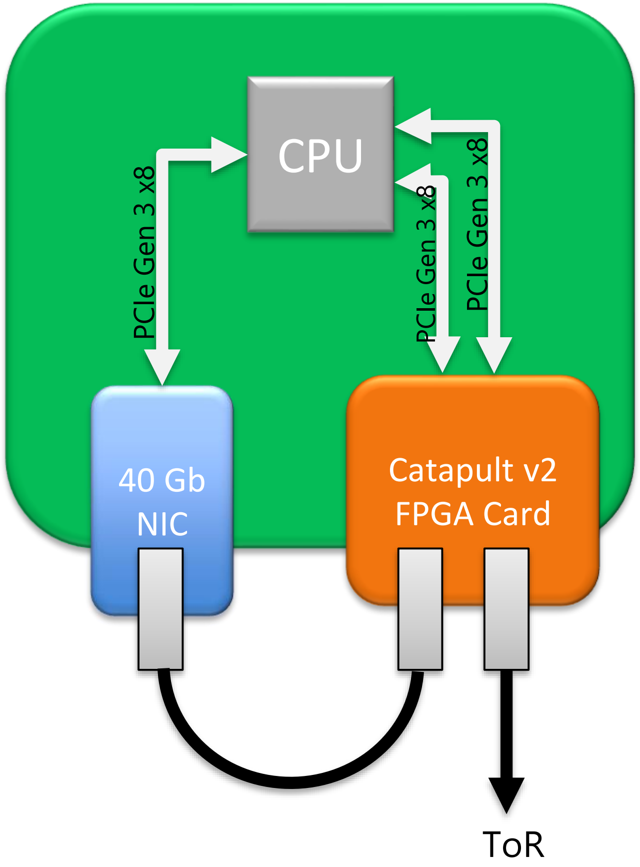
MICROSOFT SUPERCHARGES BING SEARCH WITH PROGRAMMABLE CHIPS

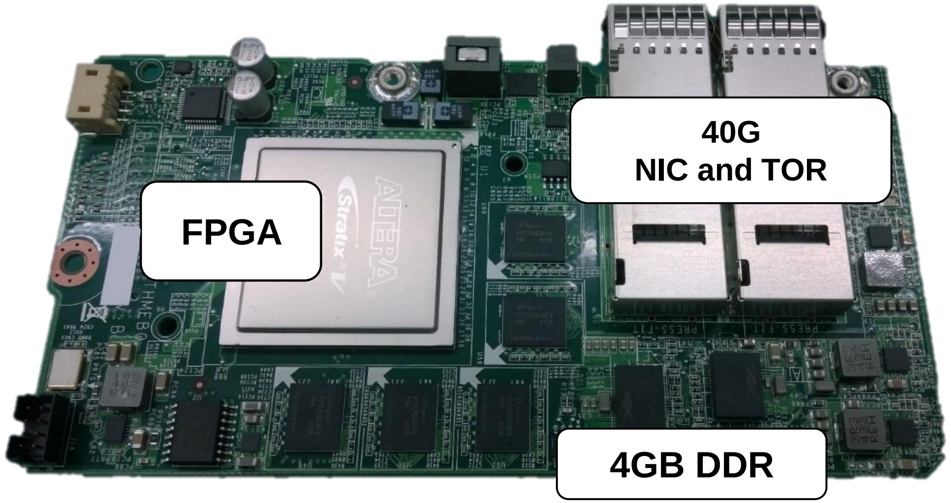
**v2 Production and ramp**
FPGAs reach production
Deployed in all new servers

**Catapult v3: Longs Peak**
DNN Platform for Bing
50Gb w/ integrated NIC

**Azure Databox Edge**
On-Site Inference

# Catapult v2 – Bump in the Wire

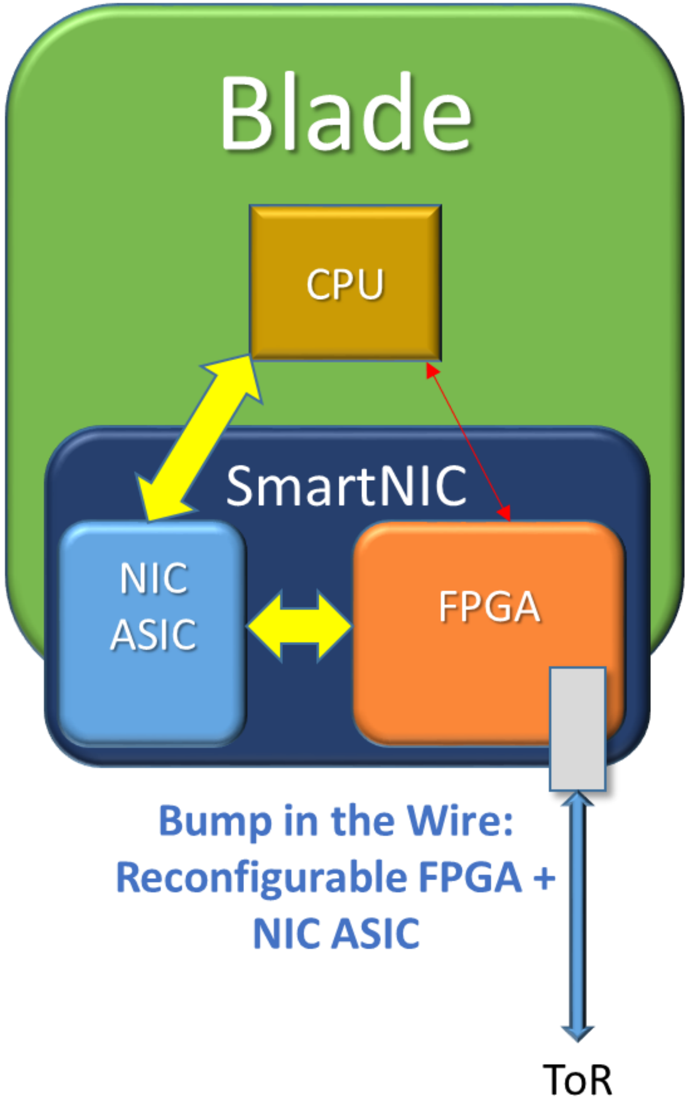




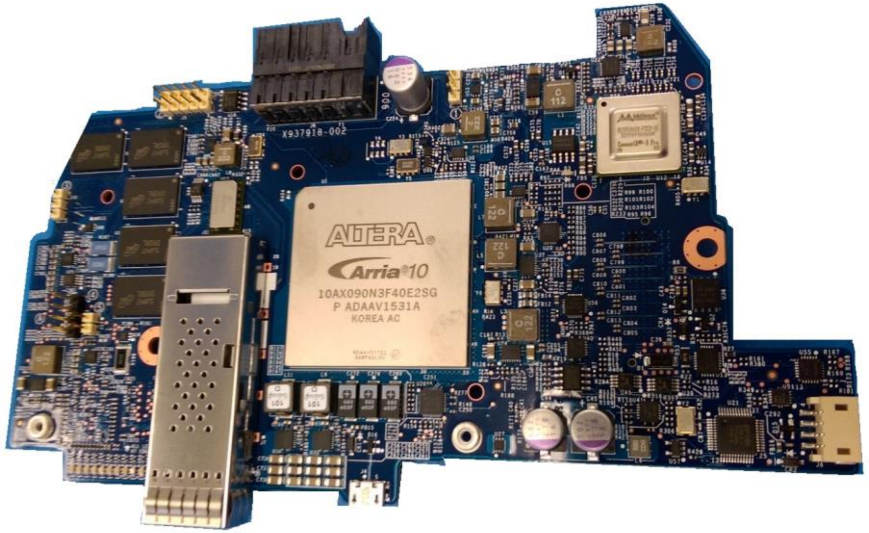
Pikes Peak

Storey Peak

# Catapult v3 – Converged Boards



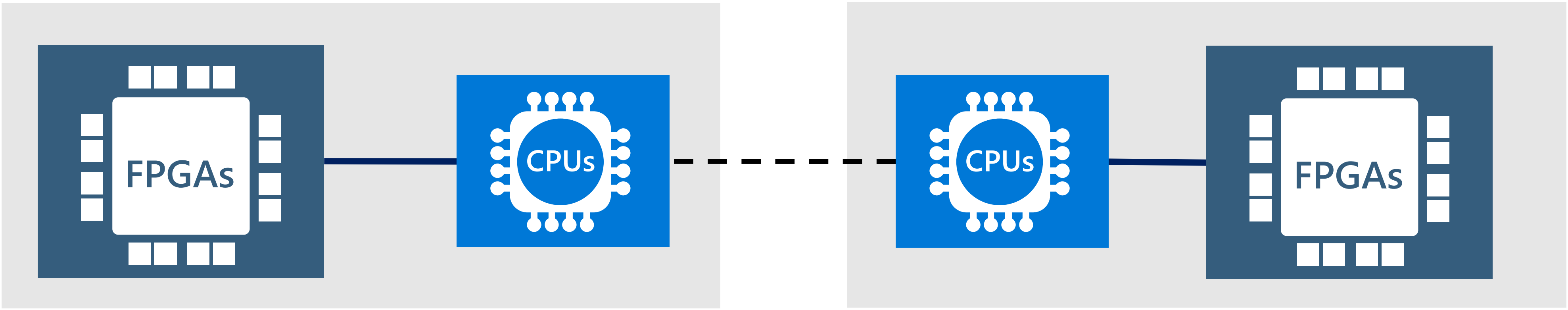
Bump in the Wire: Reconfigurable FPGA + NIC ASIC



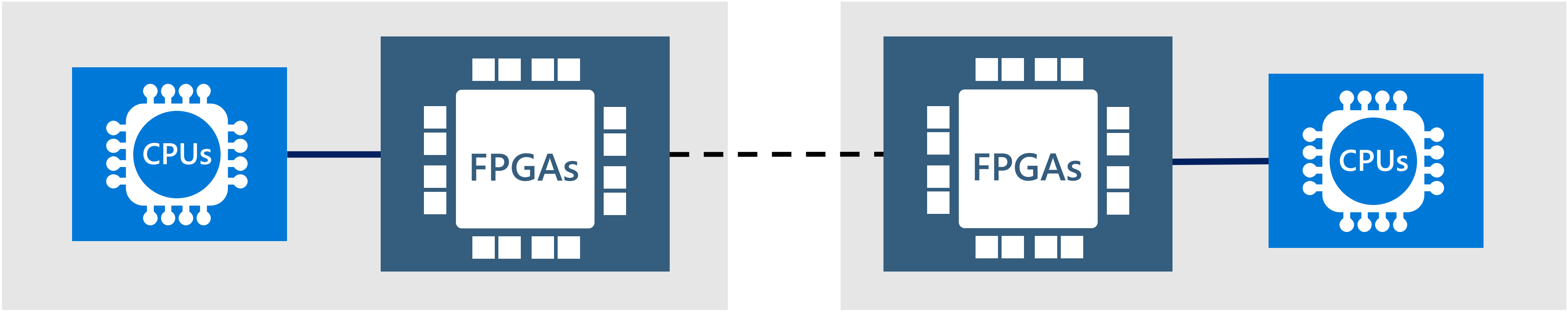
Dragontail Peak

Longs Peak

# Accelerator Integration

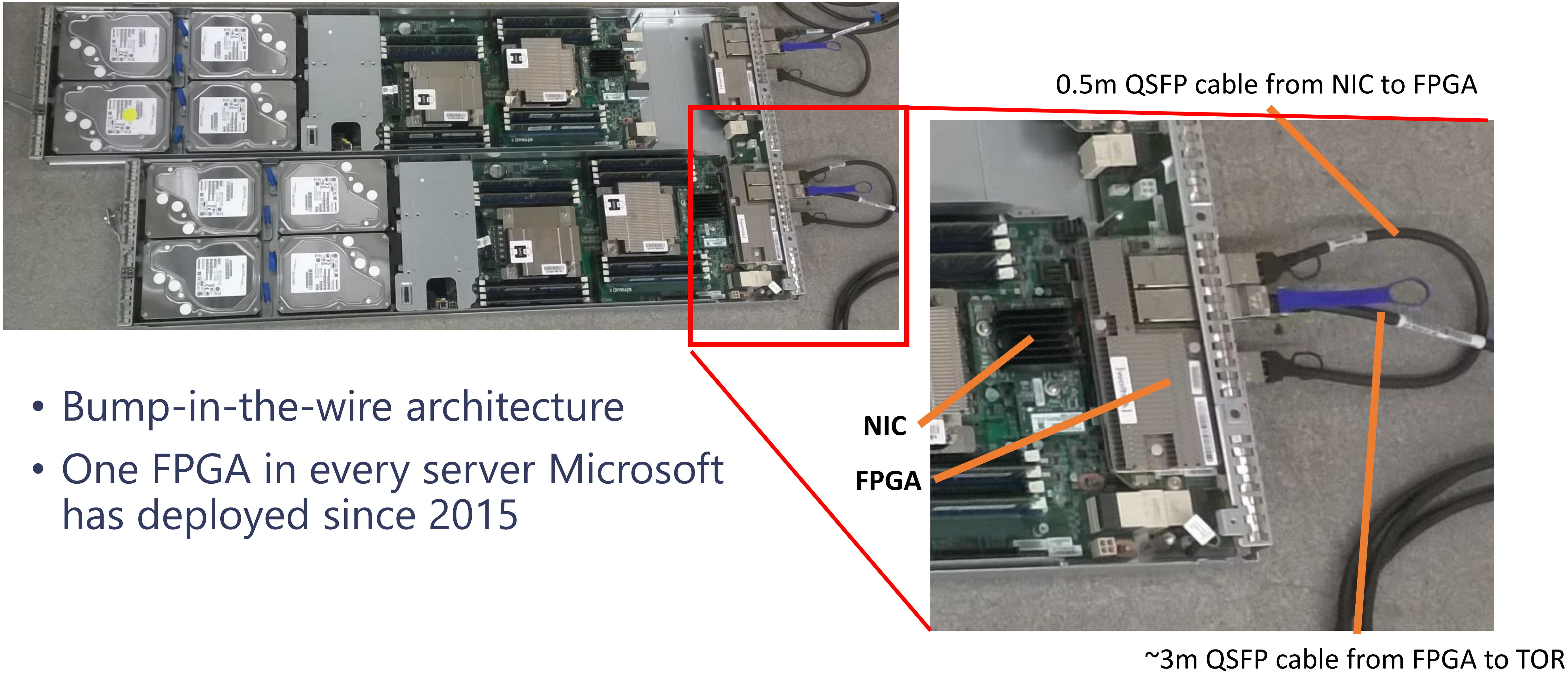


Traditional Integration



Bump in the Wire

# Basic Catapult Architecture



0.5m QSFP cable from NIC to FPGA

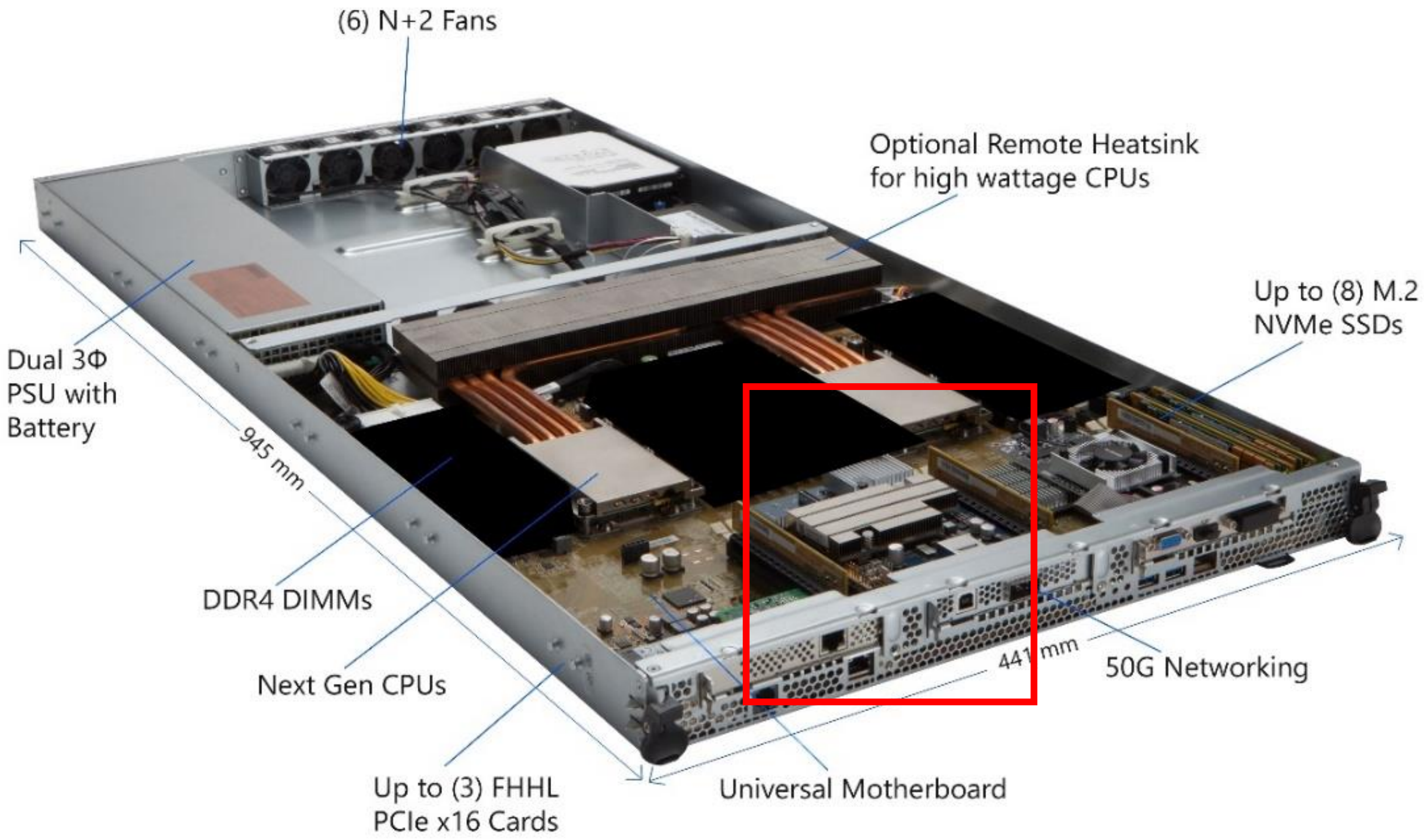
NIC

FPGA

~3m QSFP cable from FPGA to TOR

- Bump-in-the-wire architecture
- One FPGA in every server Microsoft has deployed since 2015

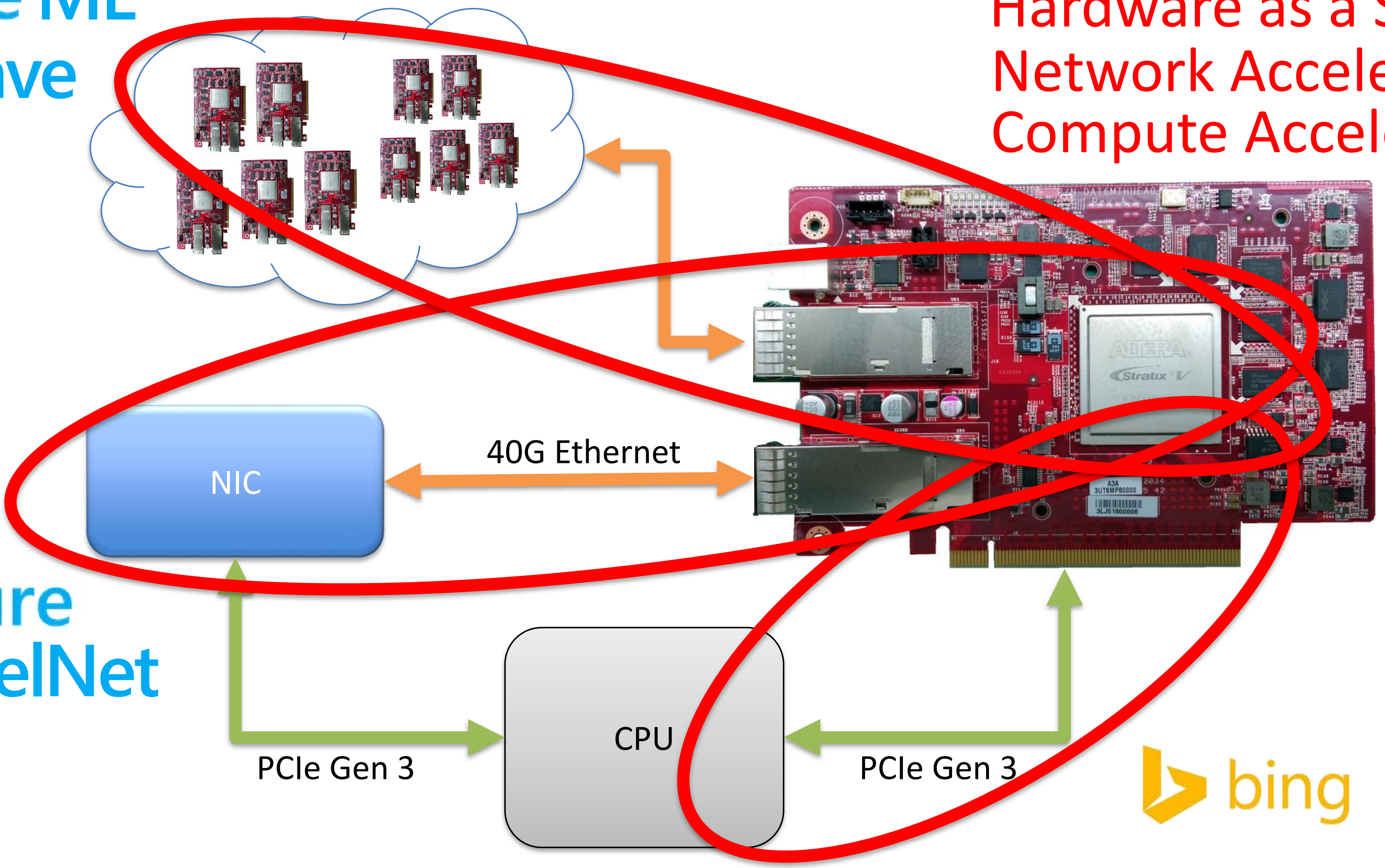
# New Generation Server Chassis



# Bump-in-the-wire Architectue

Azure ML
BrainWave

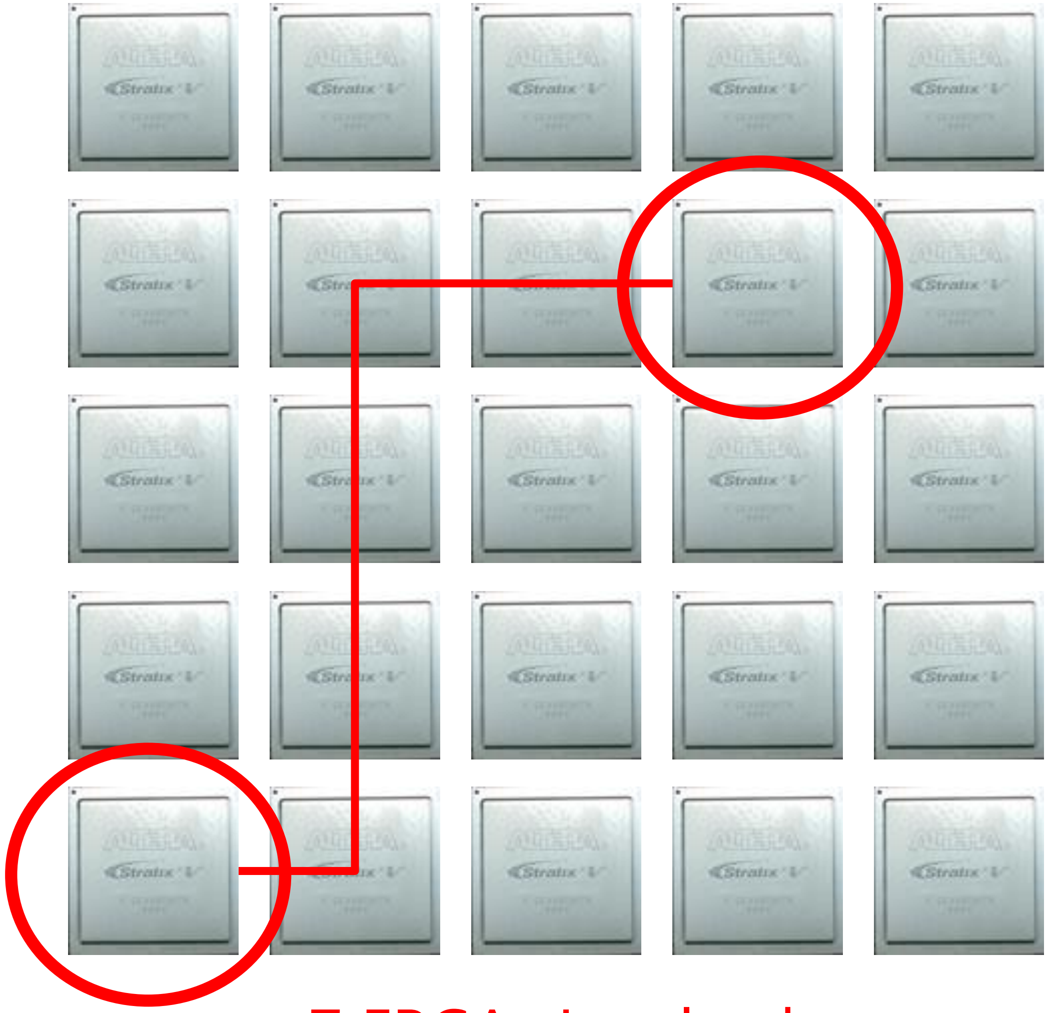
Hardware as a Service
Network Acceleration
Compute Acceleration



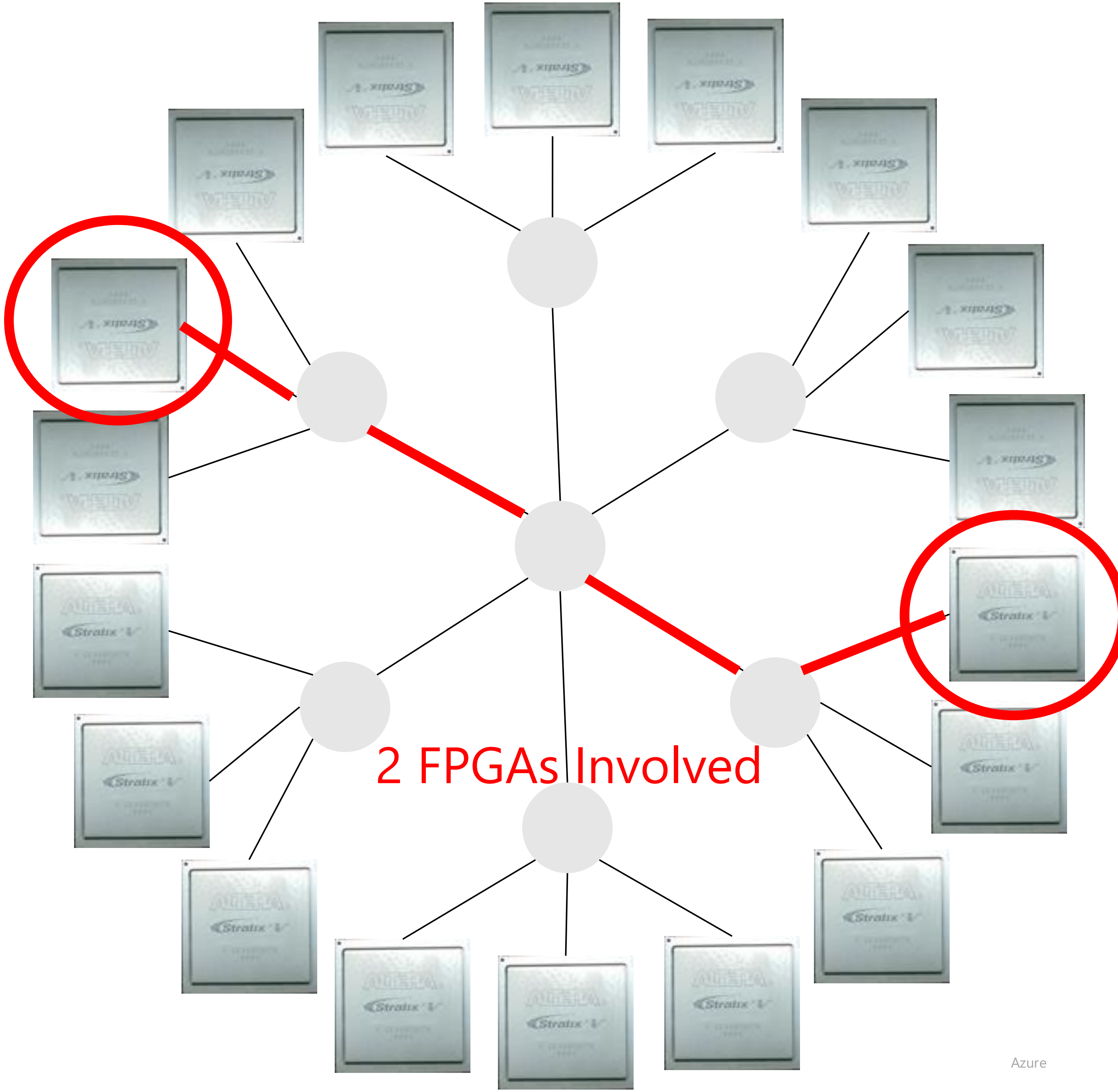
Azure
AccelNet

bing

# Global-Scale FPGA

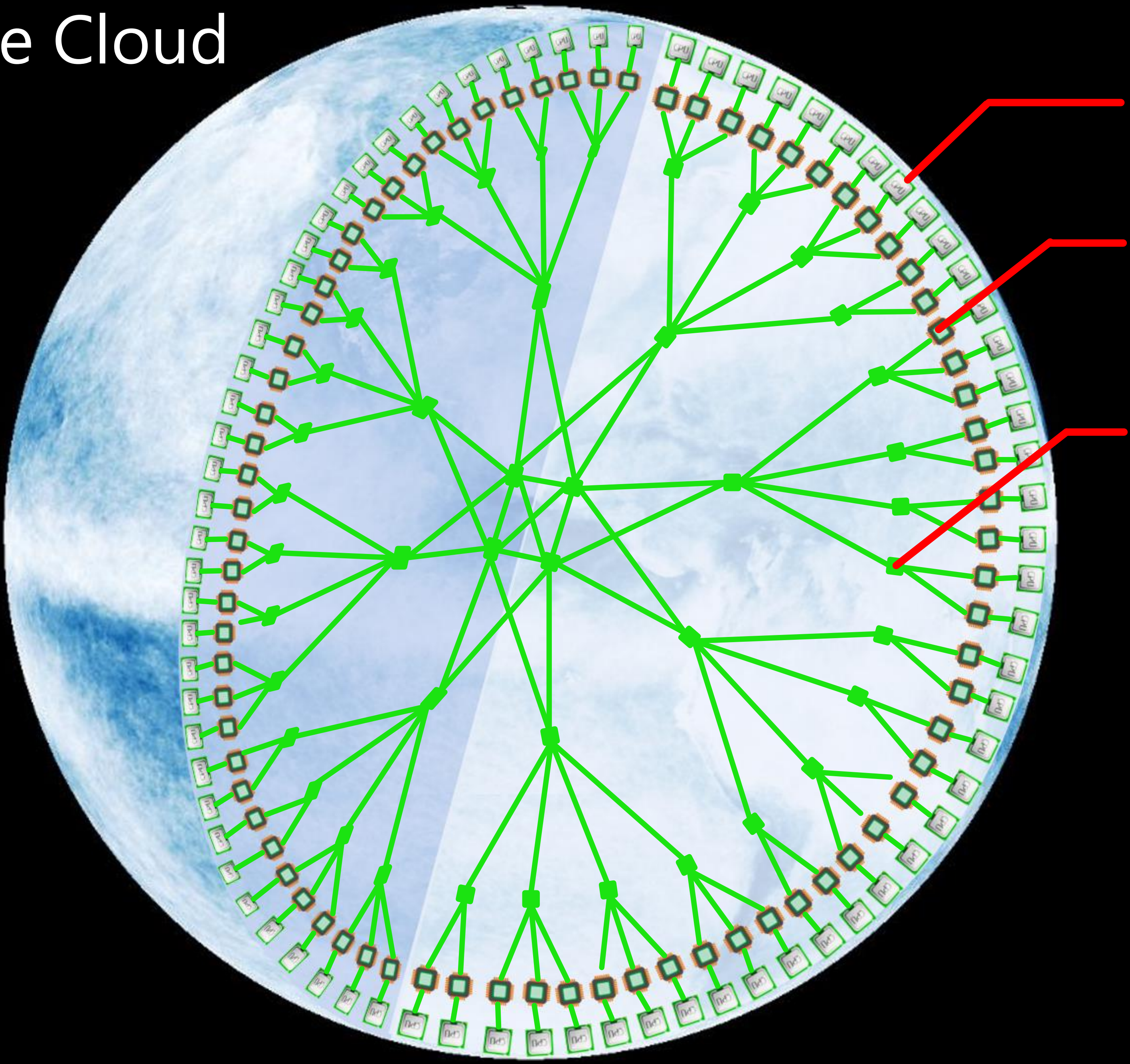


7 FPGAs Involved



2 FPGAs Involved

# Configurable Cloud

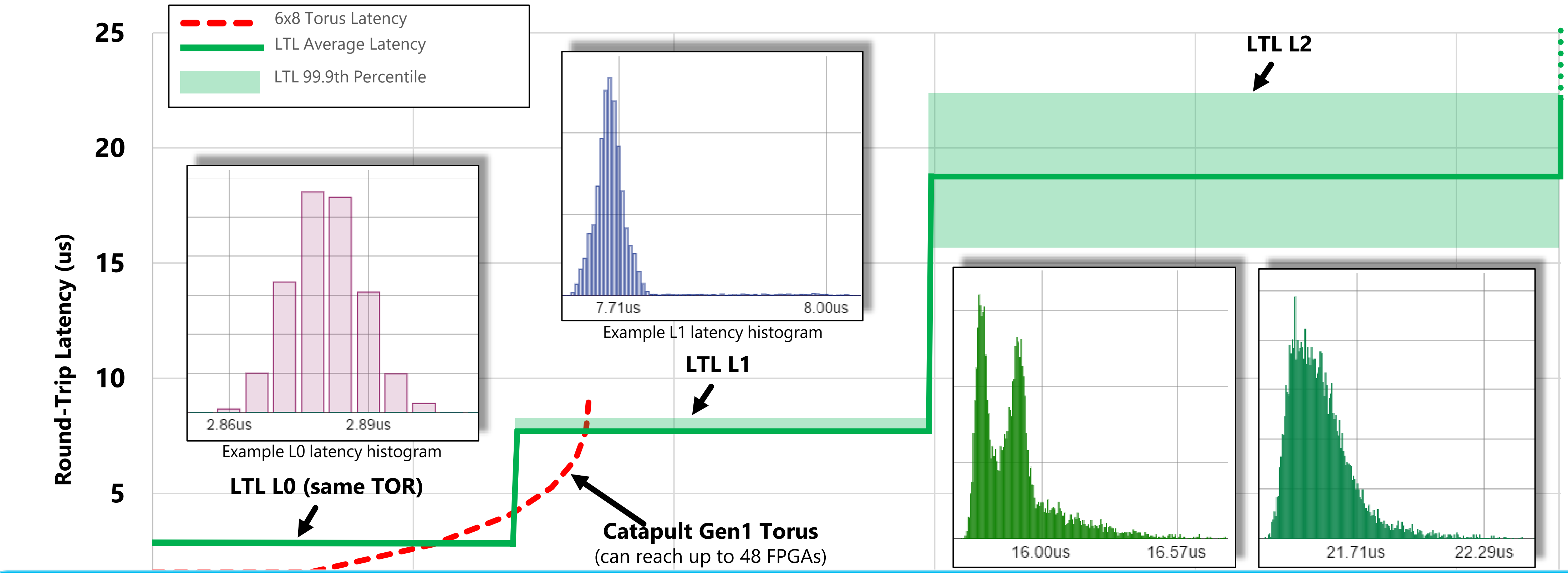


CPU compute layer

Reconfigurable compute layer

Converged network

# Network Latencies



- Extremely low latency (Similar to Infiniband)
- Global-scale FPGA

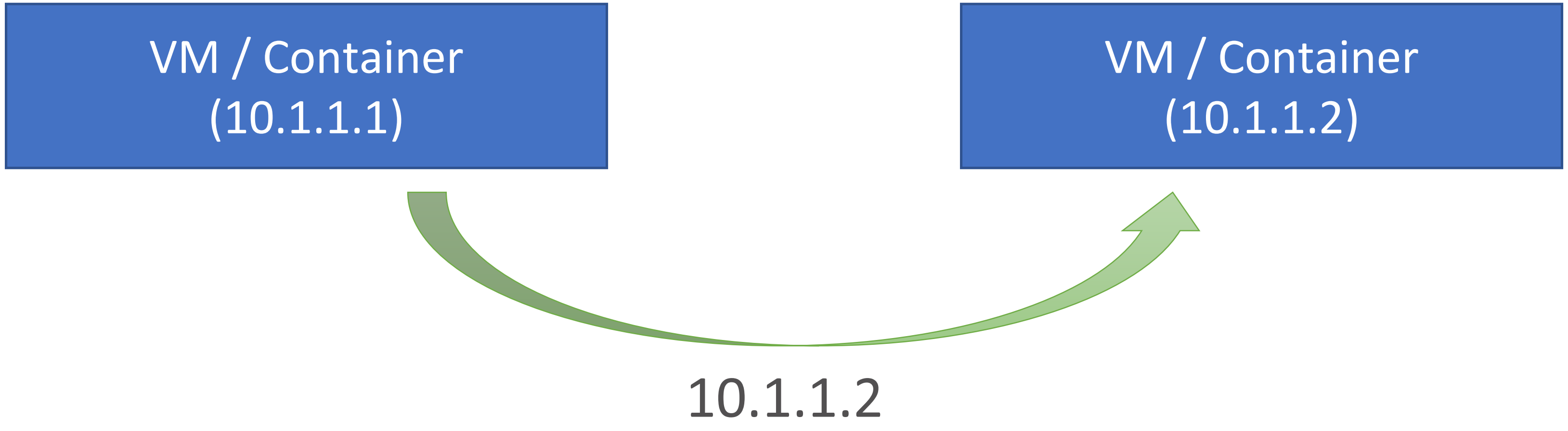
# HPC with the Cloud?

- The idea *sounds* great
- Pay for compute only when you use it
- When it breaks, it’s someone else’s problem
- No need to call the realtor / utility company when you want a bigger machine
- New hardware just shows up. No retrofits needed.

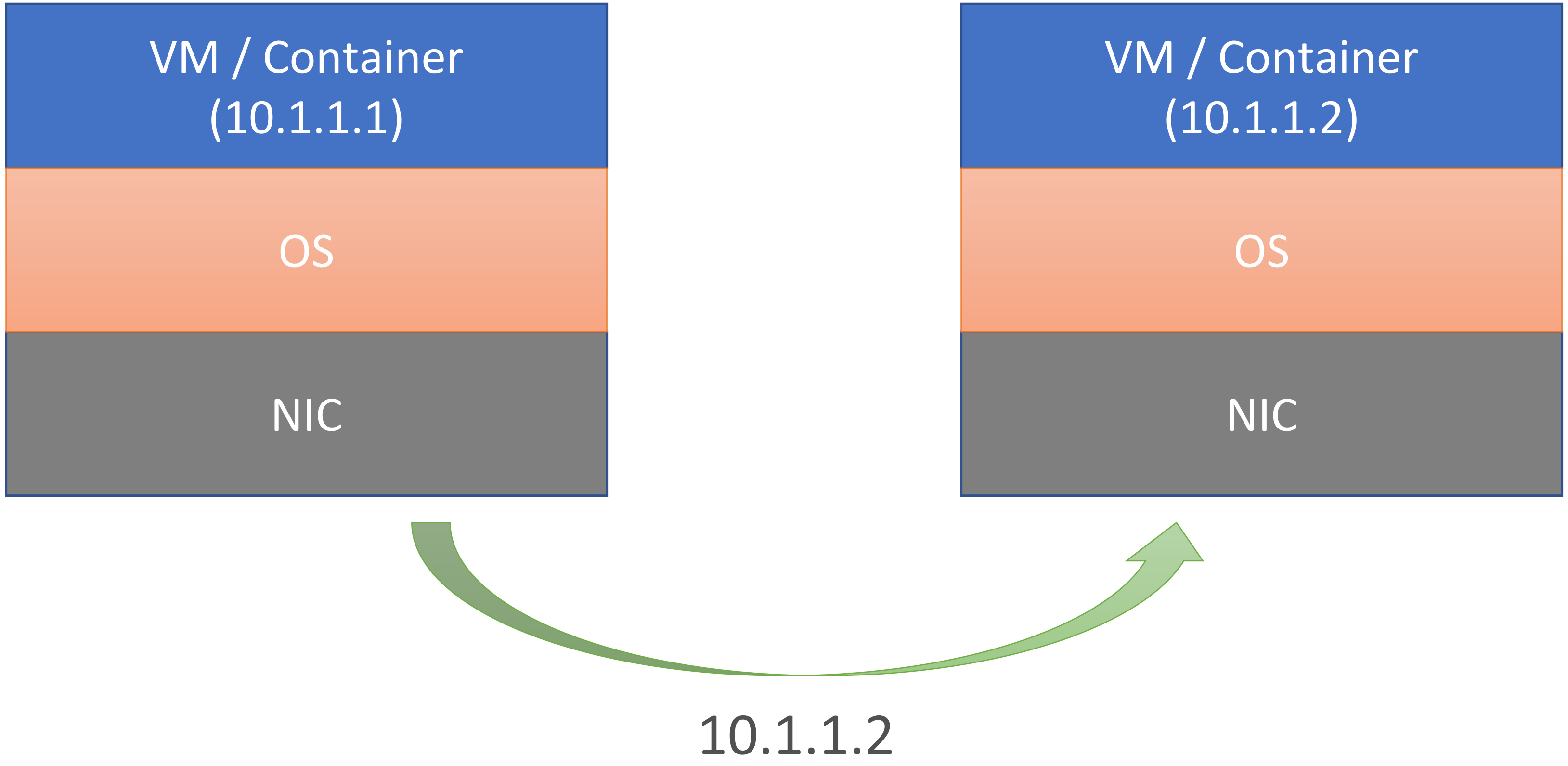
# Why hasn’t Supercomputing moved to the Cloud?

- ❑ CPUs look largely the same
- ❑ Networks are highly specialized, tuned for low-latency, high bandwidth
  - ❑ Proved this works FPGA-to-FPGA, but what about software (CPU-to-CPU)?
- ❑ Supercomputers include specialized accelerators (especially GPUs)
- ❑ Won’t running virtual machines kill performance?

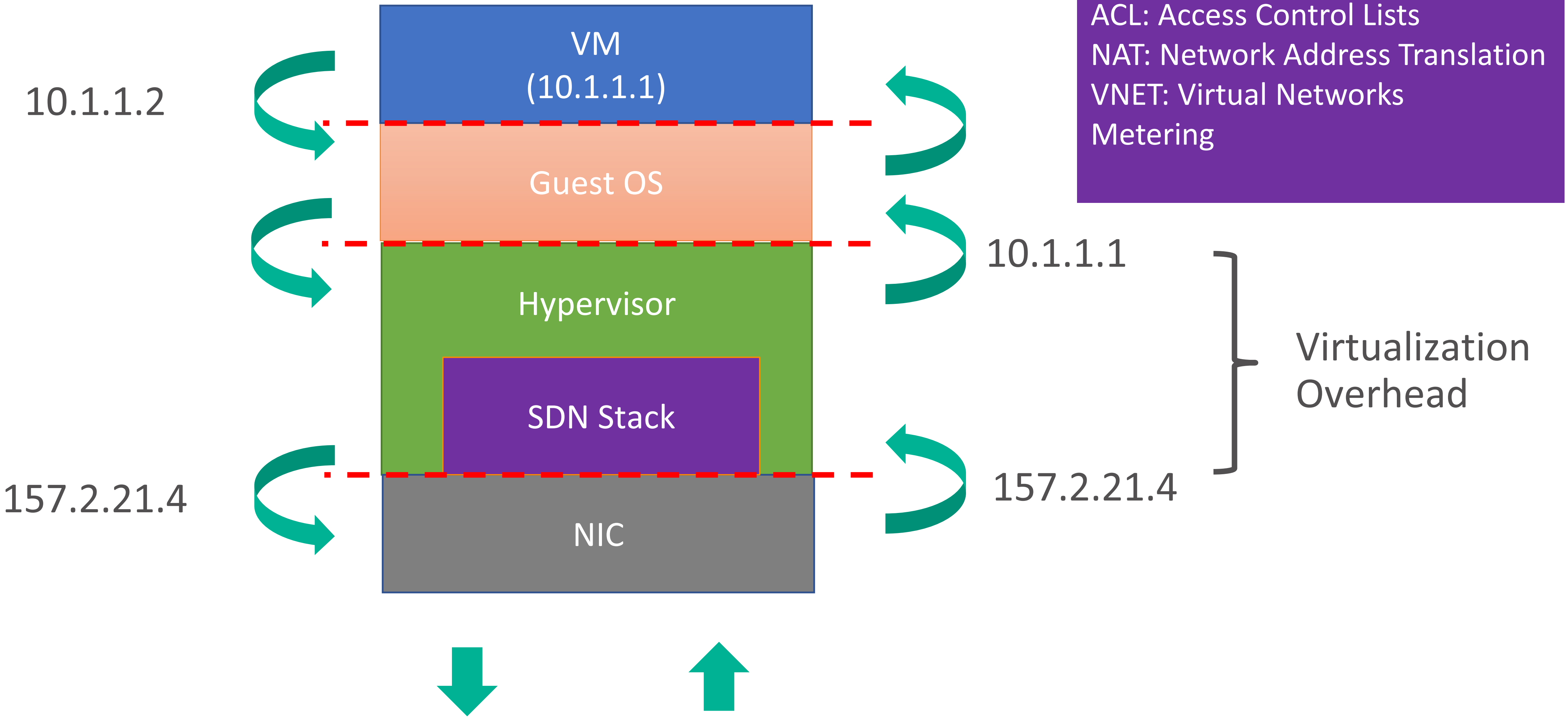
# Network Acceleration – Azure Accelerated Networking



# Network Acceleration – Azure Accelerated Networking



# Virtualization Overhead – Standard Virtual Machines



SDN Stack Functions

SLB: Software Load Balancers

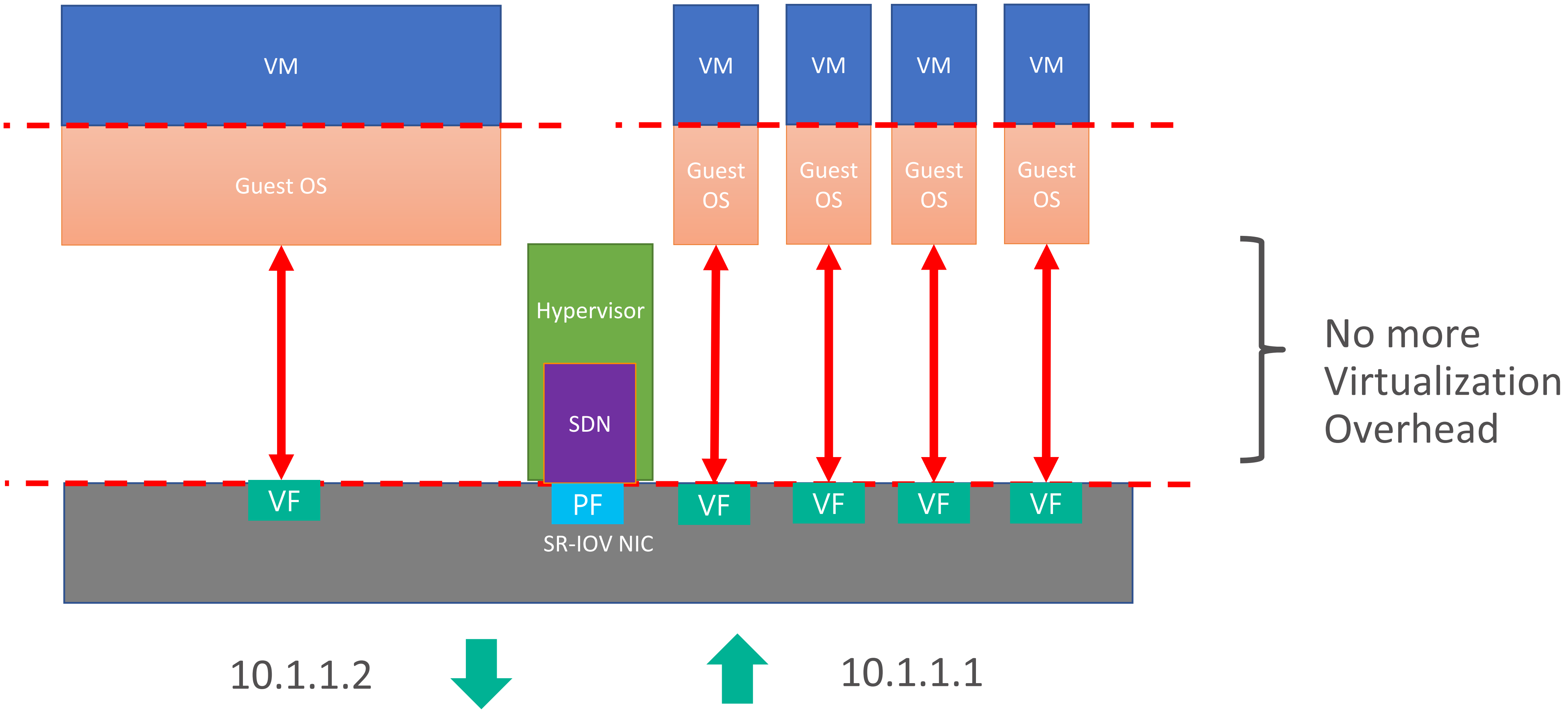
ACL: Access Control Lists

NAT: Network Address Translation

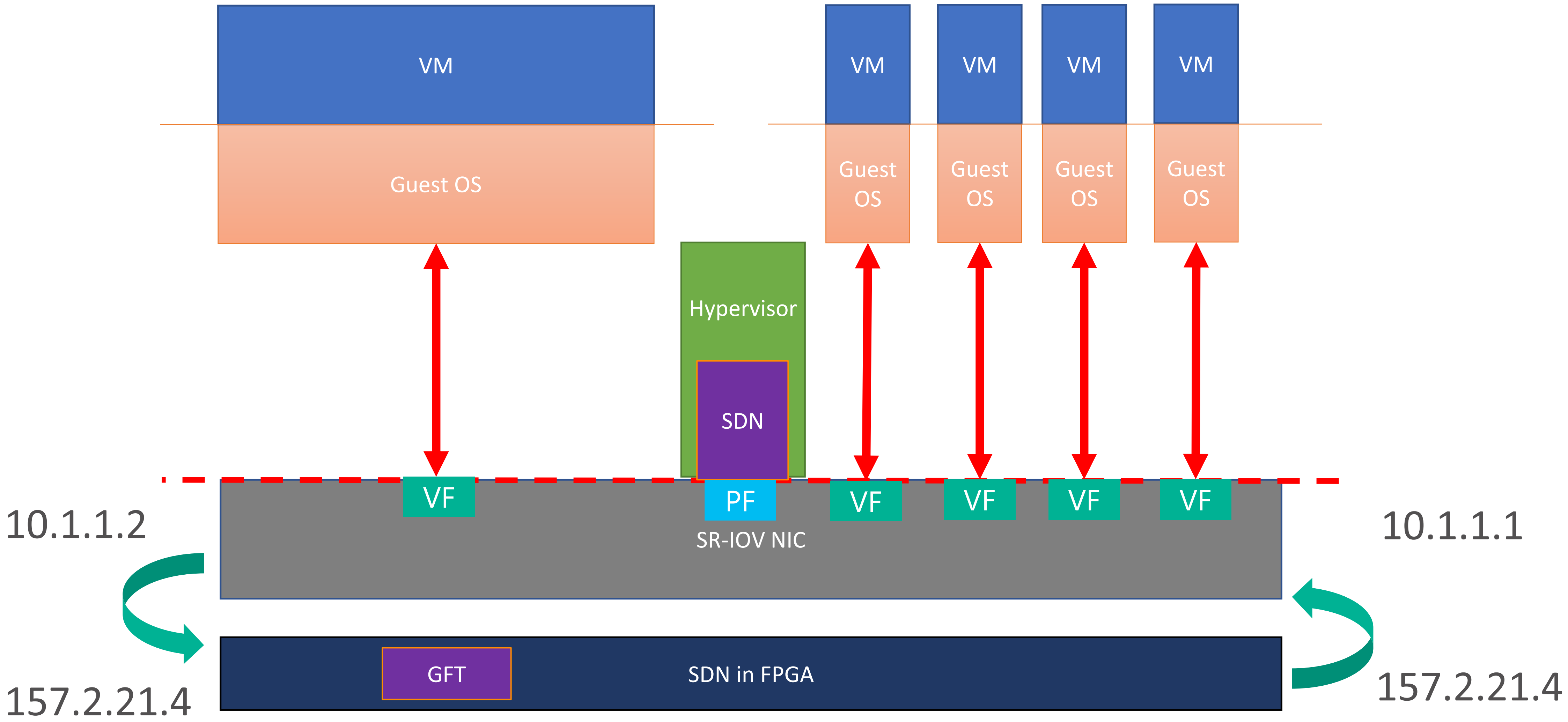
VNET: Virtual Networks

Metering

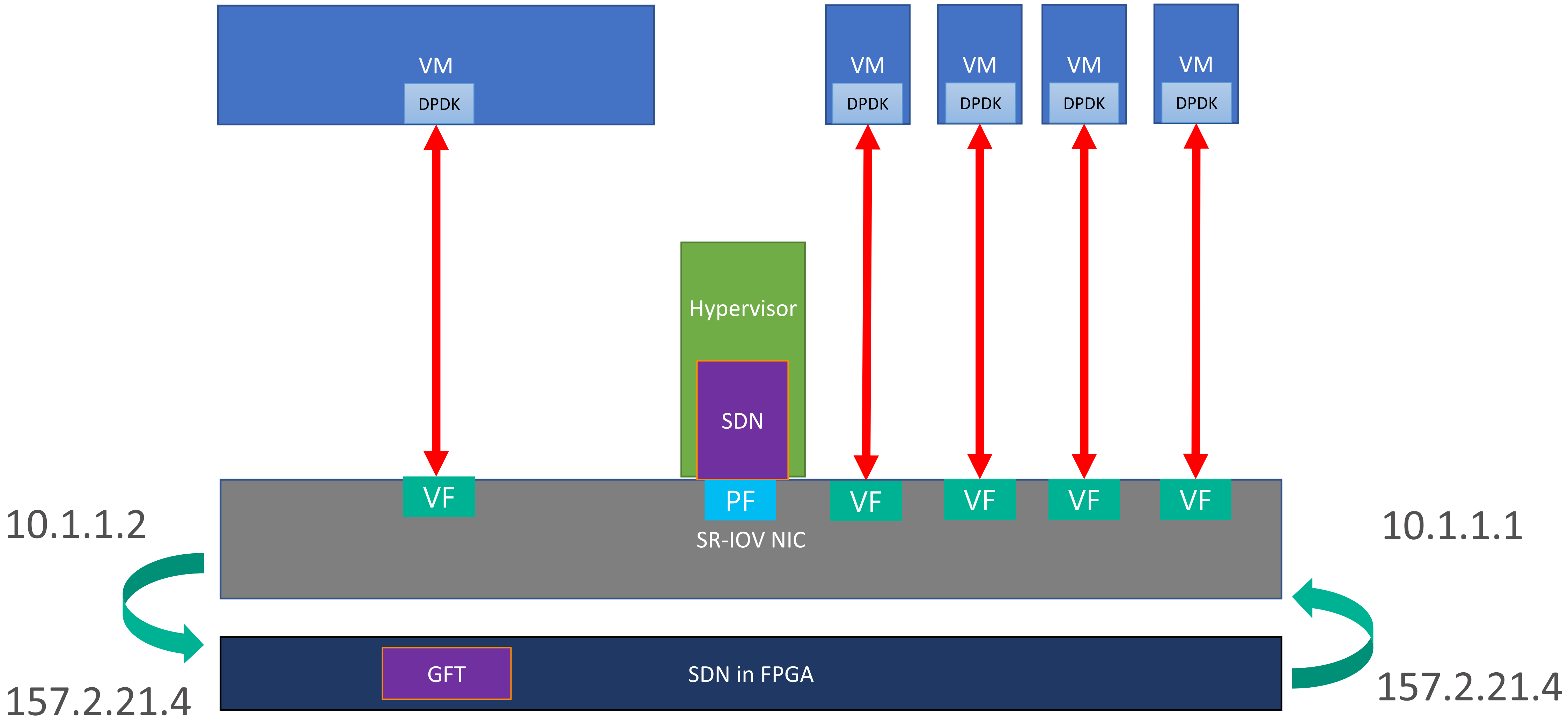
# Virtualization Overhead – SRIOV NICs



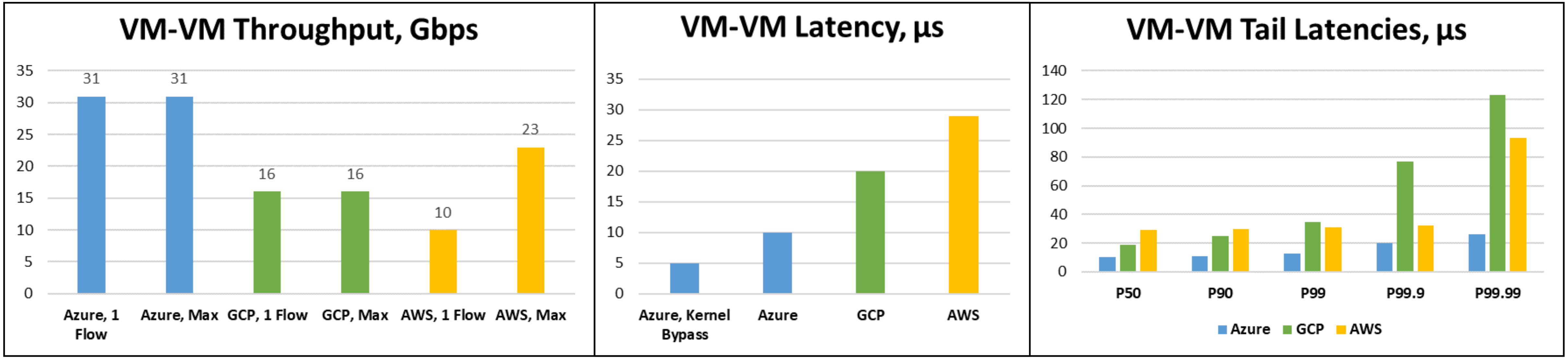
# Virtualization Overhead – Azure SmartNIC w/ FPGAs



# Virtualization Overhead – Azure SmartNIC w/ FPGAs & DPDK



# AccelNet Performance



VM-VM Throughput, Gbps

VM-VM Latency, µs

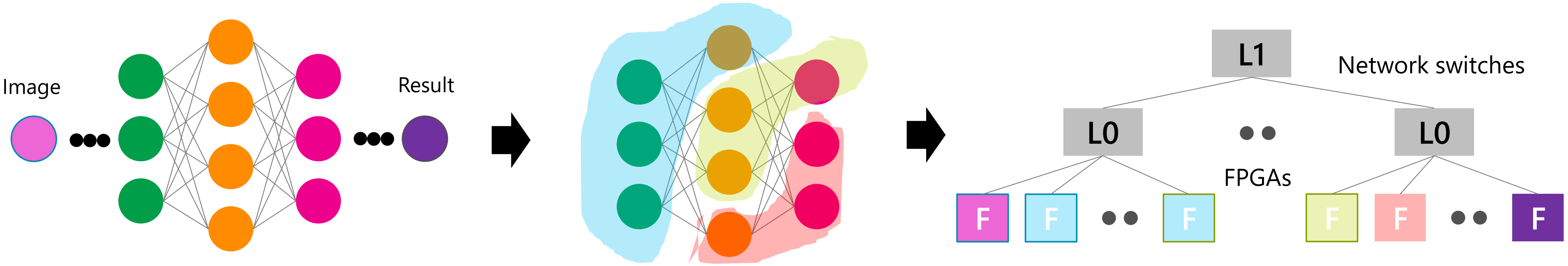
VM-VM Tail Latencies, µs

Lowest latency, highest-bandwidth network in the Cloud… for a while

Same approach can work for storage I/O

# Distributed Hardware-as-a-Service

- Use multiple FPGAs to perform a complex task
- Head node is the gateway
- Use as many nodes as necessary – calling function doesn’t need to know

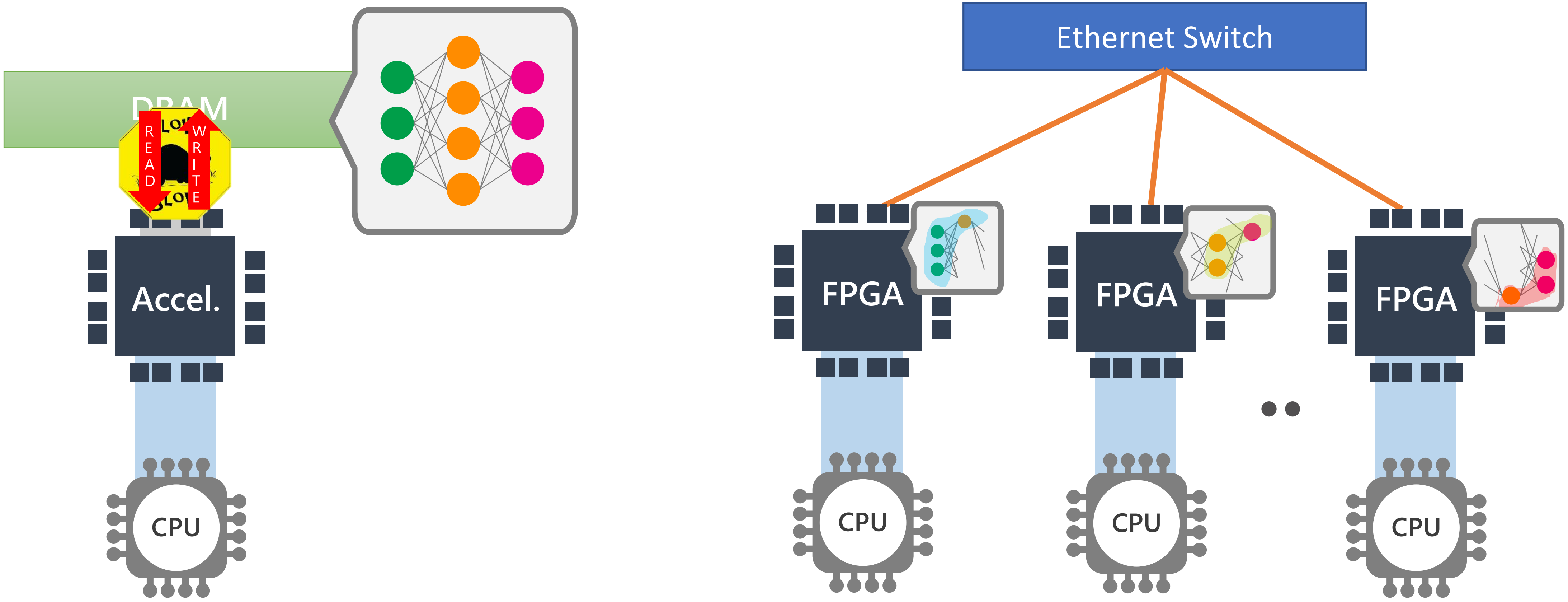


Pretrained DNN Model
(CNTK, TensorFlow, etc…)

BrainWave Compiler & Mapper

Scalable DNN Hardware
Microservice

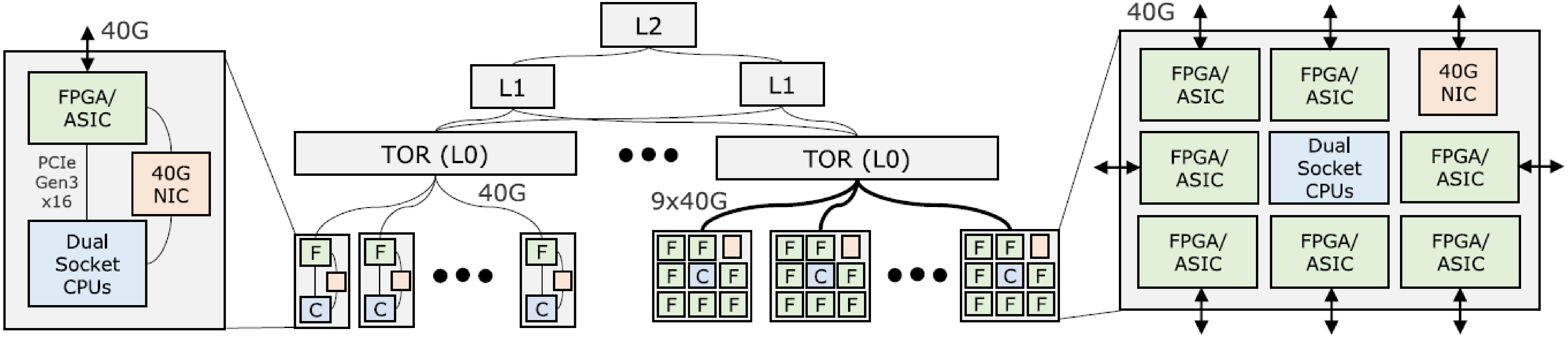
# Project BrainWave Concept



Traditional Approach

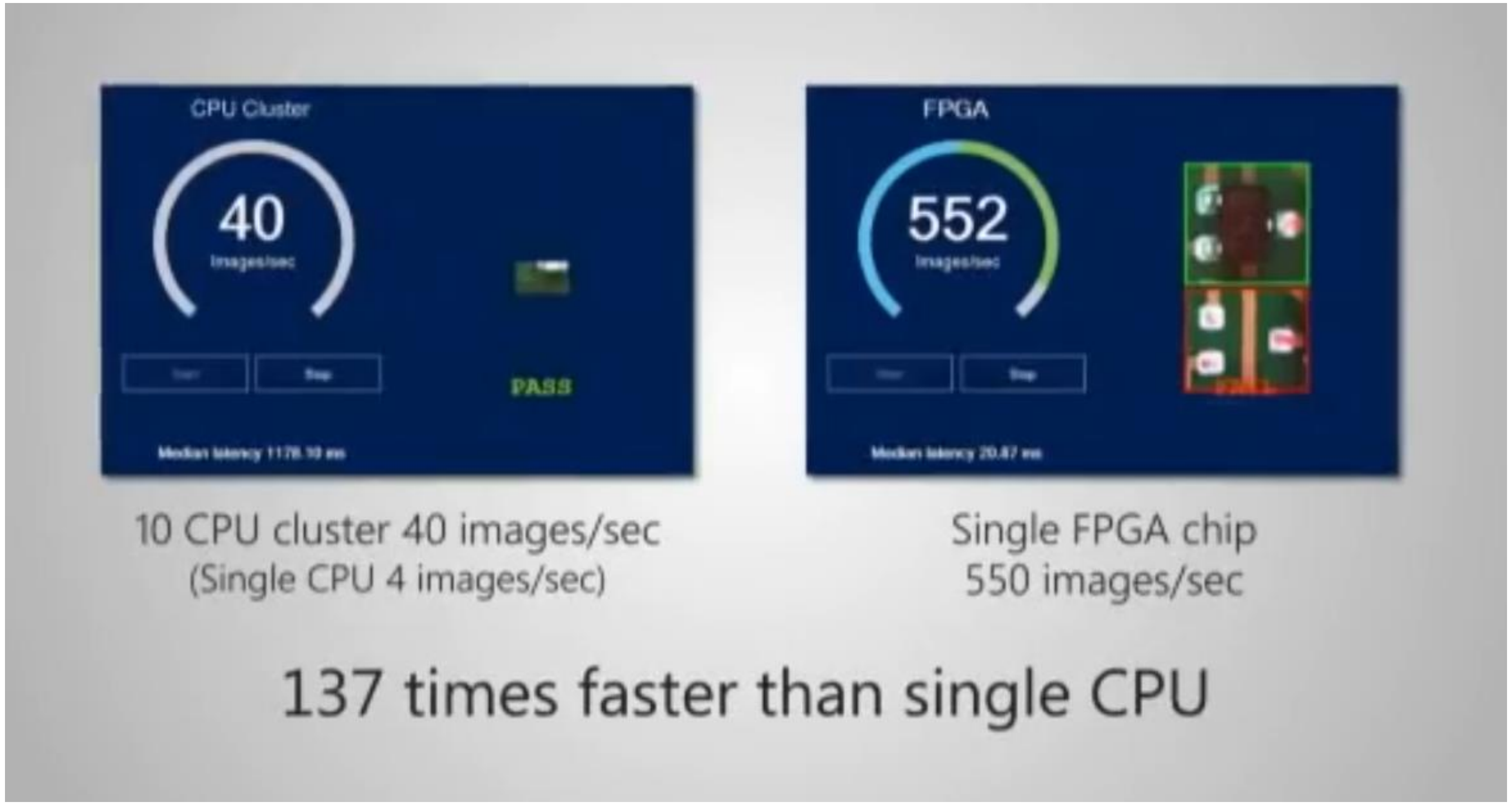
Project Brainwave

# Brainwave System at Cloud-Scale



# Project Brainwave – Fastest Image Classification

- 8 billion operations per image
- 1.5 ms on Resnet-50 with batch size of 1  (Realtime AI)
- Available for anyone to upload their models and run



Median latency 1178.10 ms

Median latency 20.87 ms

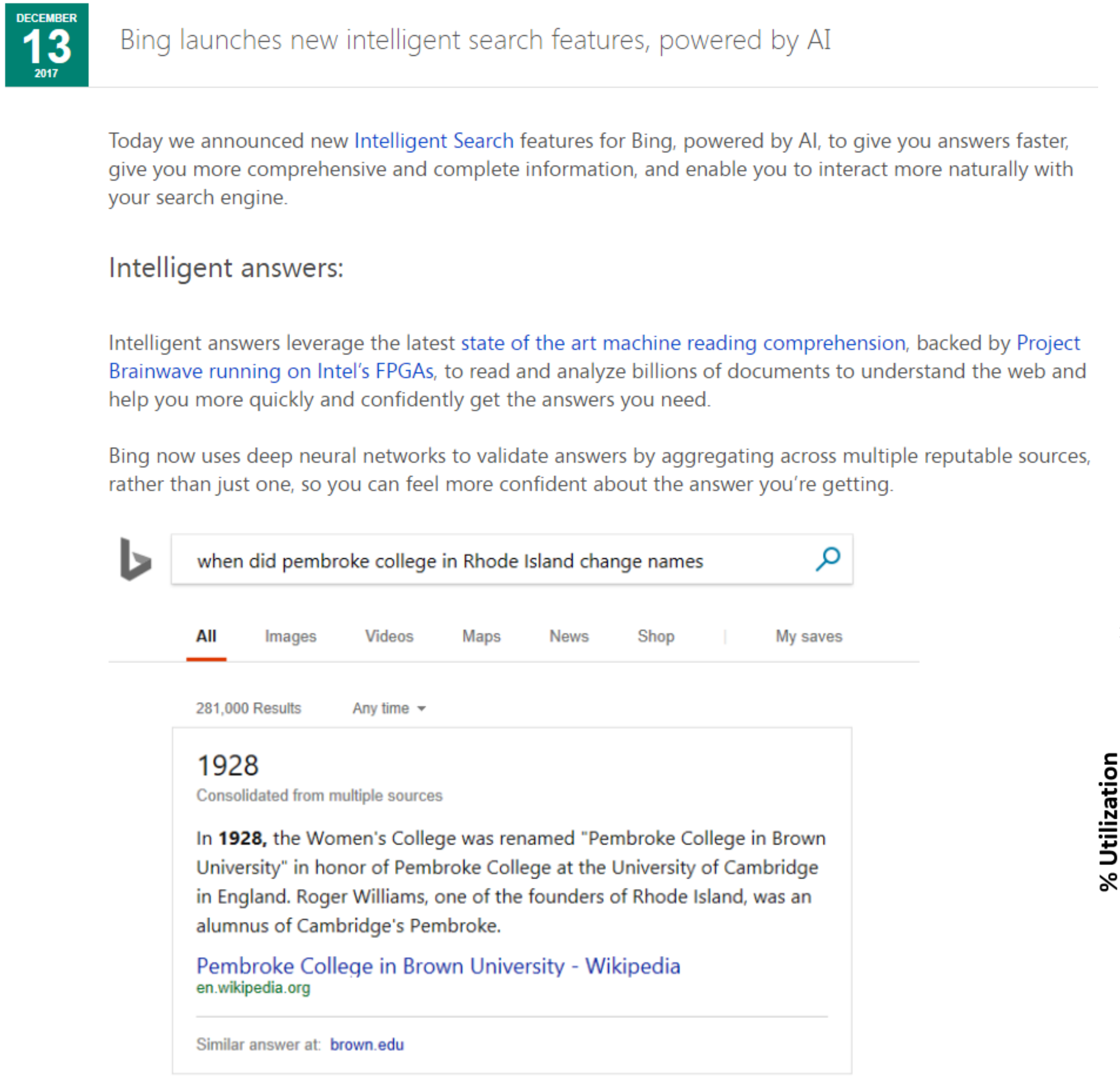
10 CPU cluster 40 images/sec
(Single CPU 4 images/sec)

Single FPGA chip
550 images/sec

137 times faster than single CPU

https://github.com/Azure/aml-real-time-ai

# Bing Intelligent Search Backed By Project Brainwave



DECEMBER
13
2017

Bing launches new intelligent search features, powered by AI

Today we announced new Intelligent Search features for Bing, powered by AI, to give you answers faster, give you more comprehensive and complete information, and enable you to interact more naturally with your search engine.

Intelligent answers:

Intelligent answers leverage the latest state of the art machine reading comprehension, backed by Project Brainwave running on Intel's FPGAs, to read and analyze billions of documents to understand the web and help you more quickly and confidently get the answers you need.

Bing now uses deep neural networks to validate answers by aggregating across multiple reputable sources, rather than just one, so you can feel more confident about the answer you're getting.

when did pembroke college in Rhode Island change names

281,000 Results    Any time

1928

Consolidated from multiple sources

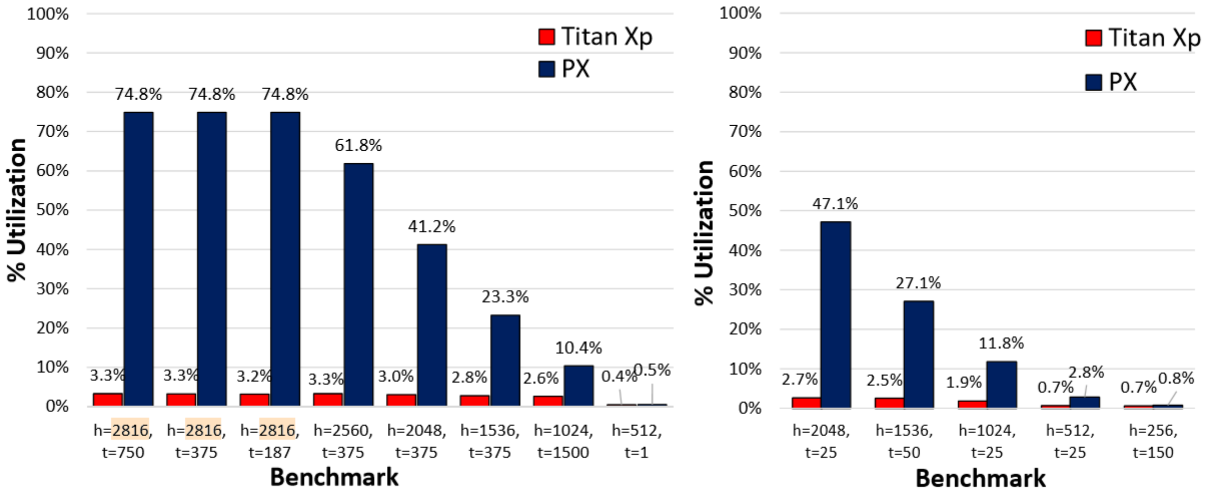
In **1928,** the Women's College was renamed "Pembroke College in Brown University" in honor of Pembroke College at the University of Cambridge in England. Roger Williams, one of the founders of Rhode Island, was an alumnus of Cambridge's Pembroke.

Pembroke College in Brown University - Wikipedia
en.wikipedia.org

Similar answer at: brown.edu

## Stratix V RNN-optimized NPUs in Scale Production

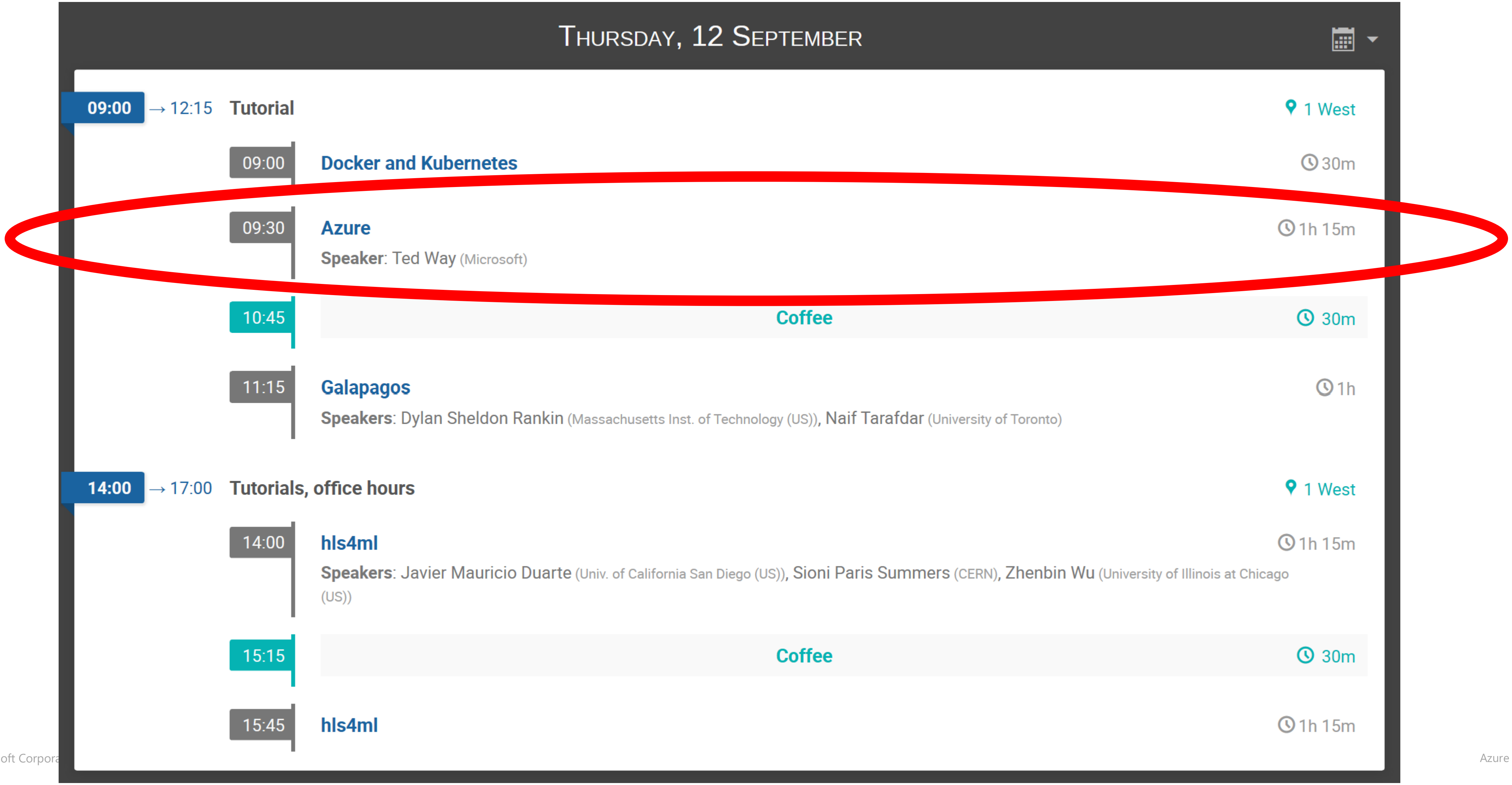
| **Bing TP1** | | | |
|---|---|---|---|
| | **CPU-only** | **Brainwave-accelerated** | **Improvement** |
| Model details | GRU 128x200 (x2)<br>+ W2Vec | LSTM 500x200 (x8)<br>+ W2Vec | Brainwave-accelerated model is > 10X larger and > 10X lower latency |
| End-to-end latency per Batch 1 request at 95% | 9 ms | 0.850 ms | |
| **Bing DeepScan** | | | |
| | **CPU-only** | **Brainwave-accelerated** | **Improvement** |
| Model details | 1D CNN + W2Vec<br>*(RNNs removed)* | 1D CNN + W2Vec<br>+ GRU 500x500 (x4) | Brainwave-accelerated model is > 10X larger and 3X lower latency |
| End-to-end latency per Batch 1 request at 95% | 15 ms | 5 ms | |



(a) GRU

(b) LSTM

# Attend the Turotial tomorrow for more detail on BrainWave



## THURSDAY, 12 SEPTEMBER

**09:00** → 12:15 **Tutorial** — 1 West

09:00 **Docker and Kubernetes** — 30m

09:30 **Azure** — 1h 15m
**Speaker**: Ted Way (Microsoft)

10:45 **Coffee** — 30m

11:15 **Galapagos** — 1h
**Speakers**: Dylan Sheldon Rankin (Massachusetts Inst. of Technology (US)), Naif Tarafdar (University of Toronto)

**14:00** → 17:00 **Tutorials, office hours** — 1 West

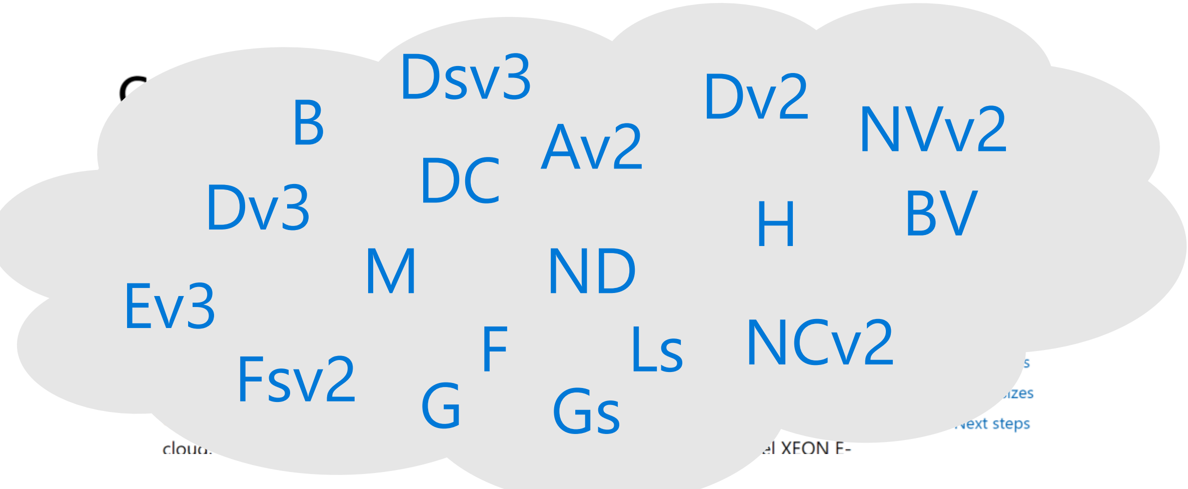
14:00 **hls4ml** — 1h 15m
**Speakers**: Javier Mauricio Duarte (Univ. of California San Diego (US)), Sioni Paris Summers (CERN), Zhenbin Wu (University of Illinois at Chicago (US))

15:15 **Coffee** — 30m

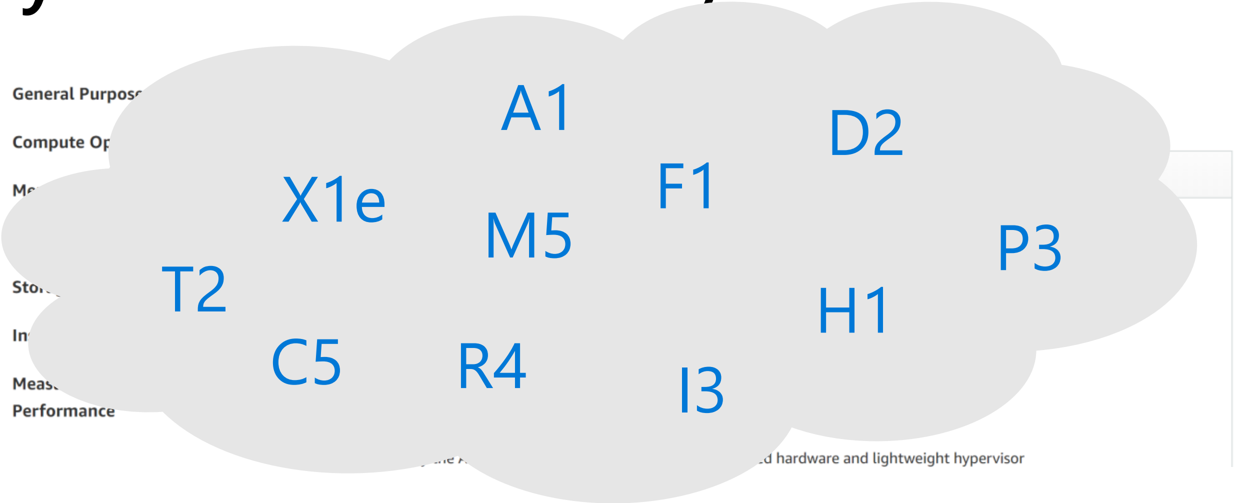
15:45 **hls4ml** — 1h 15m

# Why will FPGAs still be here tomorrow?

- **Performance? Low Power? Ubiquity? Killer features? AI/ML?**



Azure (Nov 2018)



Amazon AWS (Nov 2018)

- **Every time we change hardware, we need to make a new VM type**
- **FPGAs allow us to back new features into existing VM types**

**Developers (Customers) don’t want to spend time fighting old battles**

Microsoft