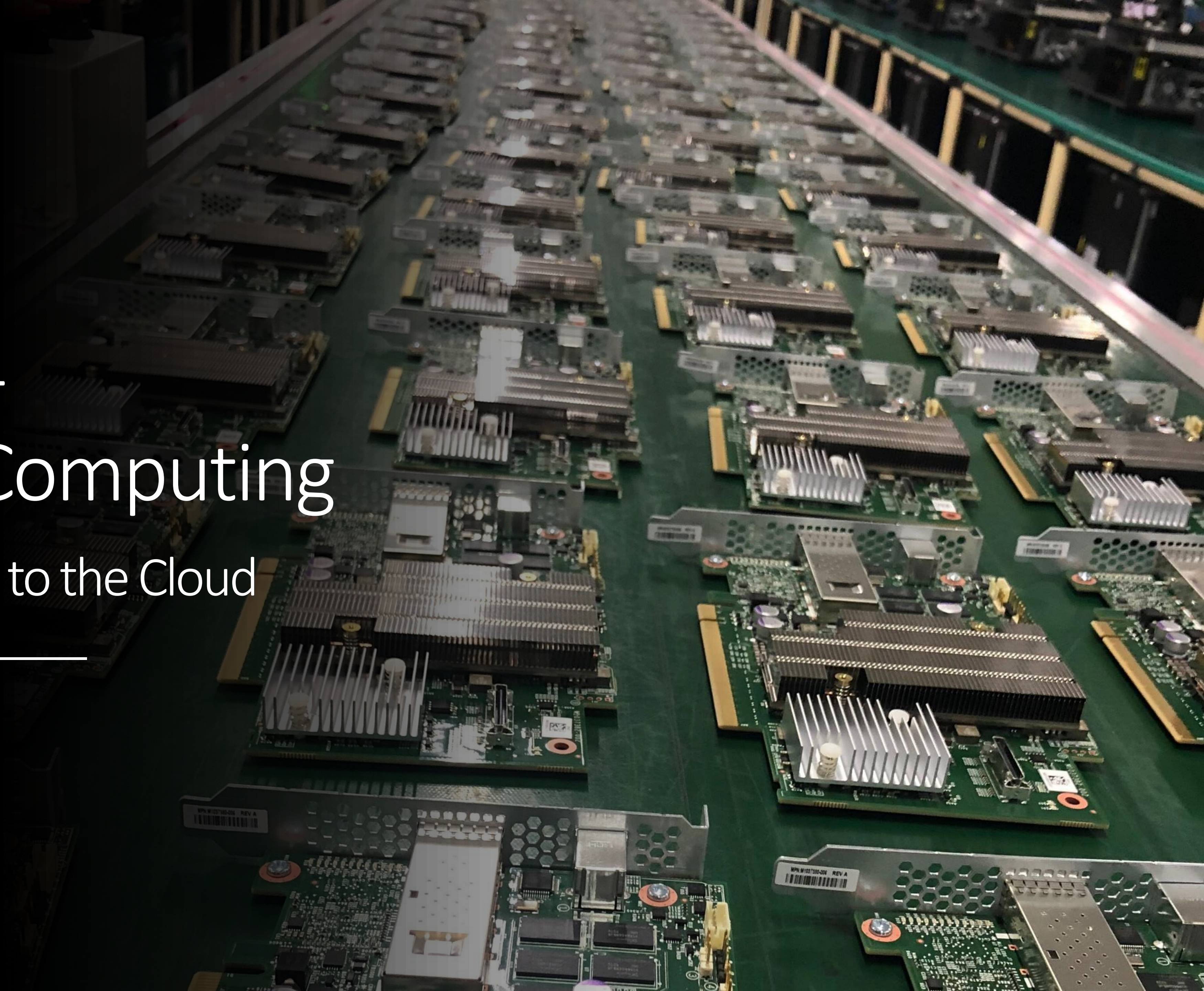
Microsoft



FPGAs in High-Performance Computing

from Data Acquisition to the Cloud

Andrew Putnam

May 25, 2022

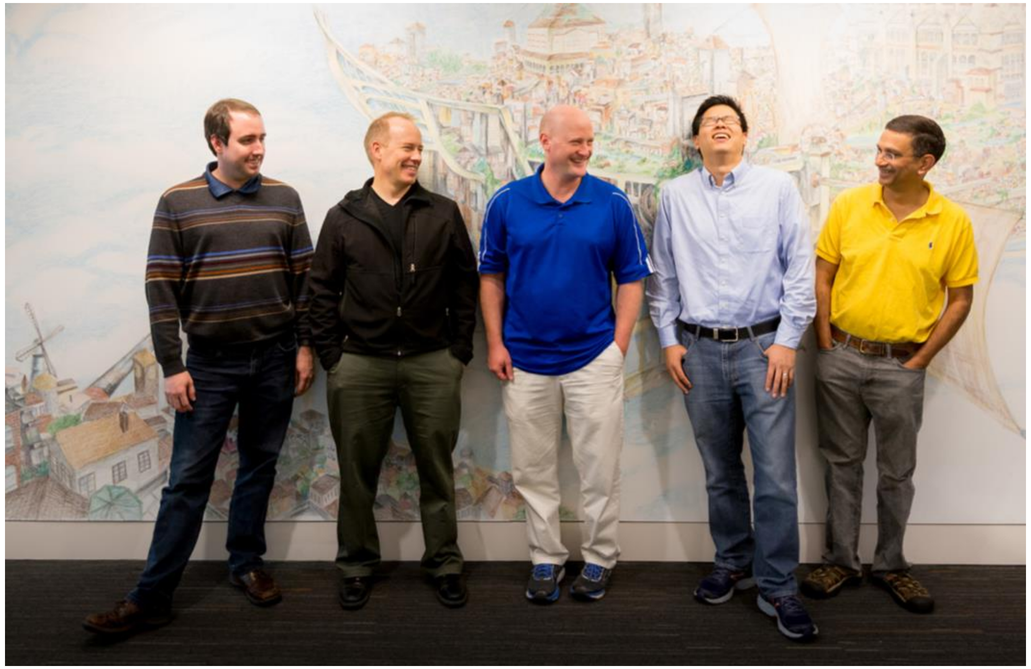
Introduction

Microsoft Research

2009 - 2016

Microsoft Azure

2016 - Present



Co-Founder of Catapult

Board Architect

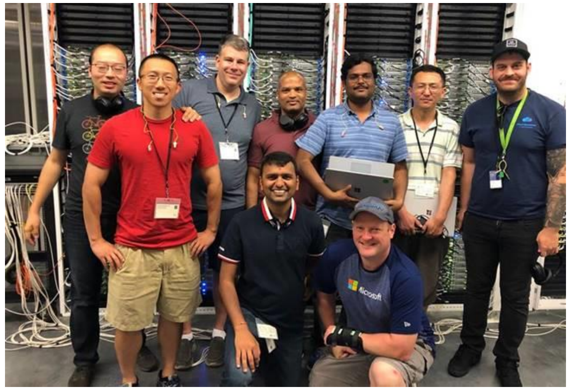
RTL Coder

Application Developer

Manufacturing Floor (Operator)

Functional Test Programmer

Rack Integrator

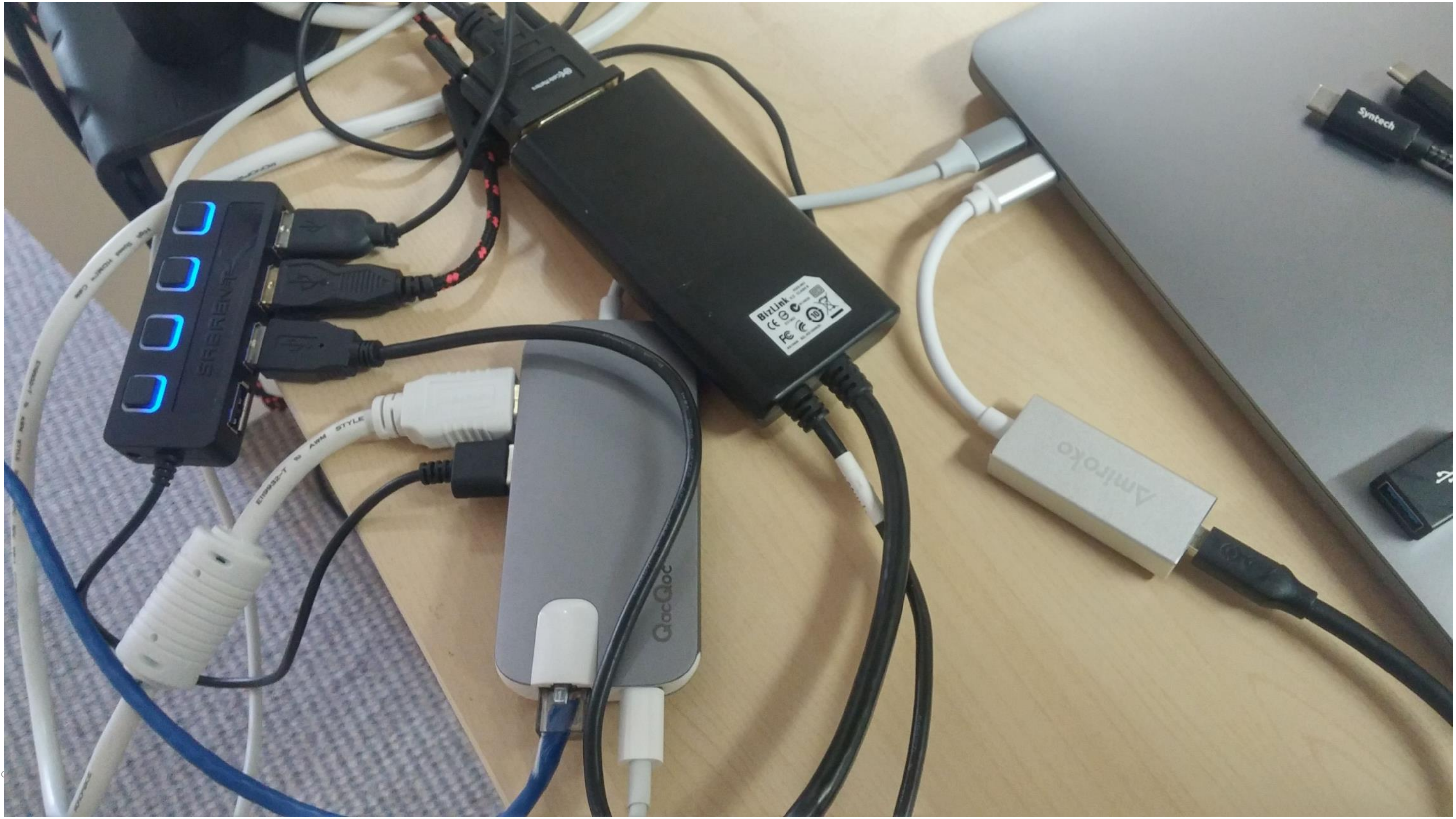


SmartNIC Lead

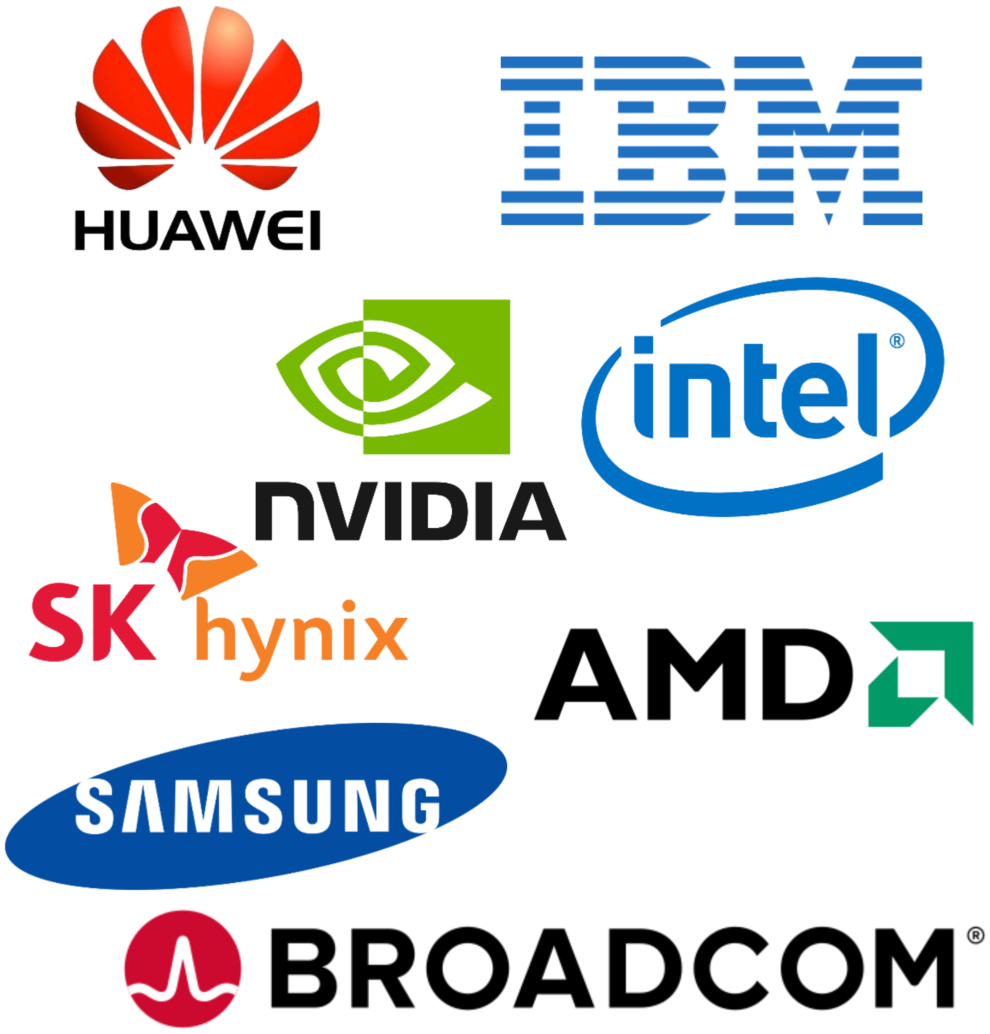
Director of 1P Accelerator Arch & Platforms

Slick new hardware

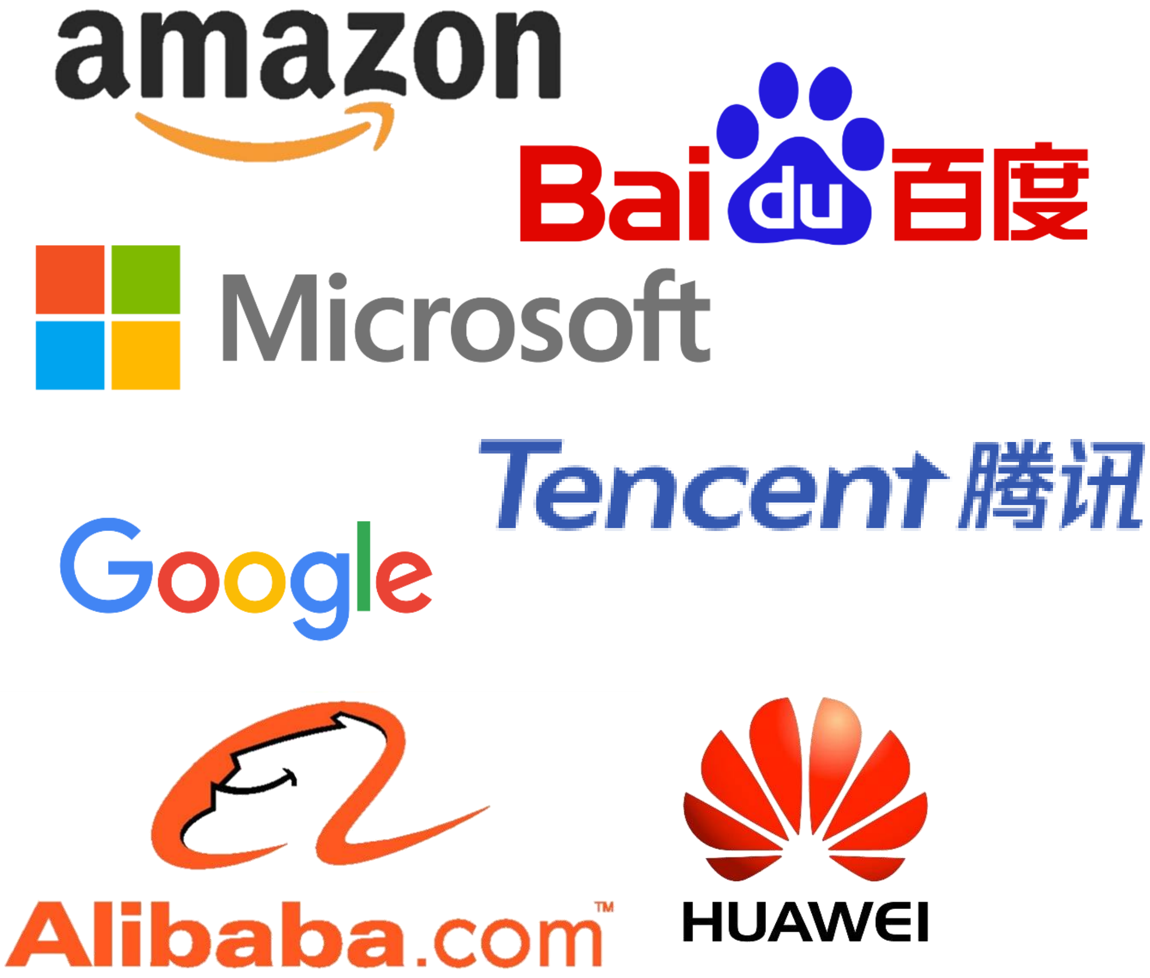
Useful hardware



What makes a public cloud company successful?



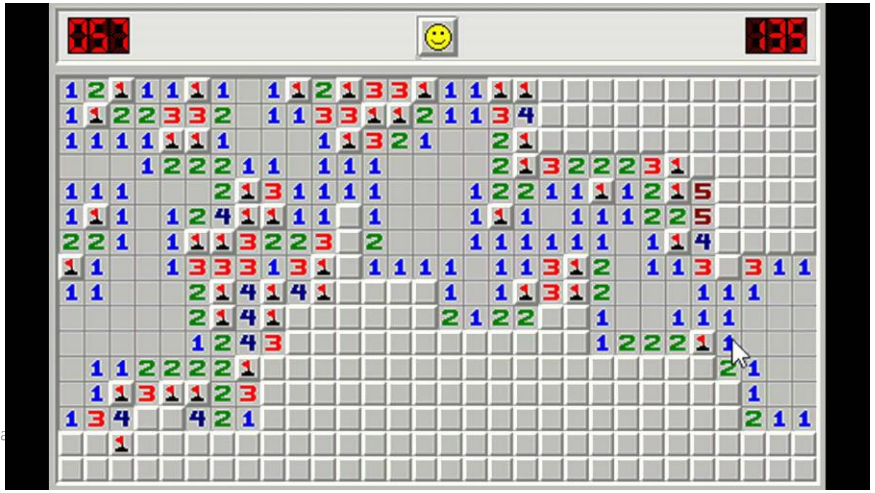
Hardware Companies

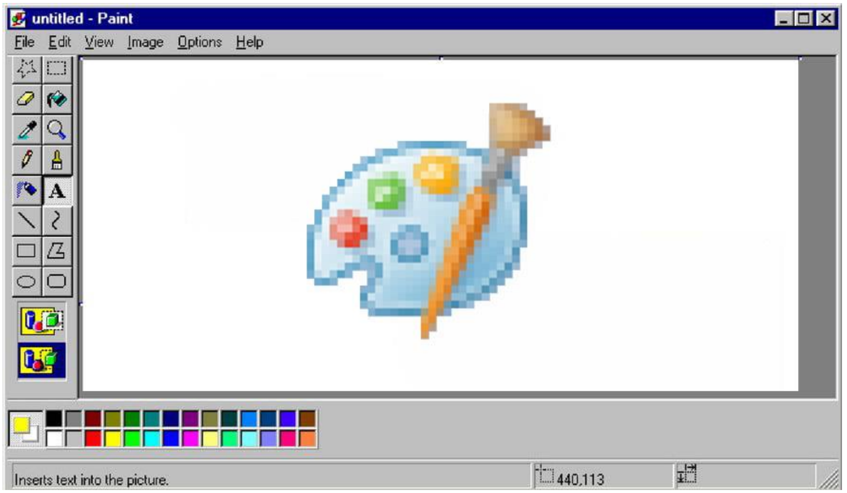


Software Companies

Innovation in Software vs. Hardware

- **SW is flexible, but is also SO much bigger**
- **You can’t lead from the bottom**
  - Just look at AMD GPUs vs. nVidia
  - x86 and Windows aren’t the leaders because they’ve *always* been the best
- **Nobody wants to do throw-away work**
  - Work needs to (plausibly) span multiple generations





Inserts text into the picture.

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.

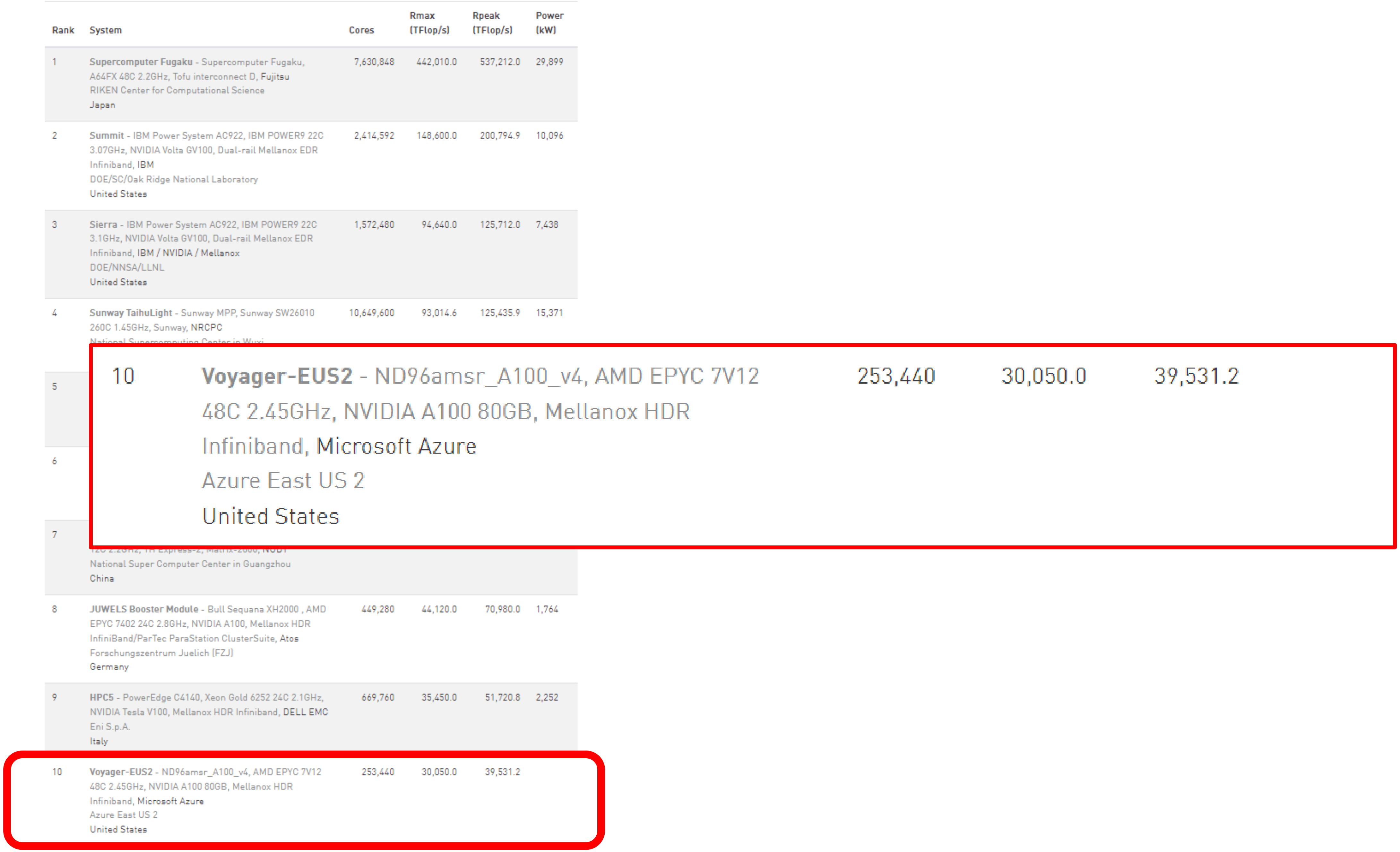
Why hasn’t Supercomputing moved to the Cloud?

CPUs look largely the same, but…

❑ Top 500 often include specialized accelerators (especially GPUs)

❑ Networks are highly specialized, tuned for low-latency, high bandwidth

❑ Won’t running virtual machines kill performance?



Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442,010.0	537,212.0	29,899
2	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148,600.0	200,794.9	10,096
3	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94,640.0	125,712.0	7,438
4	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi	10,649,600	93,014.6	125,435.9	15,371
5					
6					
7	12C 2.2GHz, TH Express-2, Matrix-2000, NUDT National Super Computer Center in Guangzhou China				
8	**JUWELS Booster Module** - Bull Sequana XH2000 , AMD EPYC 7402 24C 2.8GHz, NVIDIA A100, Mellanox HDR InfiniBand/ParTec ParaStation ClusterSuite, Atos Forschungszentrum Juelich (FZJ) Germany	449,280	44,120.0	70,980.0	1,764
9	**HPC5** - PowerEdge C4140, Xeon Gold 6252 24C 2.1GHz, NVIDIA Tesla V100, Mellanox HDR Infiniband, DELL EMC Eni S.p.A. Italy	669,760	35,450.0	51,720.8	2,252
10	**Voyager-EUS2** - ND96amsr_A100_v4, AMD EPYC 7V12 48C 2.45GHz, NVIDIA A100 80GB, Mellanox HDR Infiniband, Microsoft Azure Azure East US 2 United States	253,440	30,050.0	39,531.2	

10 **Voyager-EUS2** - ND96amsr_A100_v4, AMD EPYC 7V12 48C 2.45GHz, NVIDIA A100 80GB, Mellanox HDR Infiniband, Microsoft Azure
Azure East US 2
United States
253,440 30,050.0 39,531.2

Why hasn’t Supercomputing moved to the Cloud?

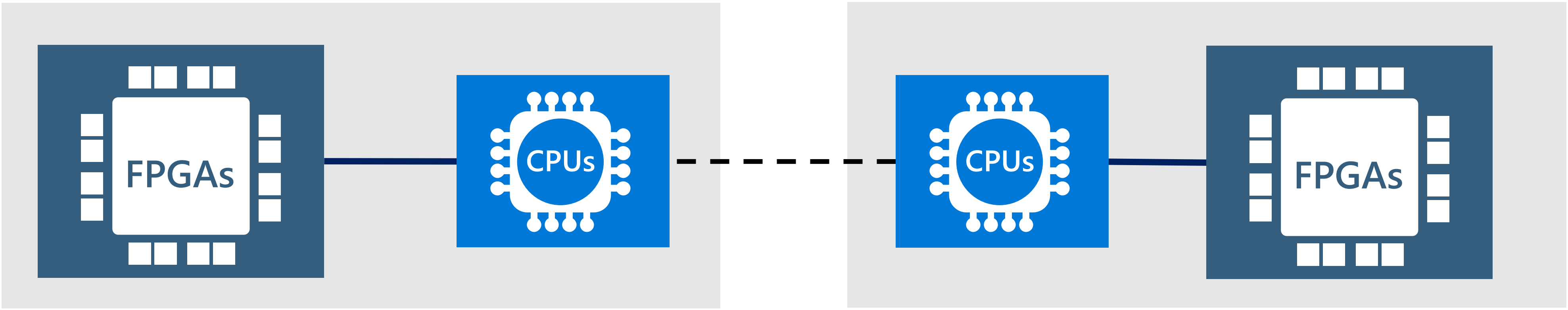
CPUs look largely the same, but…

☑ Top 500 often include specialized accelerators (especially GPUs)

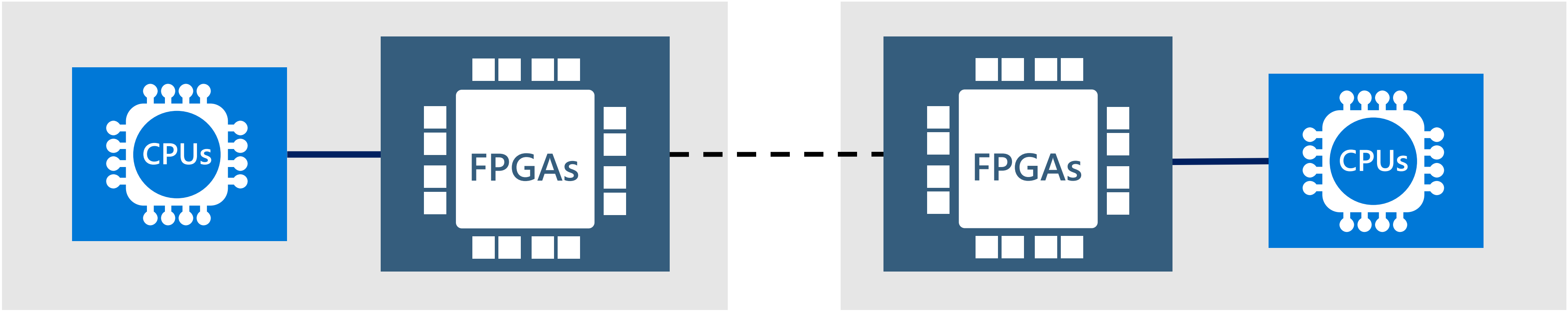
□ Networks are highly specialized, tuned for low-latency, high bandwidth

□ Won’t running virtual machines kill performance?

Accelerator Integration



Traditional Accelerator Integration

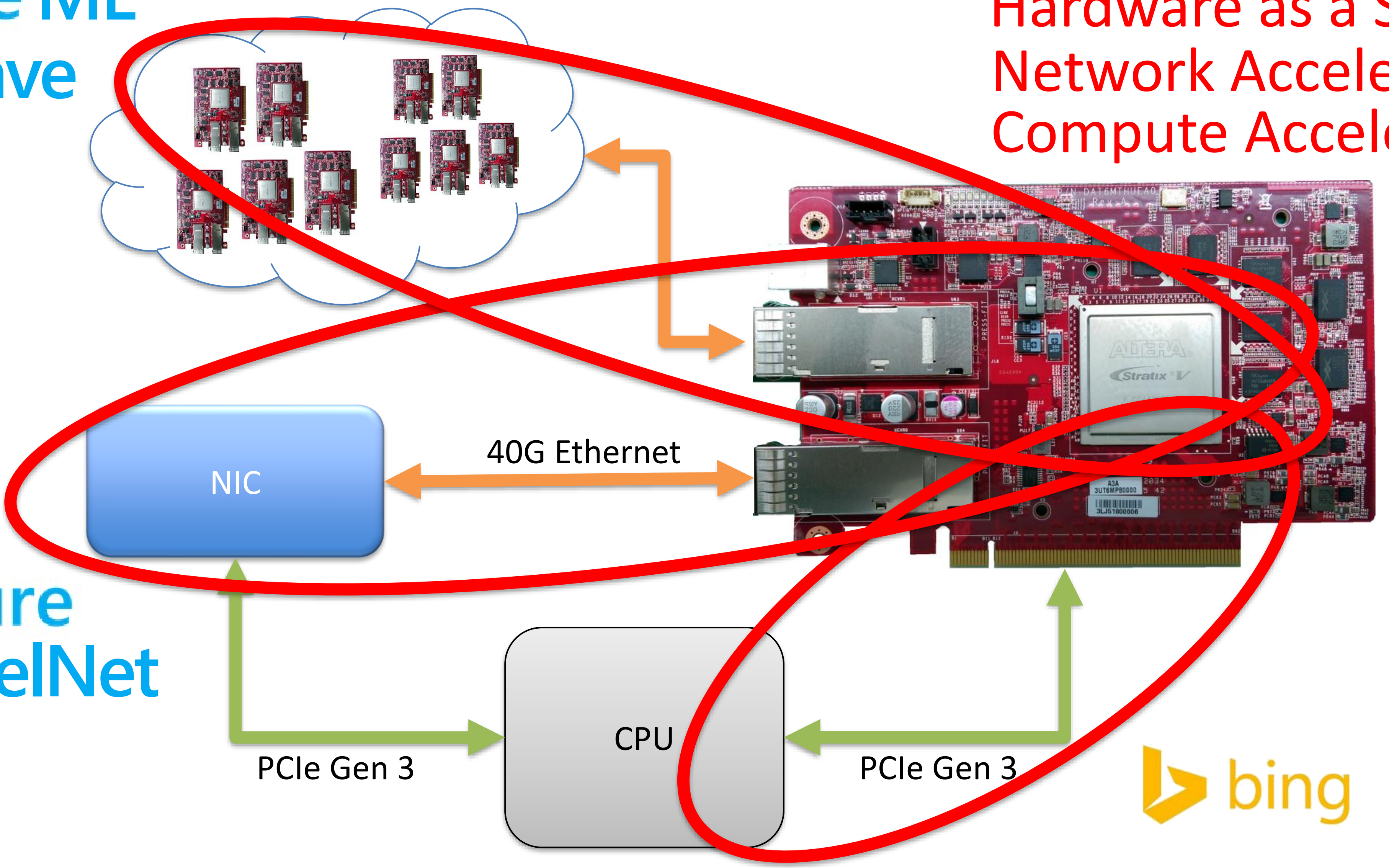


Bump in the Wire -- In-Network Acceleration

Bump-in-the-wire Architecture

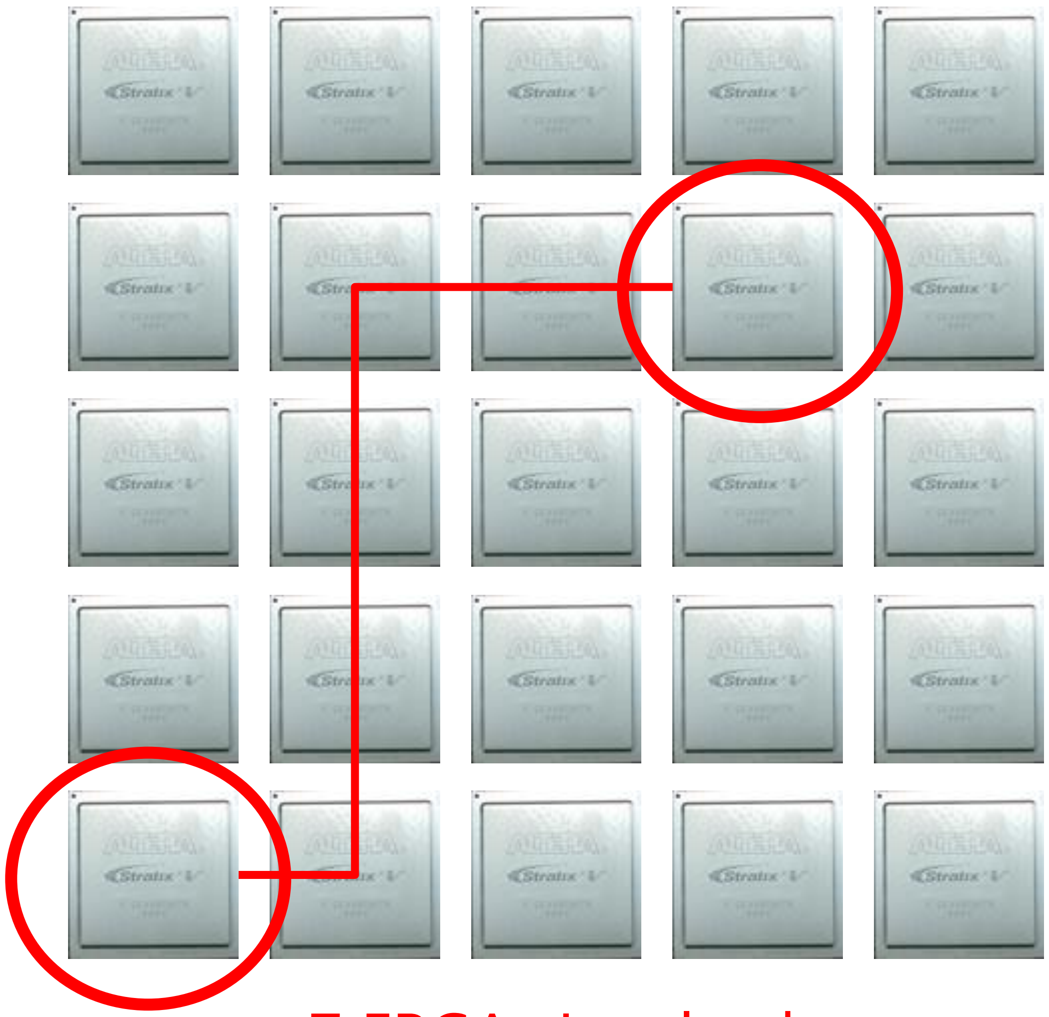
Azure ML BrainWave

Hardware as a Service
Network Acceleration
Compute Acceleration

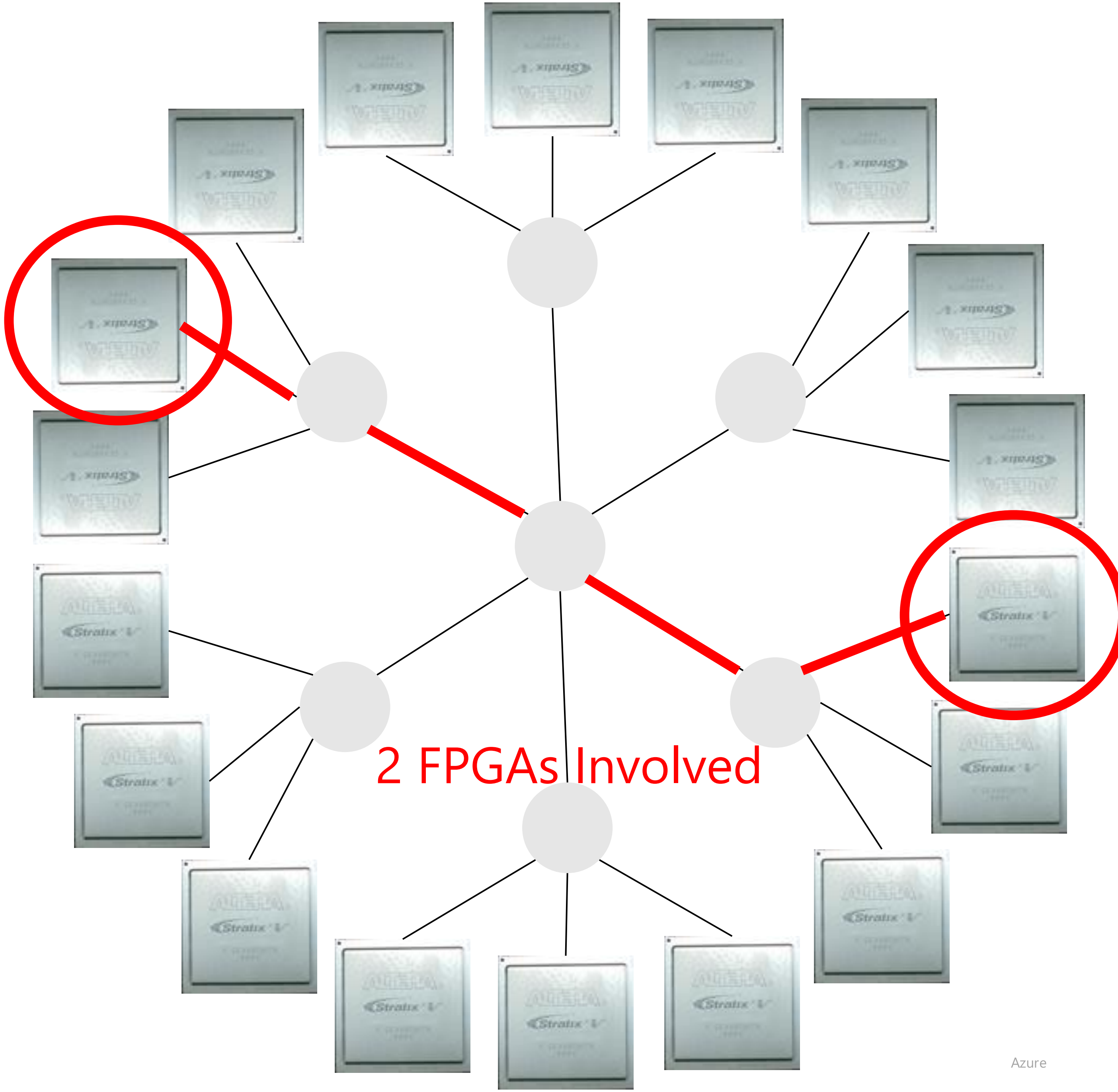


Azure AccelNet

Global-Scale FPGA

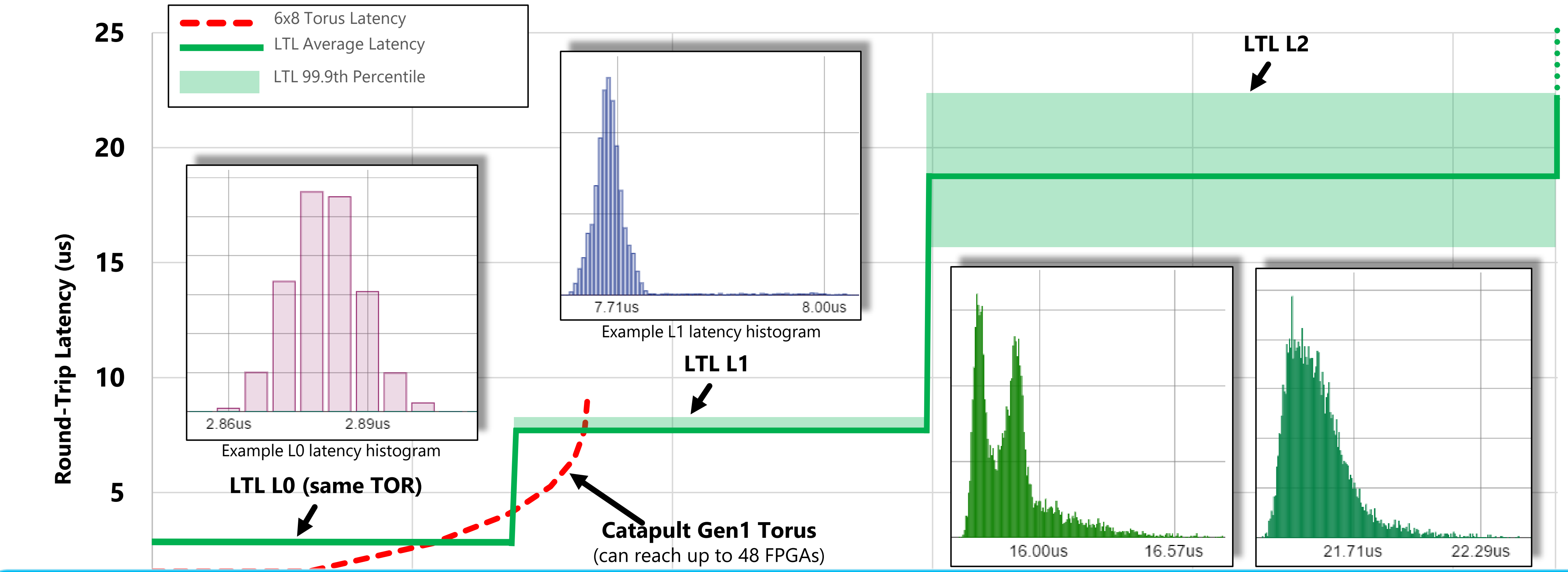


7 FPGAs Involved



2 FPGAs Involved

Network Latencies



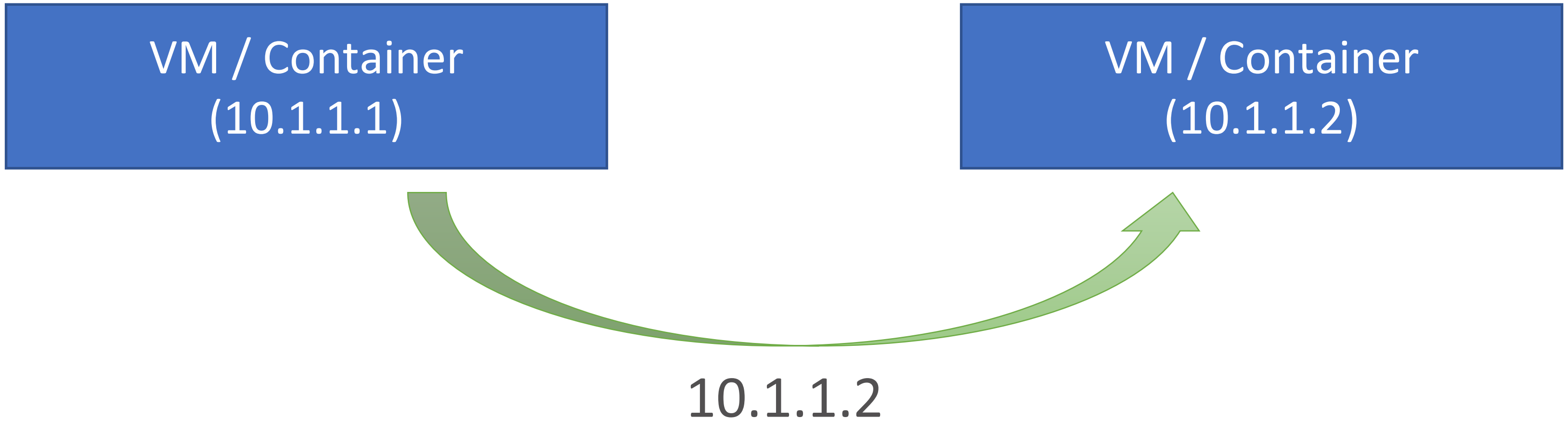
- Extremely low latency (Similar to Infiniband)
- Global-scale FPGA

Why hasn't Supercomputing moved to the Cloud?

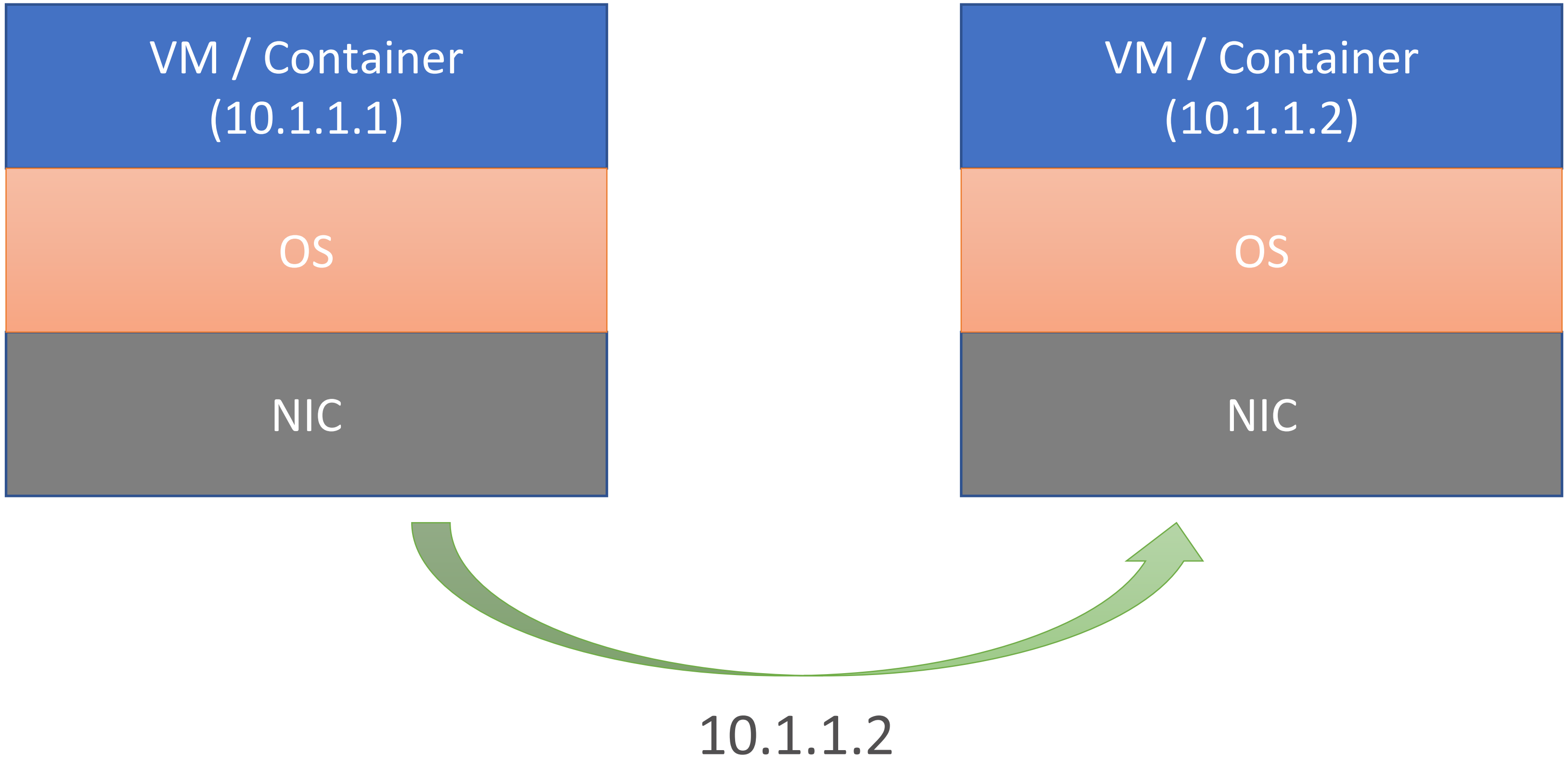
CPUs look largely the same, but…

- ☑ Top 500 often include specialized accelerators (especially GPUs)
- ☑ Networks are highly specialized, tuned for low-latency, high bandwidth
- ❑ Won't running virtual machines kill performance?

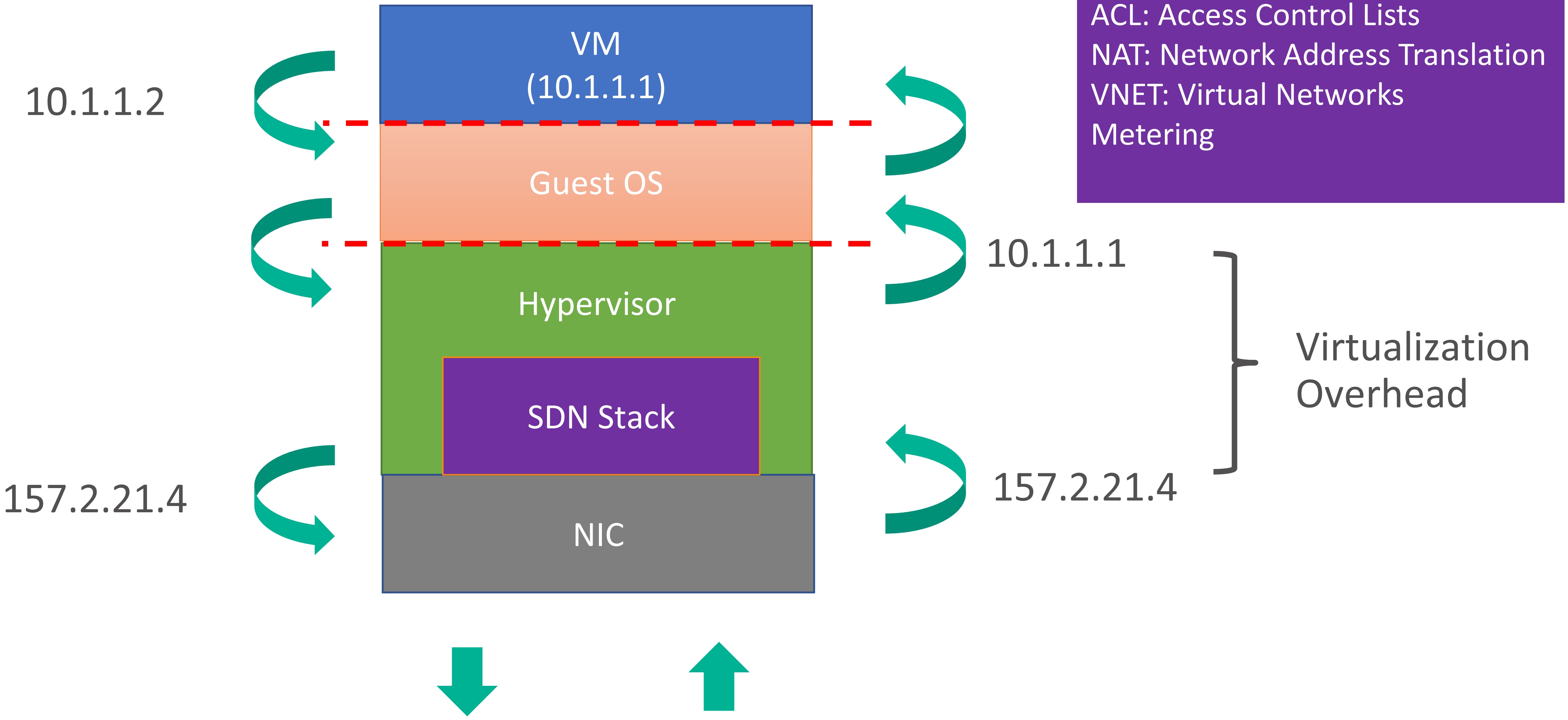
Network Acceleration – Azure Accelerated Networking



Network Acceleration – Azure Accelerated Networking



Virtualization Overhead – Standard Virtual Machines



SDN Stack Functions

SLB: Software Load Balancers

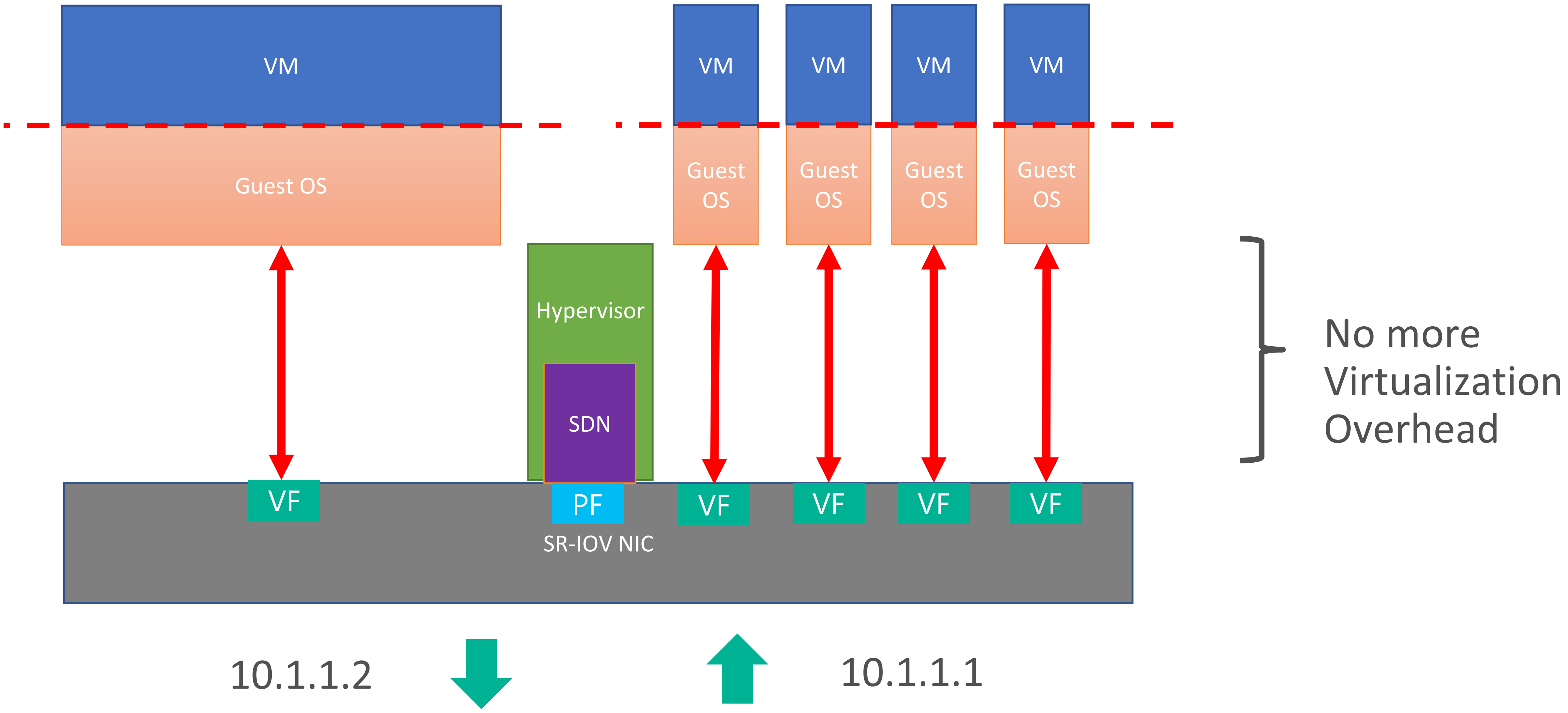
ACL: Access Control Lists

NAT: Network Address Translation

VNET: Virtual Networks

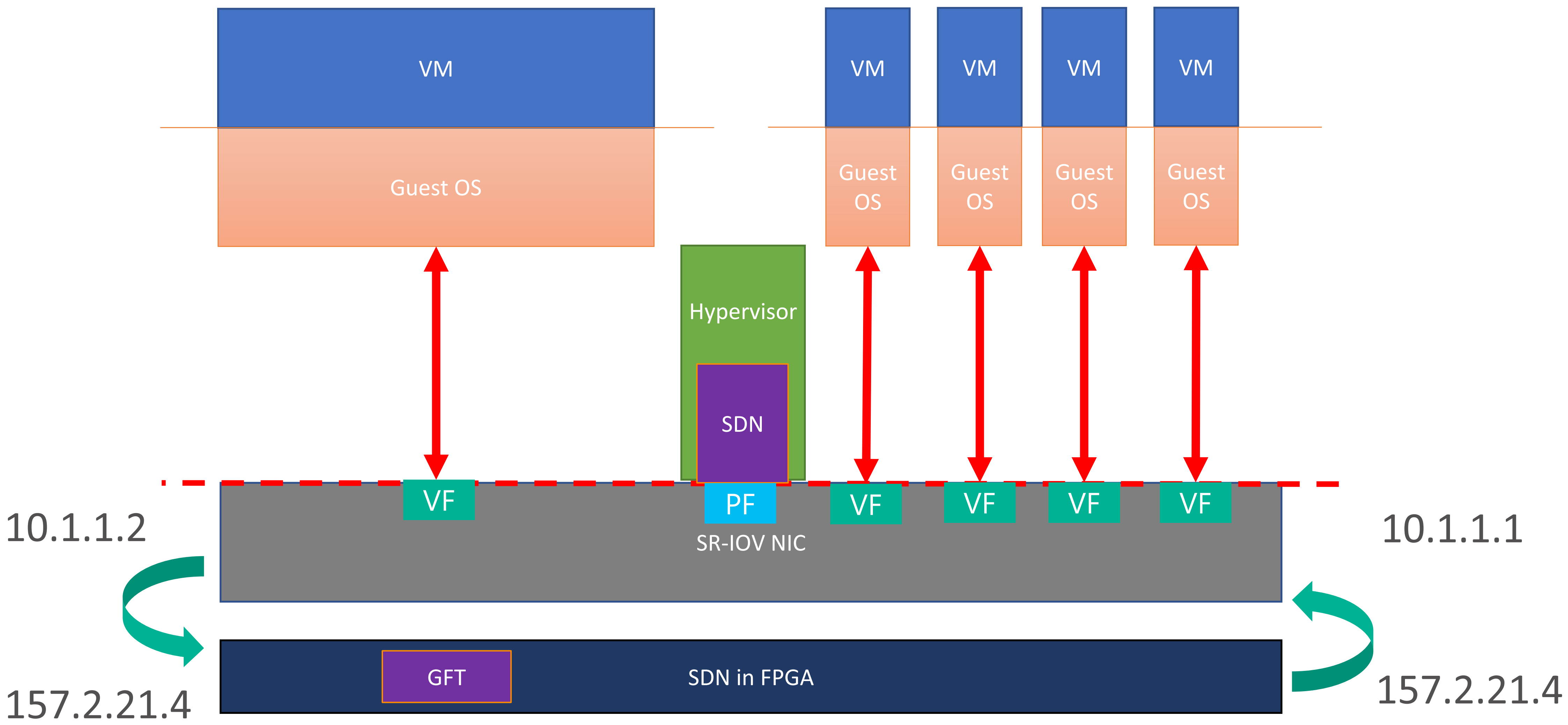
Metering

Virtualization Overhead – SRIOV NICs

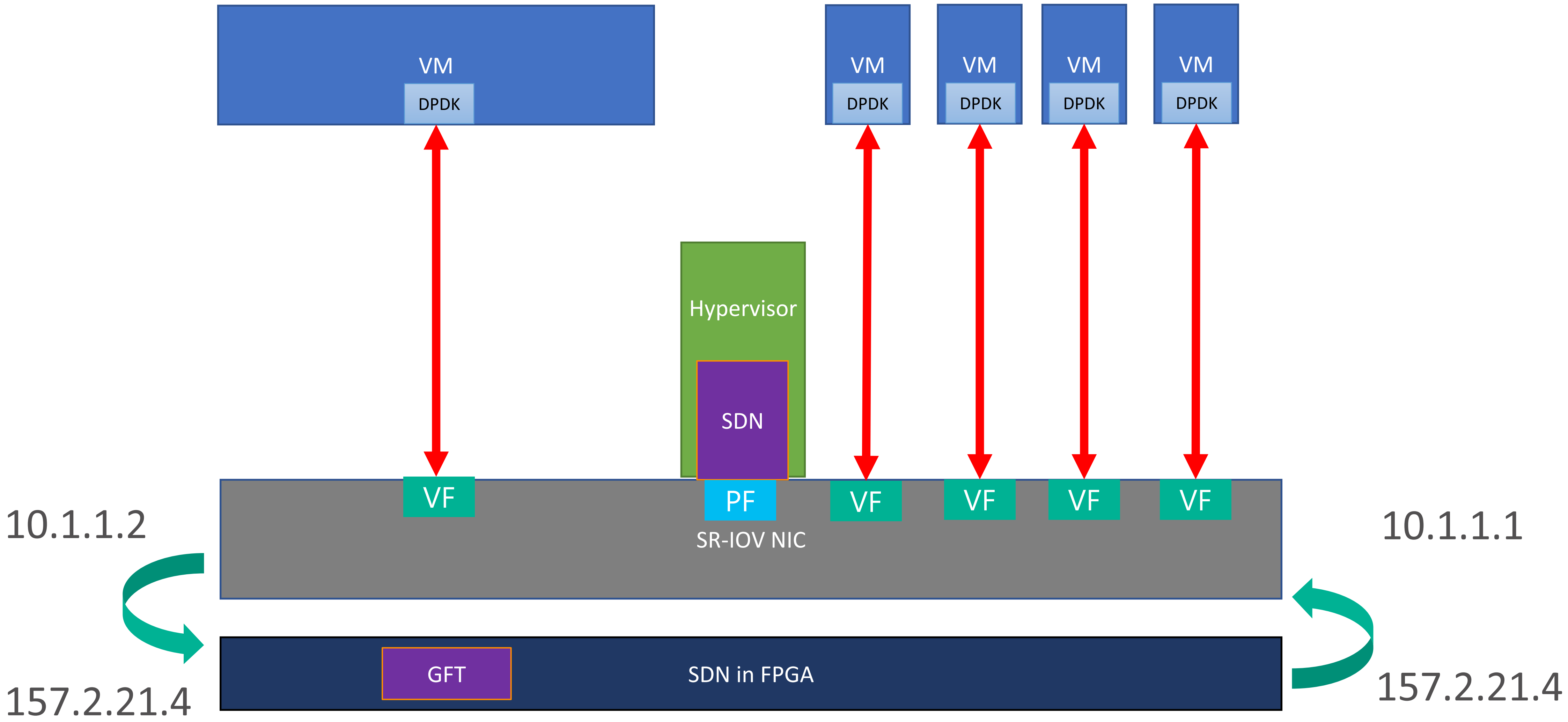


No more Virtualization Overhead

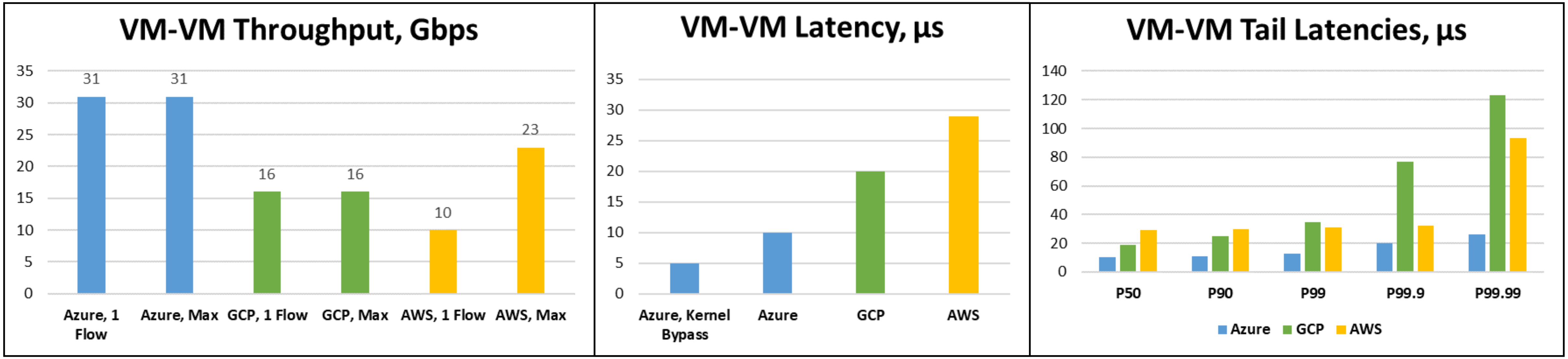
Virtualization Overhead – Azure SmartNIC w/ FPGAs



Virtualization Overhead – Azure SmartNIC w/ FPGAs & DPDK



AccelNet Performance



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

Why hasn’t Supercomputing moved to the Cloud?

CPUs look largely the same, but…

☑ Top 500 often include specialized accelerators (especially GPUs)

☑ Networks are highly specialized, tuned for low-latency, high bandwidth

☑ Won’t running virtual machines kill performance?

Will HPC developers adopt the cloud?

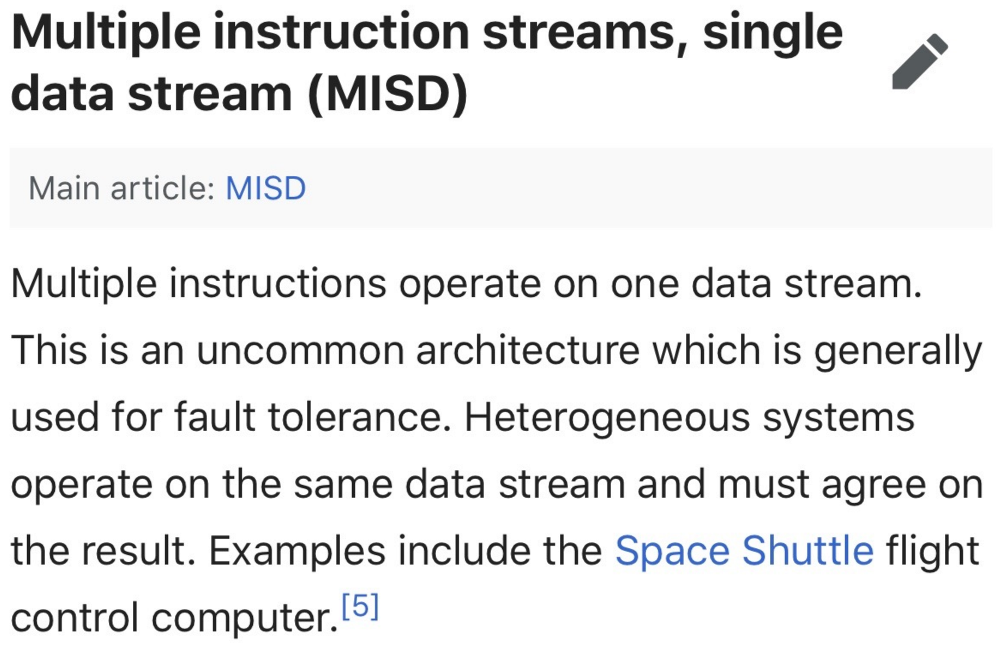
Developer Experience

- **Focus on the Customer**
- In Supercomputing, developers are often the customer
- Traditional HPC machines require long, in-advance reservations
- Cloud allows for gradual scaling, 24/7/365 availability

- *Enabling physicists / chemists / biologists / etc.. to experiment* ***is far more important to impact*** *than peak performance*

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**



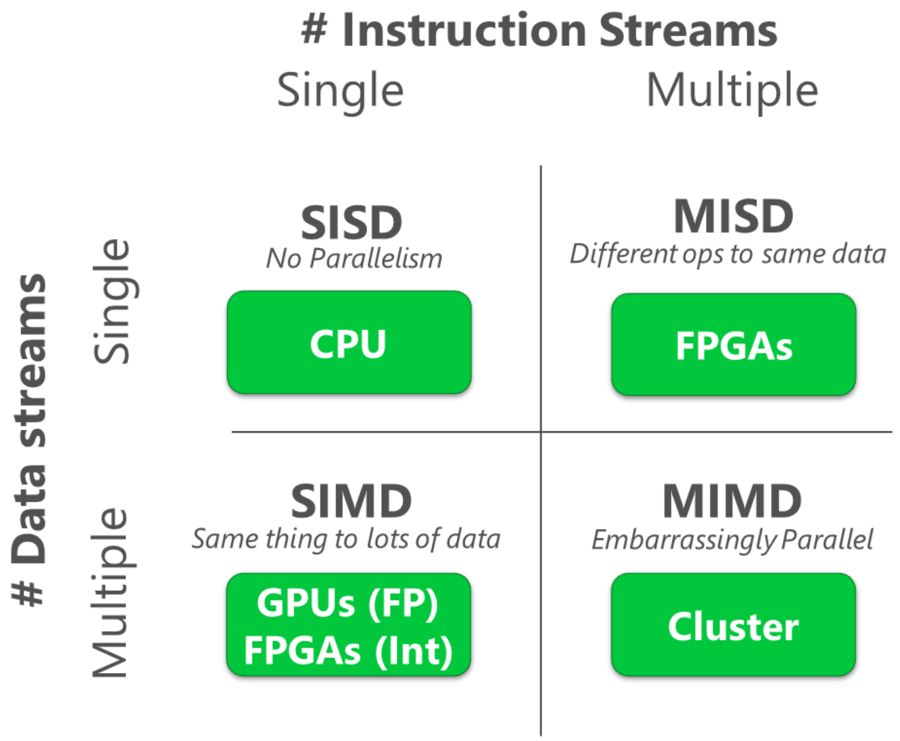
Multiple instruction streams, single data stream (MISD)

Main article: MISD

1

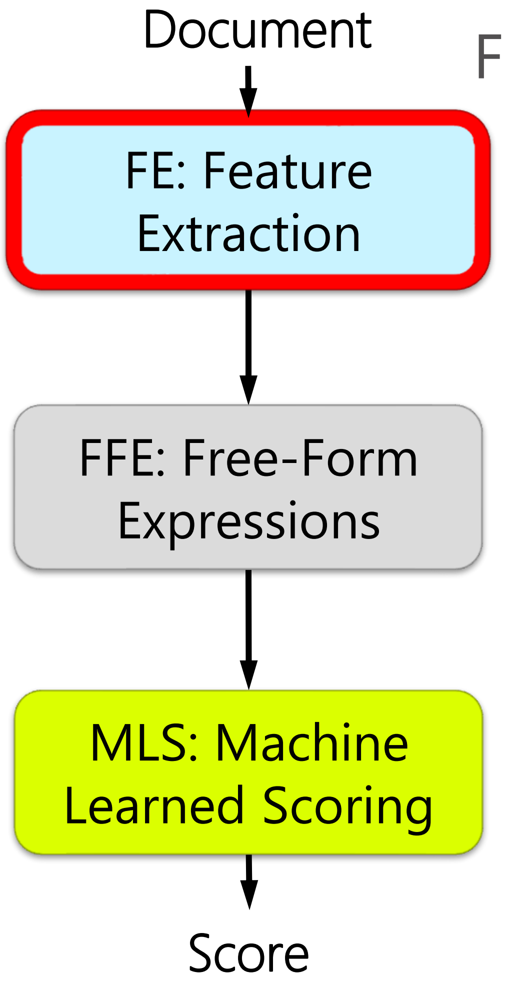
Multiple instructions operate on one data stream. This is an uncommon architecture which is generally used for fault tolerance. Heterogeneous systems operate on the same data stream and must agree on the result. Examples include the Space Shuttle flight control computer.[5]

27



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

FE: Feature Extraction



Document

FE: Feature Extraction

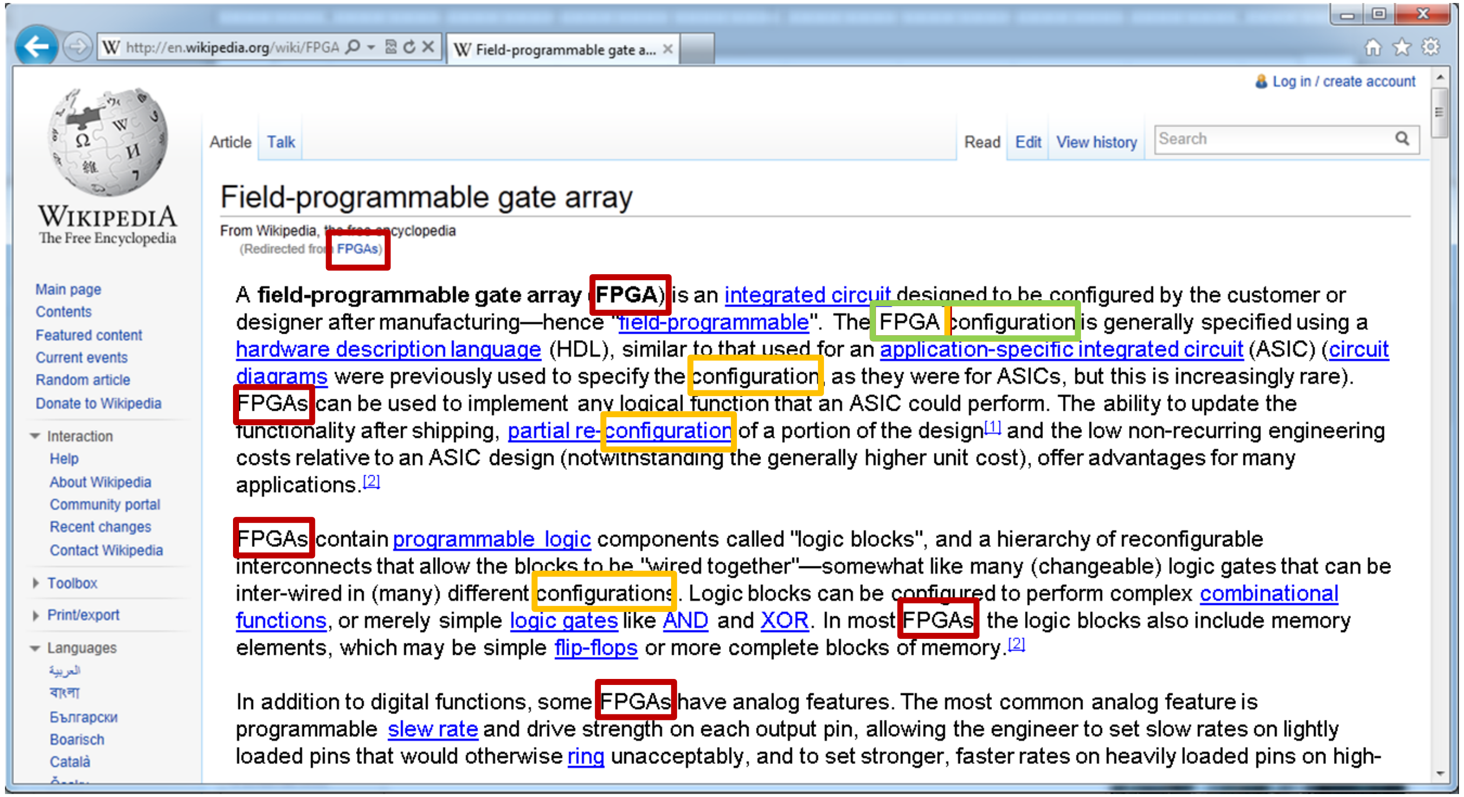
FFE: Free-Form Expressions

MLS: Machine Learned Scoring

Score

Query: “FPGA Configuration”

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



Field-programmable gate array

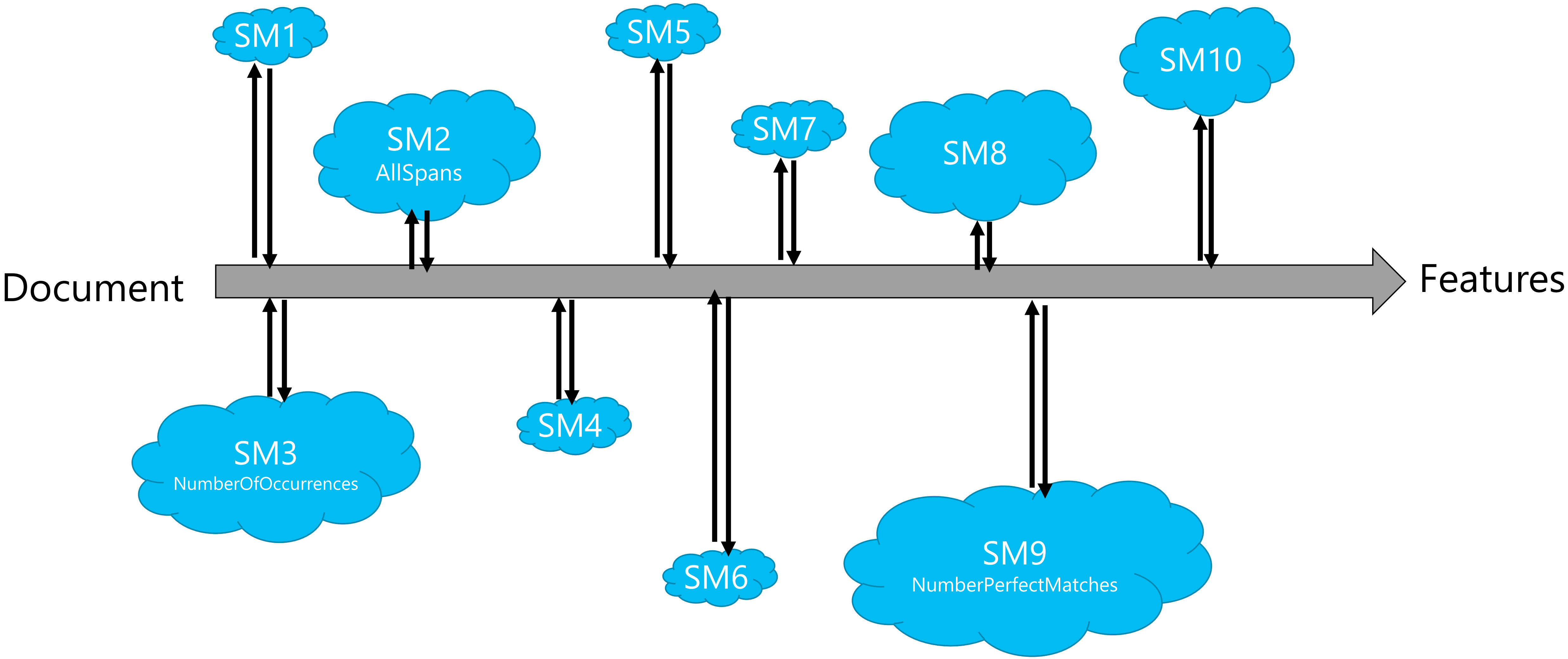
From Wikipedia, the free encyclopedia
(Redirected from FPGAs)

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 Cosmology

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

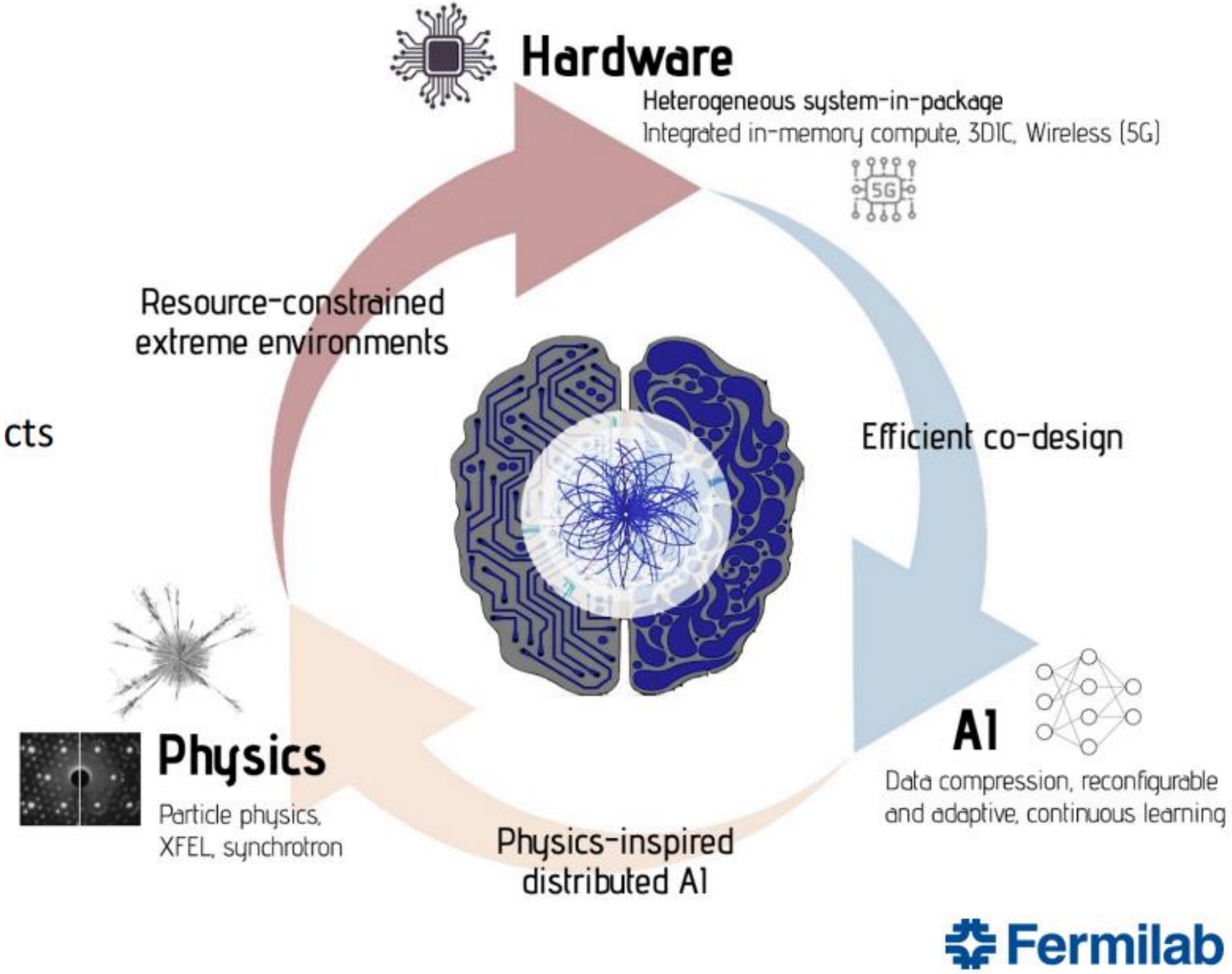
Cosmologists often refer to their telescopes as
“software telescopes”

FPGAs in Physics Applications

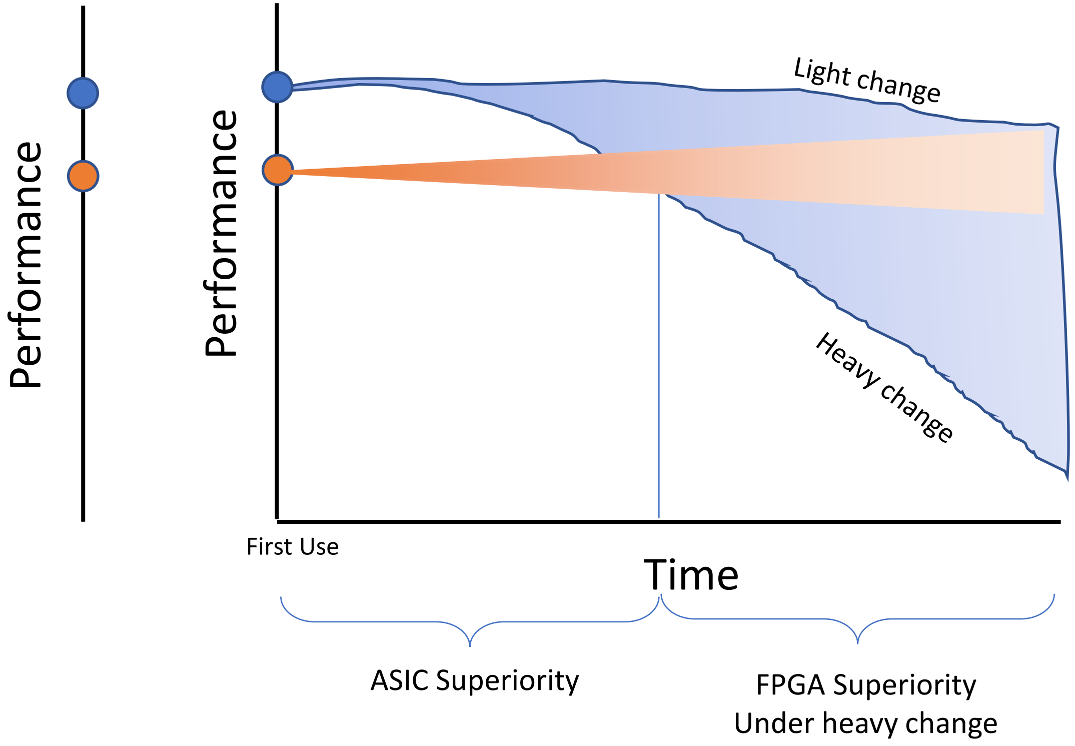
CERN

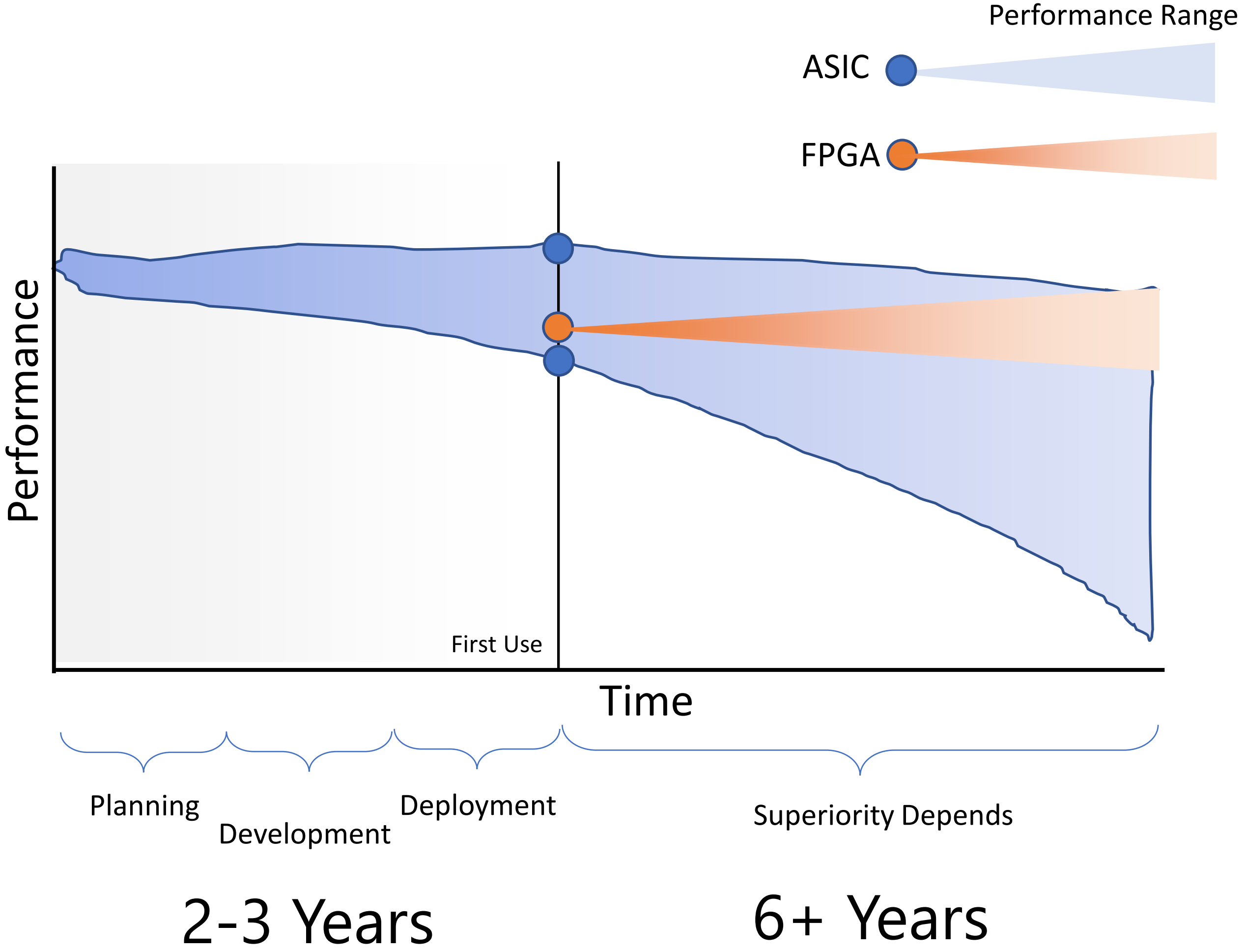
SETI

cts


Hardware
Heterogeneous system-in-package
Integrated in-memory compute, 3DIC, Wireless (5G)
5G
Resource-constrained extreme environments
Efficient co-design
Physics
Particle physics, XFEL, synchrotron
Physics-inspired distributed AI
AI
Data compression, reconfigurable and adaptive, continuous learning
Fermilab

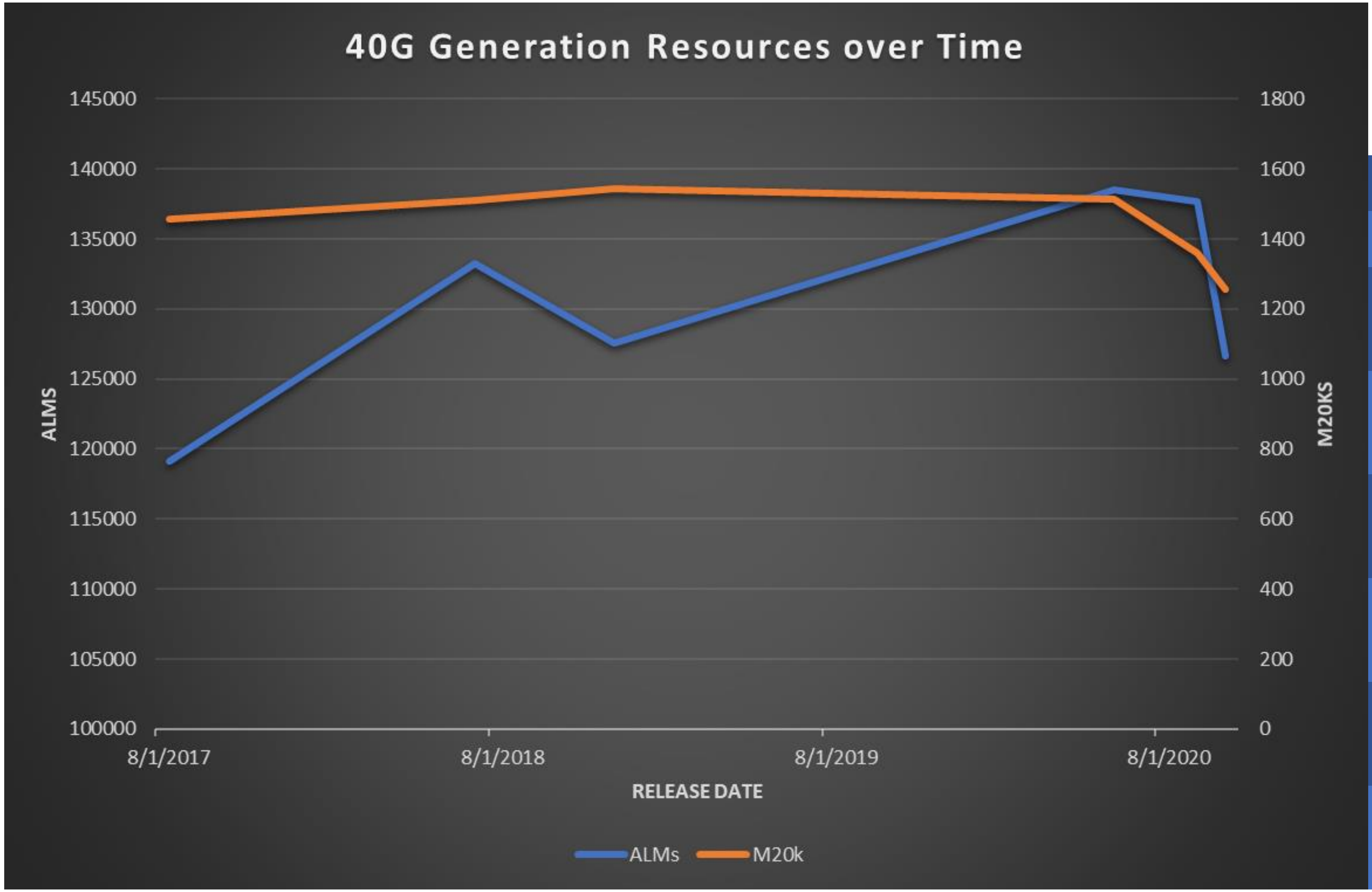
ASIC vs. FPGA





Requirement lock to decommission is over a decade
(as long as Azure itself has existed)
A lot changes over a decade

Resource Functionality Over Time for 40Gbps Generation



Pkts/sec	Description
22.5M PPS	
22.5M PPS	PFC Added
22.5M PPS	Fast Offload, new Lookup
22.5M PPS	PdParser, multi-tenancy, Flow Scaling to 4Million+
100M PPS	GFT-V2, 100MPPS, Shell Update
100M PPS	PCAP-V3, Filtering

Conclusion

- **Software is more important than hardware when you want to make an impact on the world**
- **The Cloud will replace dedicated supercomputers**
  - In large part due to developer experience
- **Think of FPGAs as a *complement* to GPGPUs, not just a competitor**
- **FPGAs play a role in all parts of the HPC stack**
- **High Flexibility enables a much longer lifetime, especially in new areas**

Microsoft