Microsoft

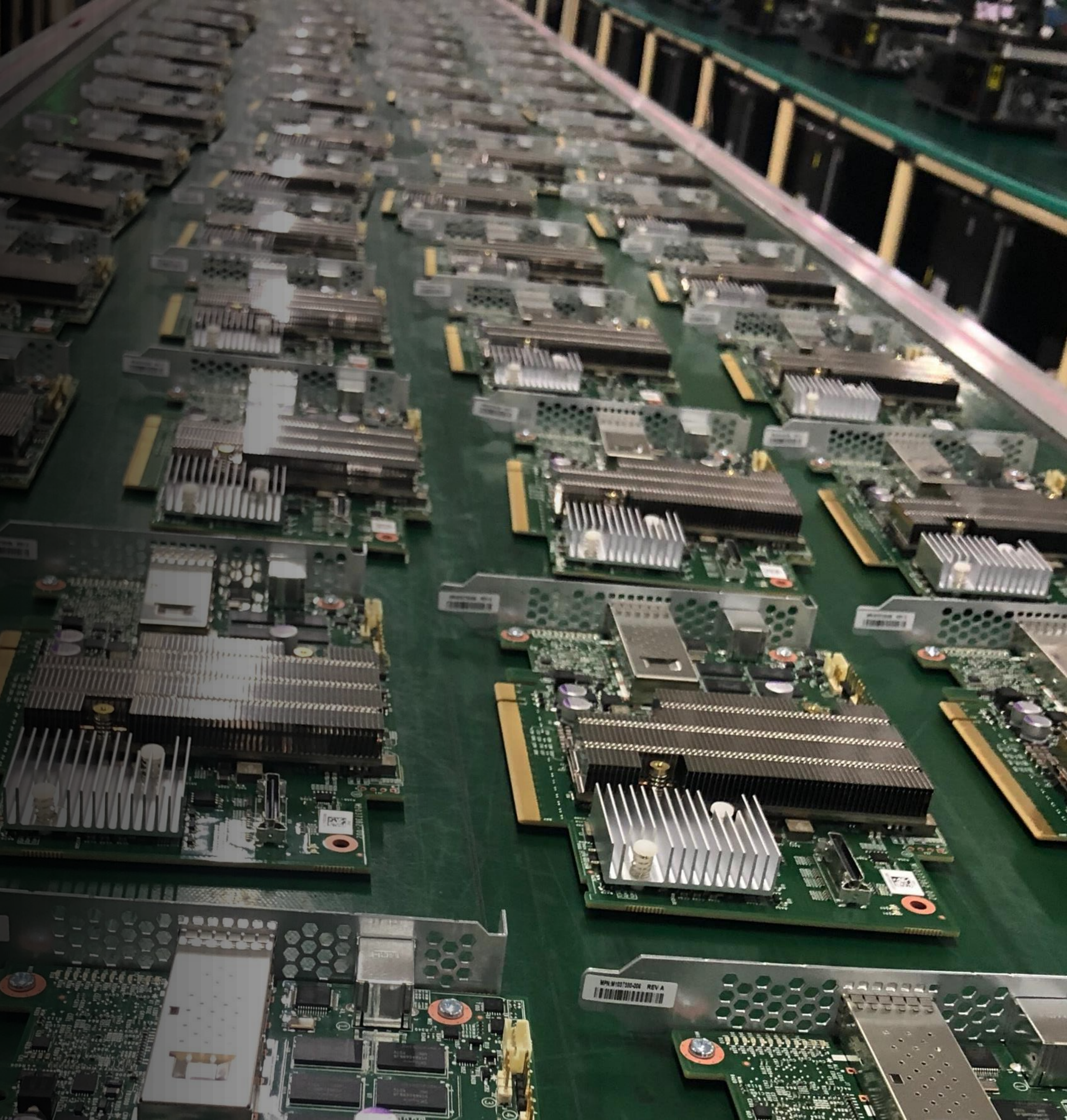
# Cloud and FPGAs in Physics and HPC

## from Data Acquisition to the Cloud

Andrew Putnam

July 22, 2022

Community Summer Study
SNOWMASS
July 17-26 2022, Seattle



# FPGAs in Cosmology



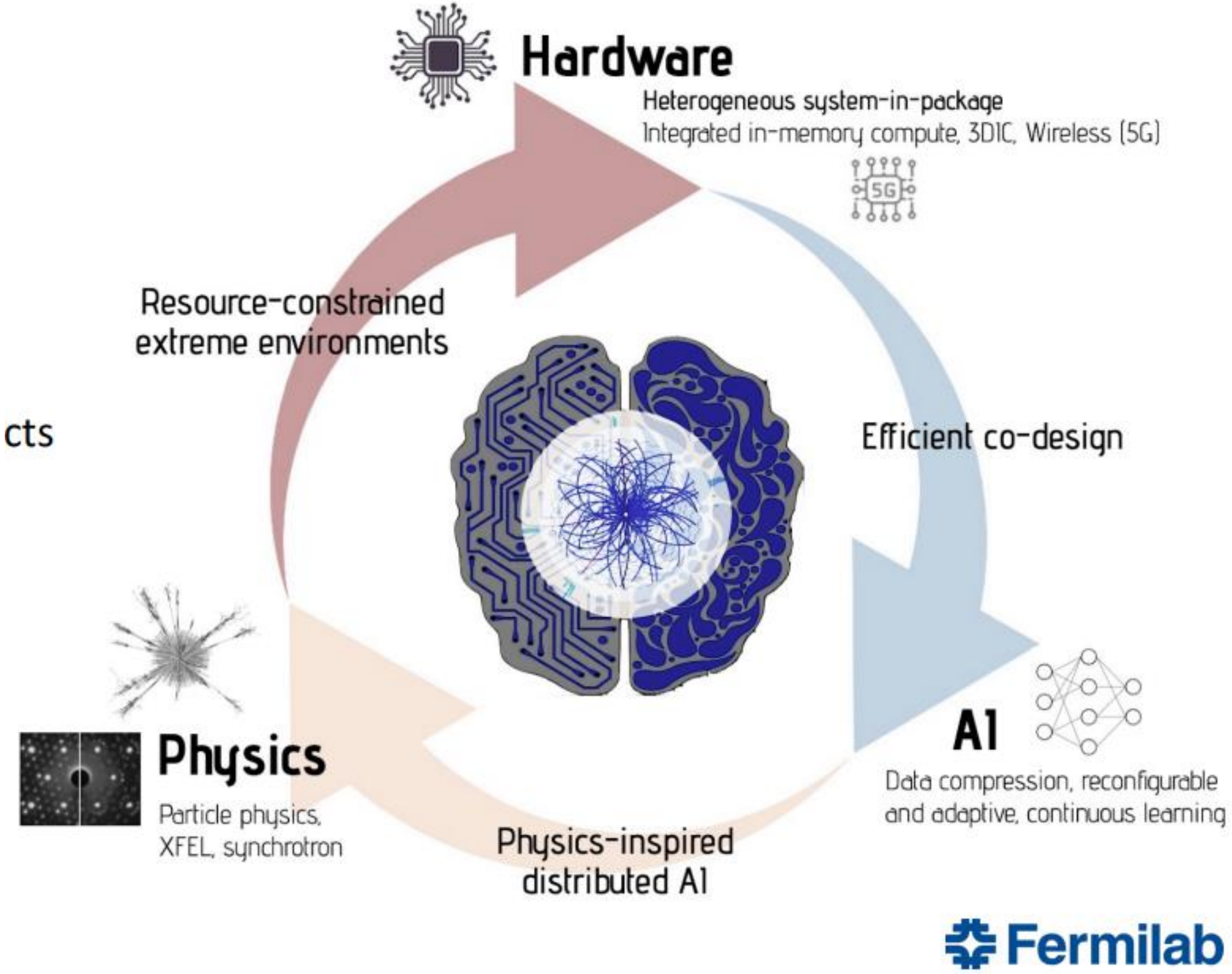
EOR Science can be done with a paperclip and a supercomputer
-- Don C. Backer

Cosmologists often refer to their telescopes as "software telescopes"

# FPGAs in Physics Applications


SETI

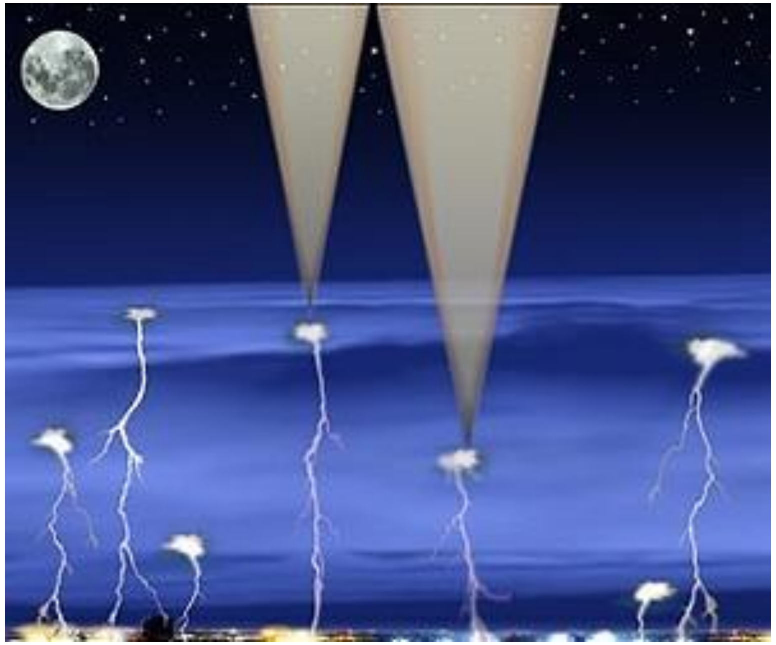
cts




Nhan Tran (FermiLab), Philip Harris (MIT), Javier Duarte (UCSD)

# Teaching Old Technology New Tricks

- Physicists are familiar for data acquisition and near-sensor processing
- You're going to have an FPGA developer on the project...
- But what else can you do with FPGAs?



# Catapult: Long, Fruitful FPGA Investment

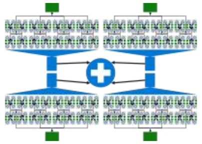
**Pre-History:**
May 2009: Bing Launched
Feb 2010: Azure Launched
Dec 2010: Catapult concept

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

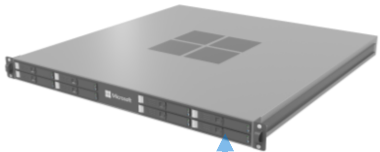
**v2: Pikes Peak**
Integrated Bing + Azure design
Bump-in-the-wire introduced

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


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



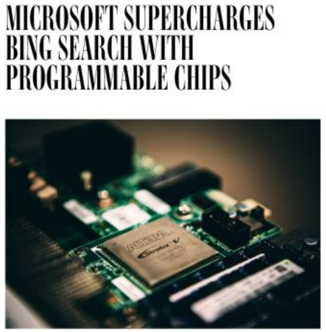
**Azure Databox Edge**
On-Site Inference



… 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




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

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

**Overlake: Celestial Peak**
100G w/ SoC
Networking + Storage

# Catapult: Long, Fruitful FPGA Investment

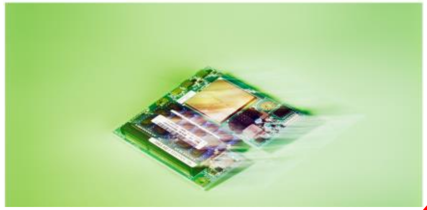
Cloud DNNs

Systolic Arrays and Feature Extraction

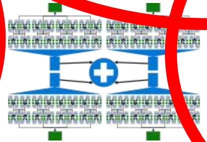
…ikes Peak
…rated
…-in-th…

SDN Offload

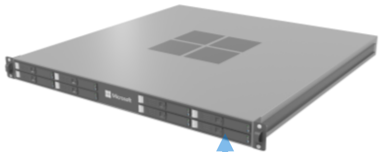
**Azure AccelNet Unveiled**
Azure production launch
…er demo



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



**Azure Databox Edge**
On-Site Inference



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

Designed for Decision Tree Scoring

| 2013 | 2014 | 2015 | 2016 | 2017 |
|---|---|---|---|---|

AI at the Edge

**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…**
DNN Platfo…
50Gb w/ int…

Hypervisor Offload

…Peak
…e

# Dominant state-of-the-art models evolving rapidly

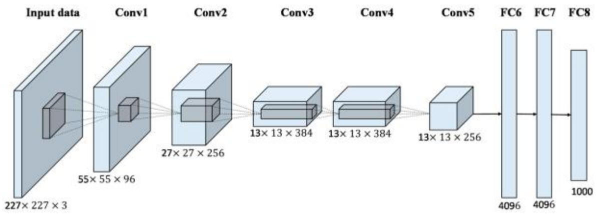
Classic ML — Deep CNNs (2012) — LSTMs/ GRUs — Transformers (2019+)

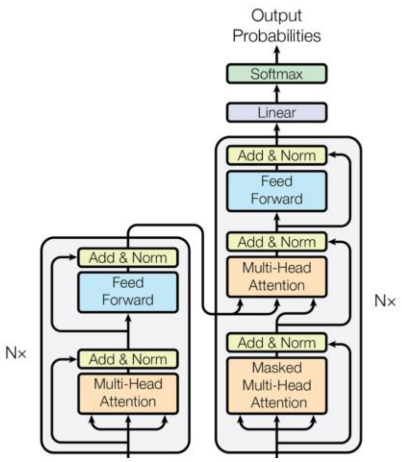
*Sparse transformers?*

*Sparse mixture of experts?*

*Dynamic, recursive networks?*

*Graph neural networks?*



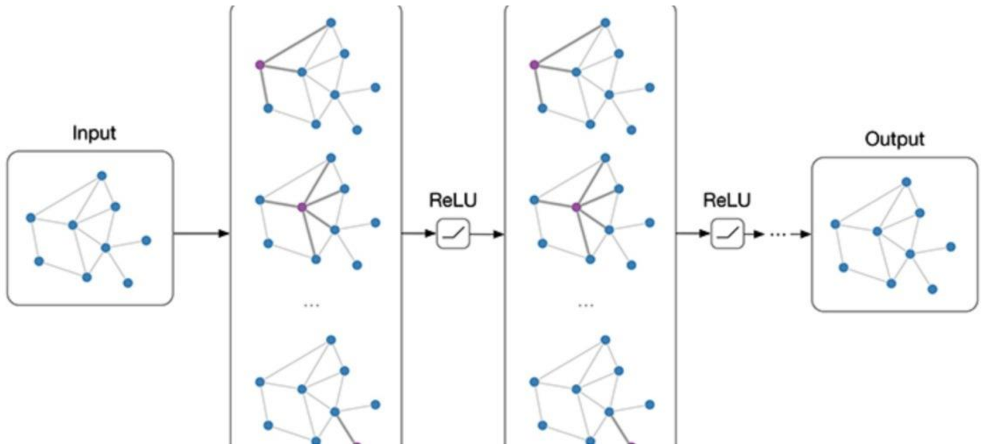
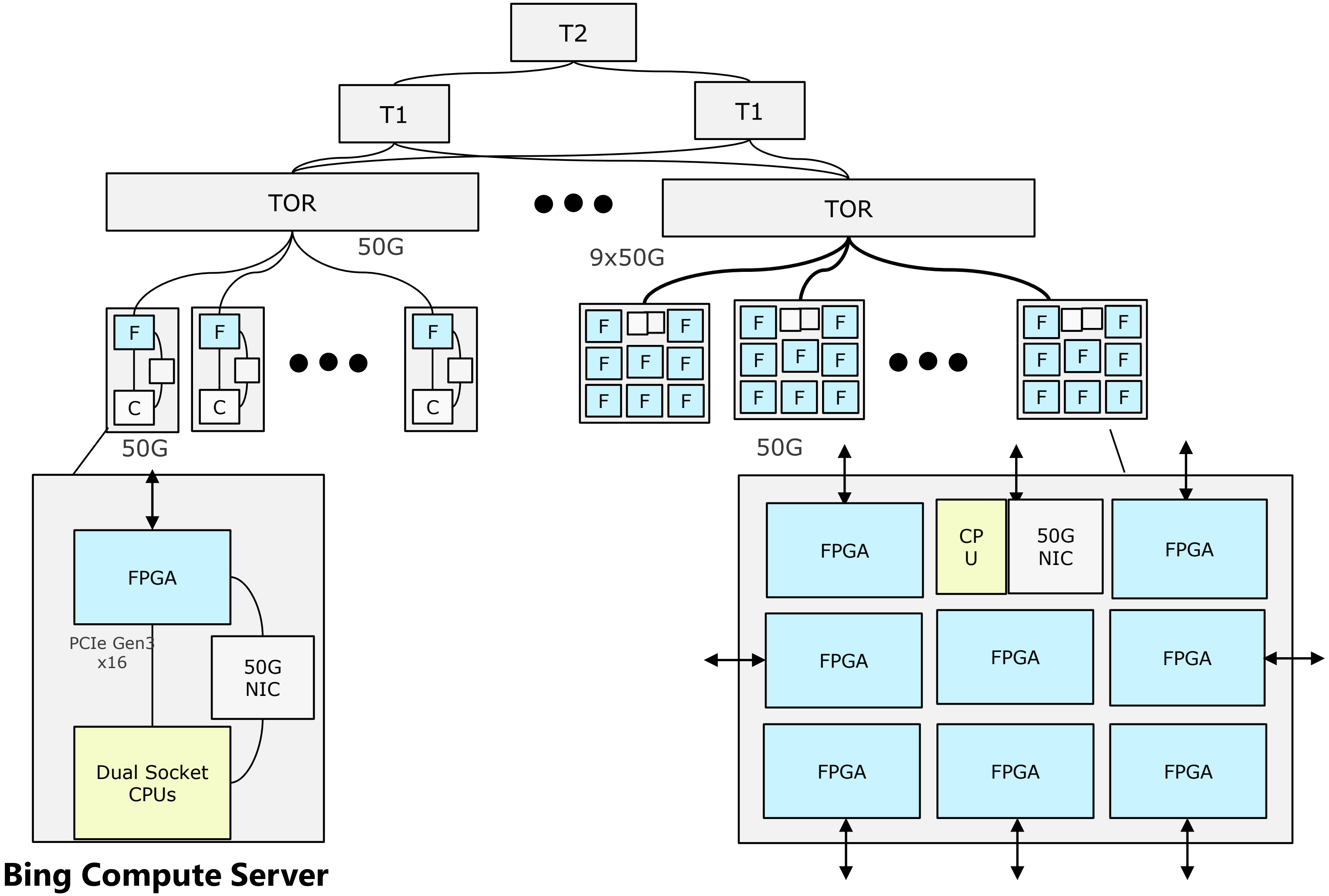




Figure sources:
1. Han et al., Pre-Trained AlexNet Architecture with Pyramid Pooling and Supervision for High Spatial Resolution Remote Sensing Image Scene Classification
2. Vaswani et al., “Attention is all you need”
3. https://tkipf.github.io/graph-convolutional-networks/

# Rapid Iteration and Deployment

| Brainwave v1 | Brainwave v2 | Brainwave v3 | Brainwave v4 |
| --- | --- | --- | --- |
| Low latency LSTM inference | Narrow Precision Breakthrough | Convolution Optimizations | Generalized ISA, Transformers |
| 2016 | 2017 | 2018 | 2019 |

# HyperScale -- Bing's 500 Petaflops Inference Supercomputer




Bing Compute Server

Bing FPGA Appliances

# Slick new hardware

# Useful hardware

# What makes a public cloud company successful?

HUAWEI
IBM
NVIDIA
intel
SK hynix
AMD
SAMSUNG
BROADCOM

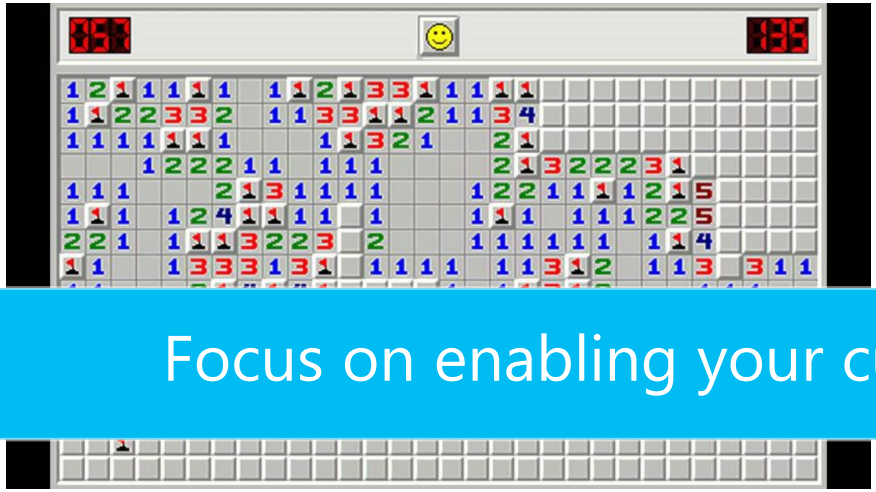
Hardware Companies

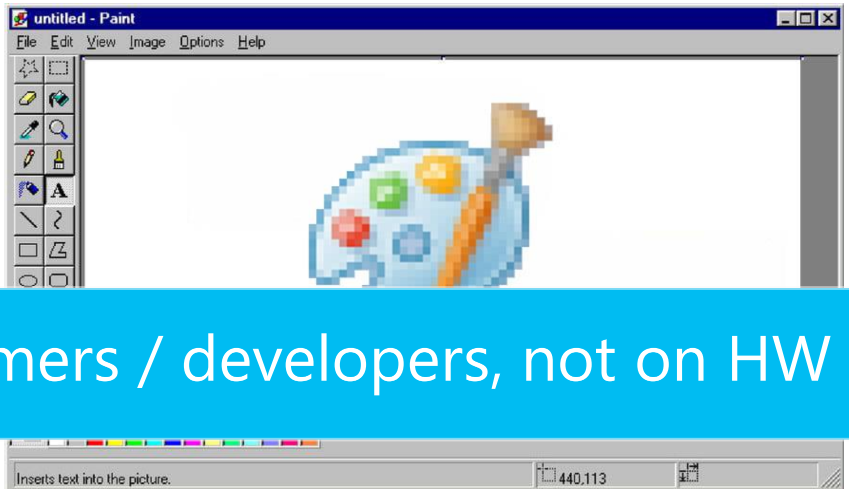
amazon
Baidu 百度
Microsoft
Tencent 腾讯
Google
Alibaba.com
HUAWEI

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



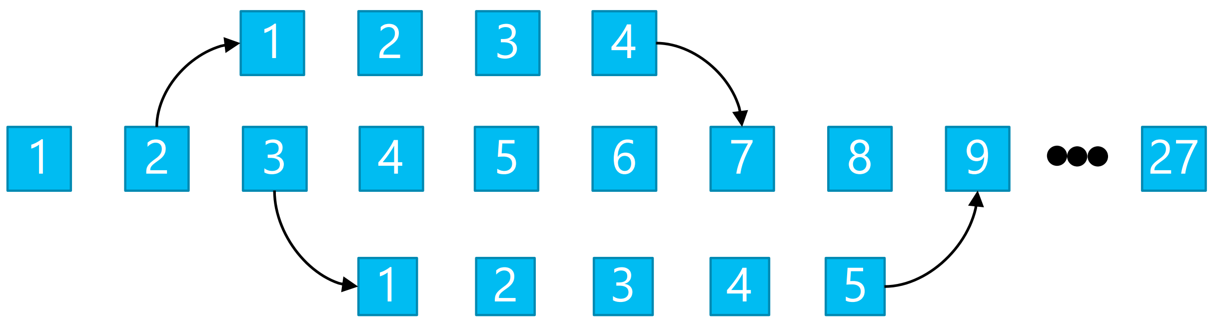


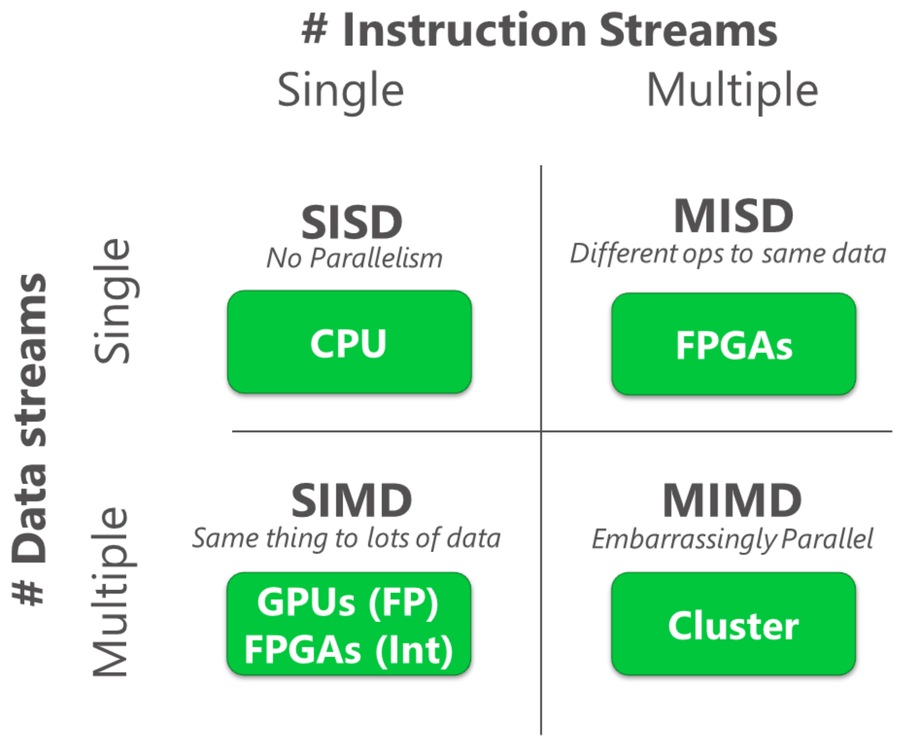

Focus on enabling your customers / developers, not on HW

# 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**

Pipeline Parallelism



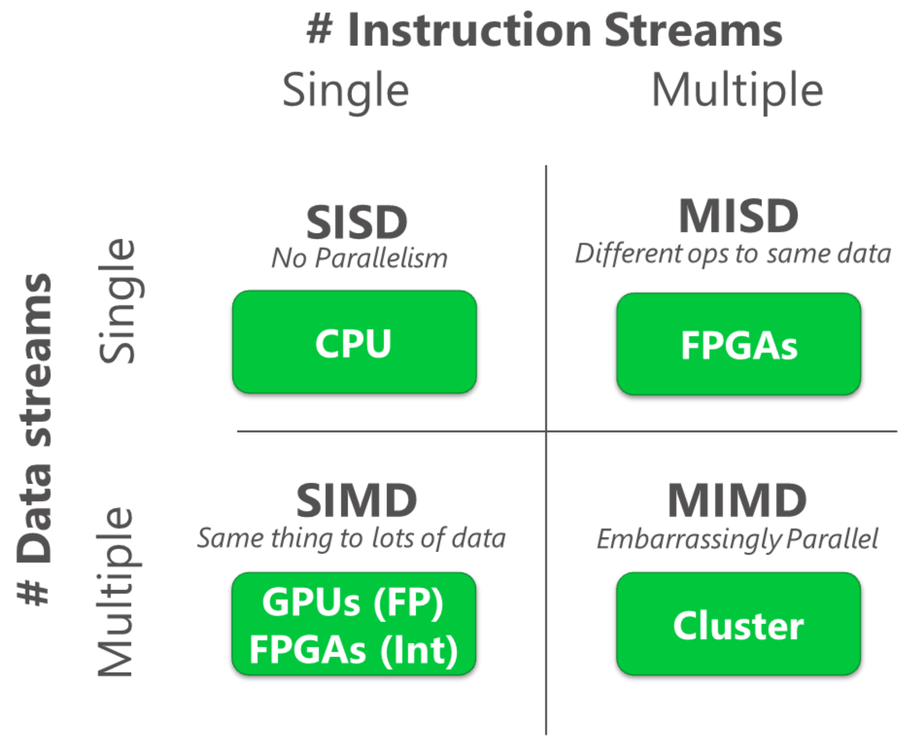
# Why is the FPGA a good choice as an accelerator?

- **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**

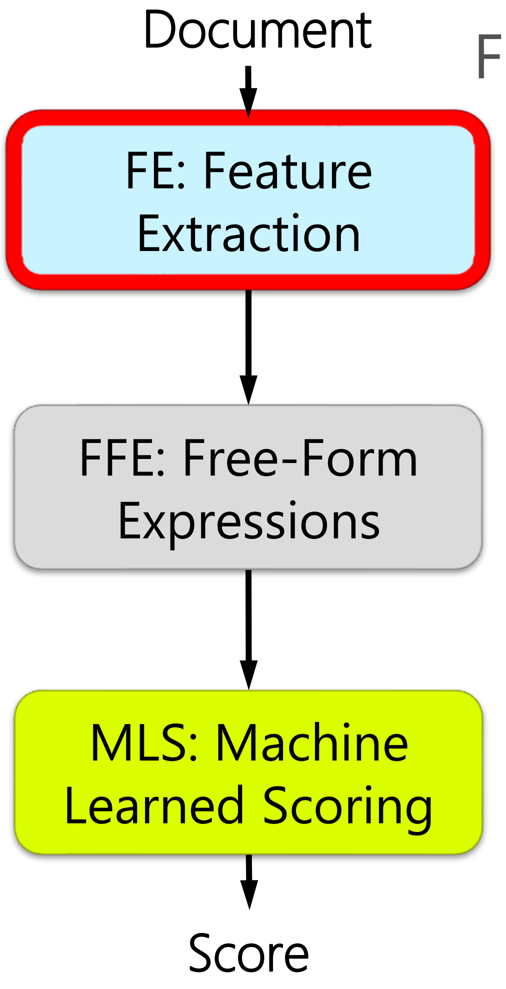
**Multiple instruction streams, single data stream (MISD)**

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

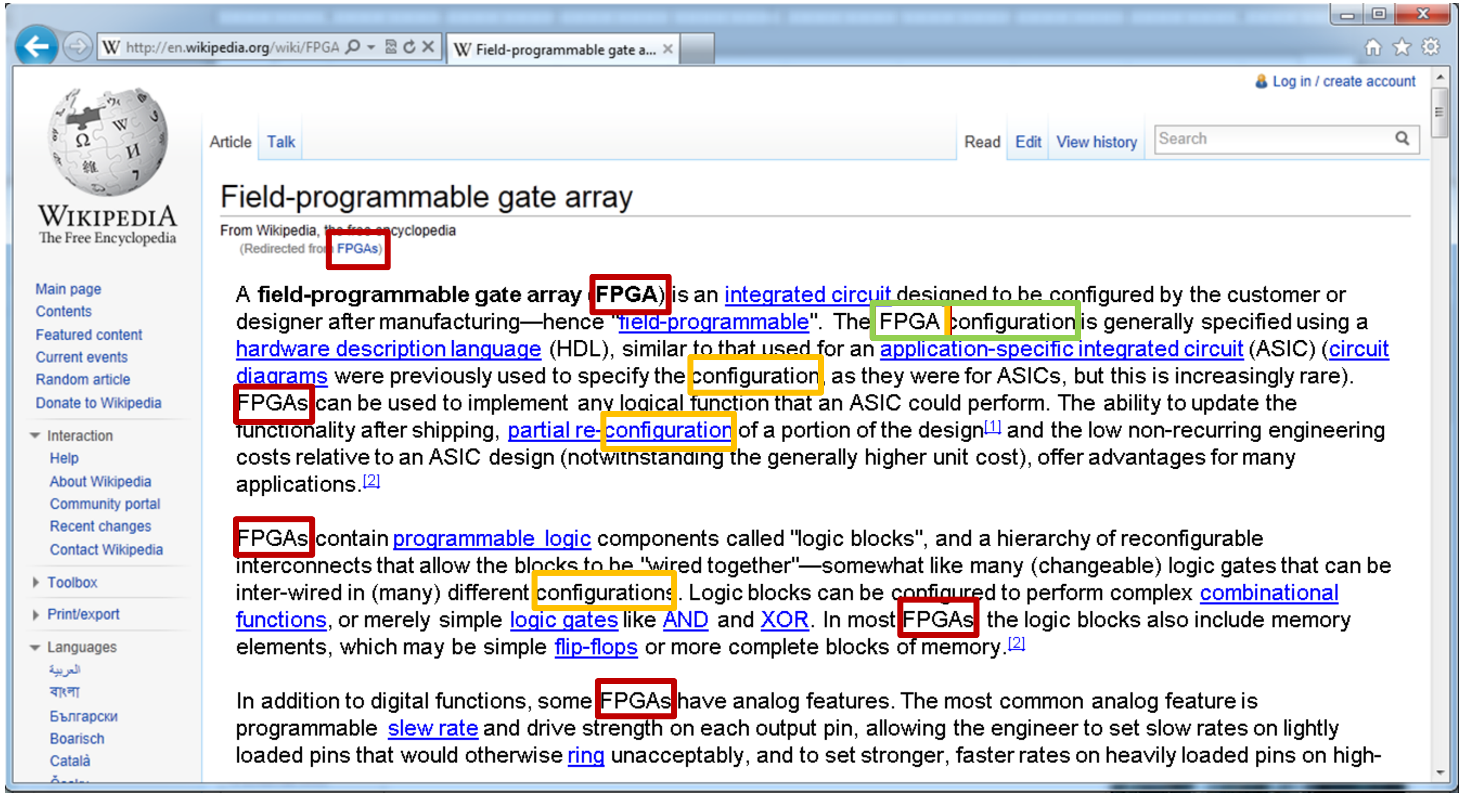


# FE: Feature Extraction

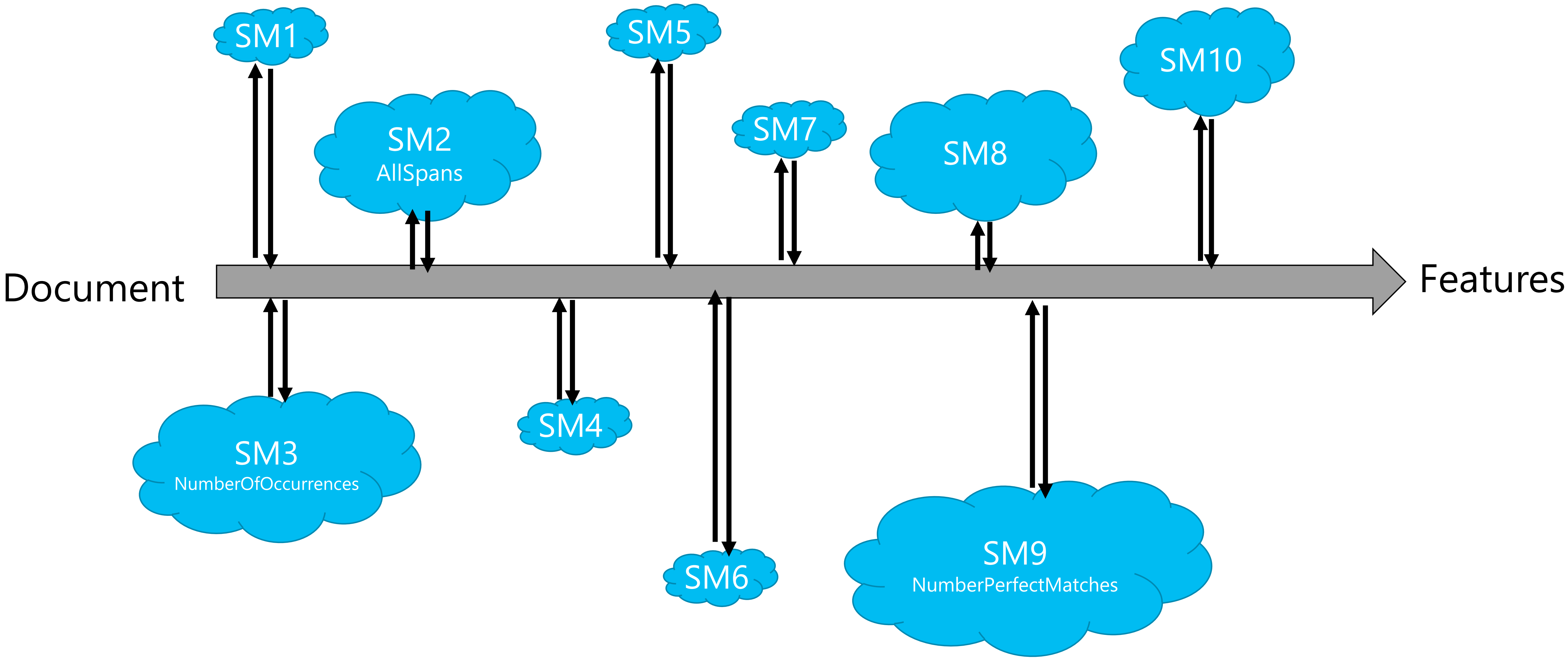



**Query: “FPGA Configuration”**

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

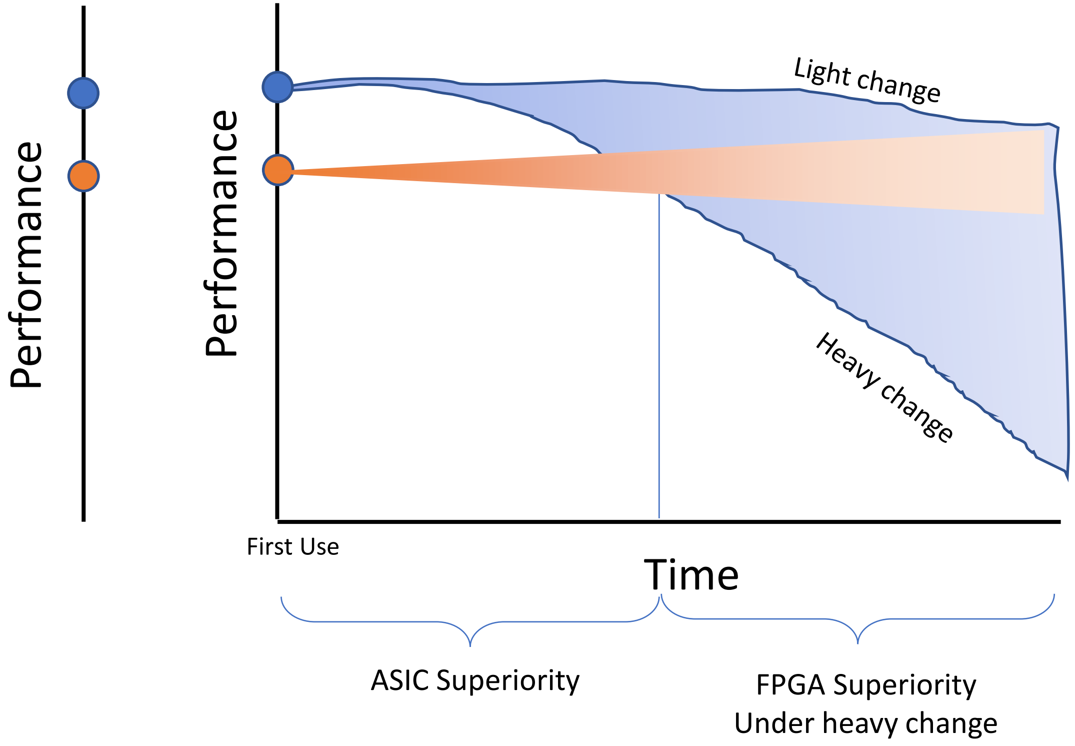


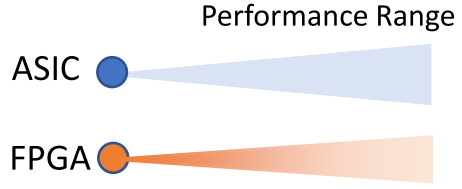
# Feature Extraction Accelerator



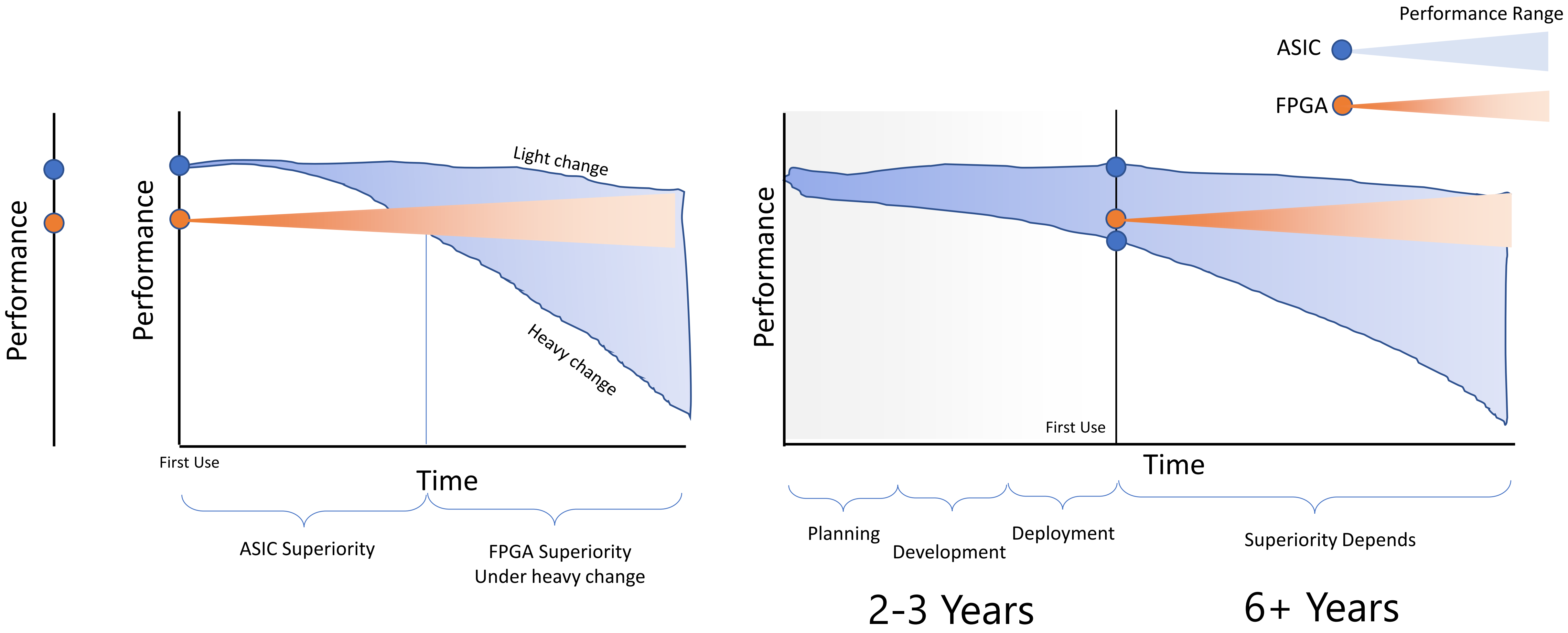
# ASIC vs. FPGA

# ASIC vs. FPGA


Performance
Performance
Light change
Heavy change
First Use
Time
ASIC Superiority
FPGA Superiority
Under heavy change

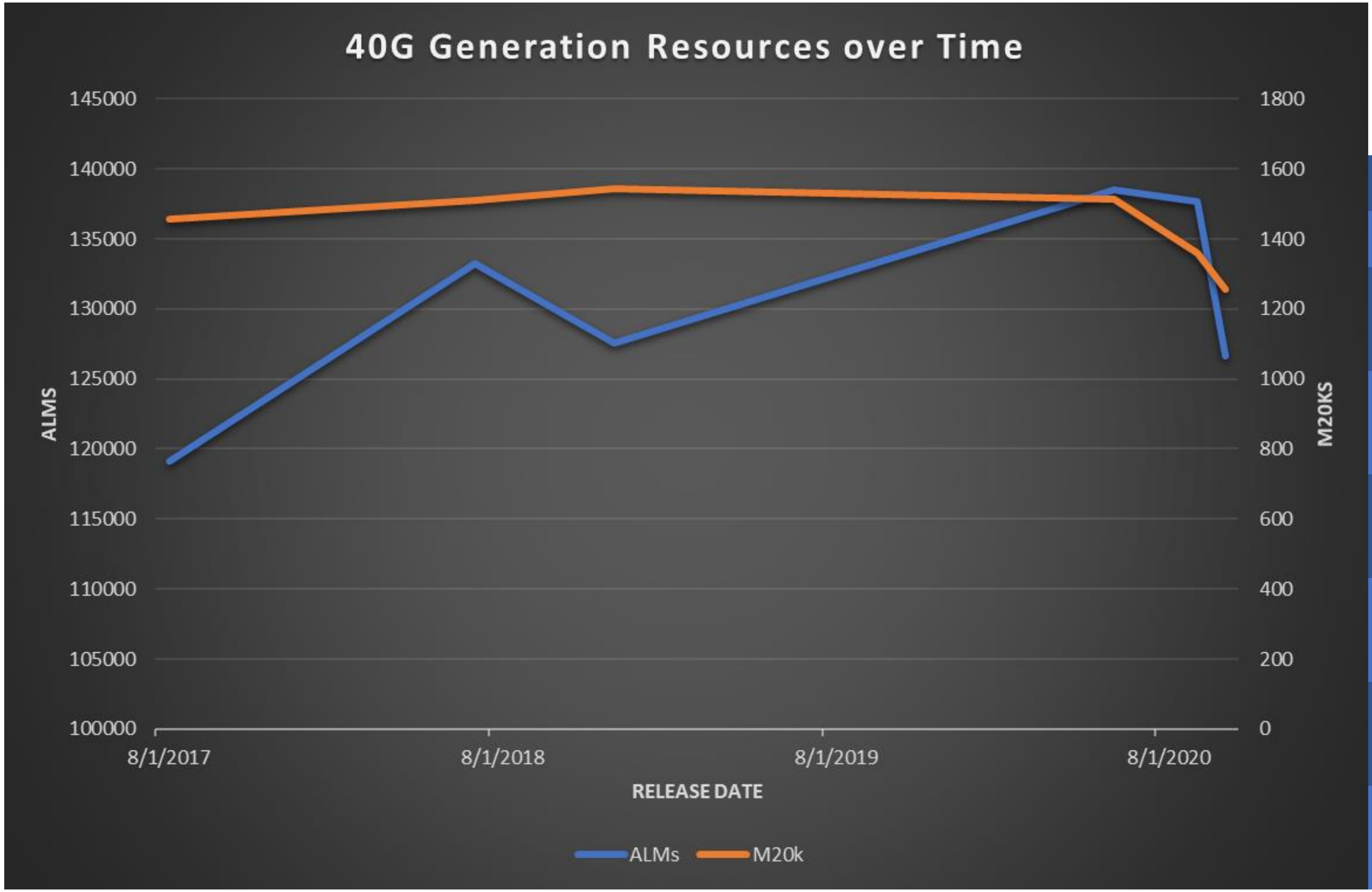

Performance Range
ASIC
FPGA

# 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 |

# 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<br>RIKEN Center for Computational Science<br>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<br>DOE/SC/Oak Ridge National Laboratory<br>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<br>DOE/NNSA/LLNL<br>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<br>National Supercomputing Center in Wuxi | 10,649,600 | 93,014.6 | 125,435.9 | 15,371 |
| 5 | | | | | |
| 6 | | | | | |
| 7 | National Super Computer Center in Guangzhou<br>China | | | | |
| 8 | **JUWELS Booster Module** - Bull Sequana XH2000 , AMD EPYC 7402 24C 2.8GHz, NVIDIA A100, Mellanox HDR InfiniBand/ParTec ParaStation ClusterSuite, Atos<br>Forschungszentrum Juelich (FZJ)<br>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<br>Eni S.p.A.<br>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<br>Azure East US 2<br>United States | 253,440 | 30,050.0 | 39,531.2 | |

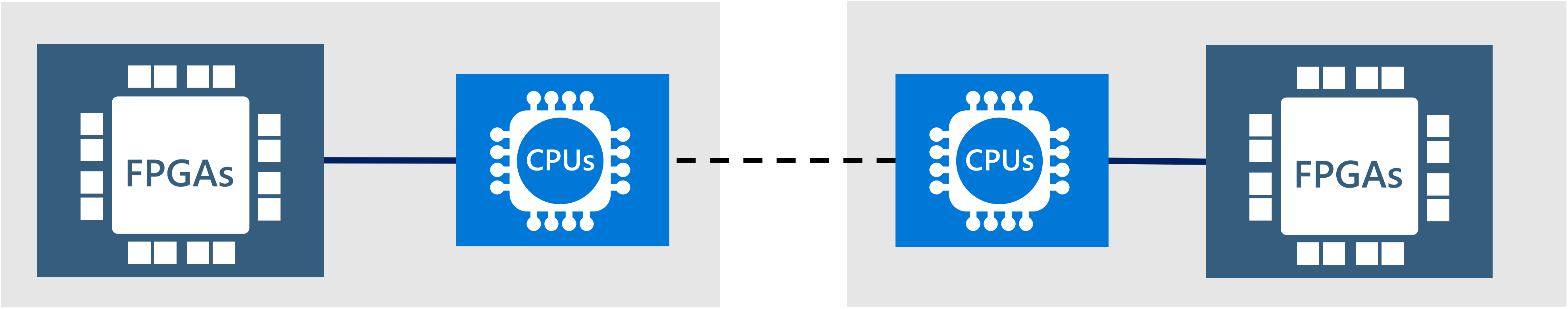
10 **Voyager-EUS2** - ND96amsr_A100_v4, AMD EPYC 7V12 48C 2.45GHz, NVIDIA A100 80GB, Mellanox HDR Infiniband, Microsoft Azure
Azure East US 2
United States
253,440 30,050.0 39,531.2

# Why hasn't Supercomputing moved to the Cloud?

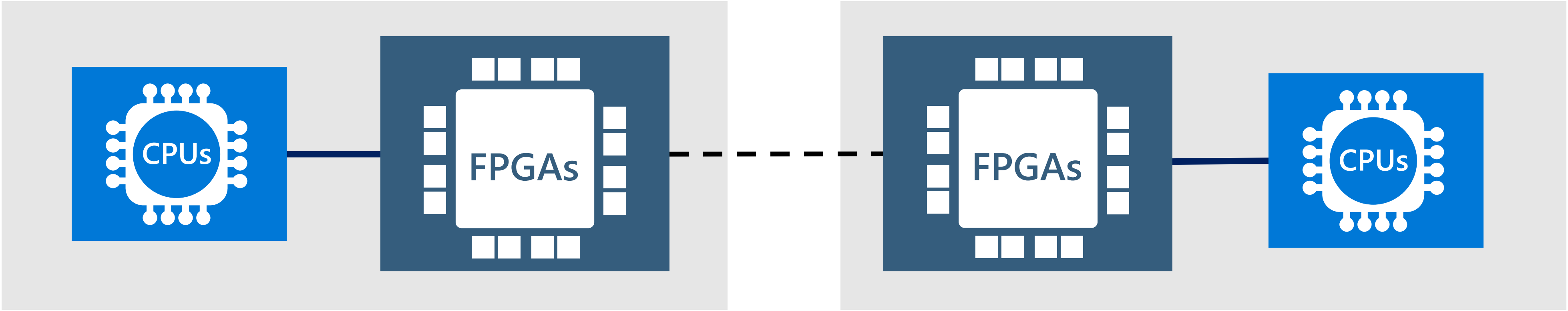
CPUs look largely the same, but…

- ☑ 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

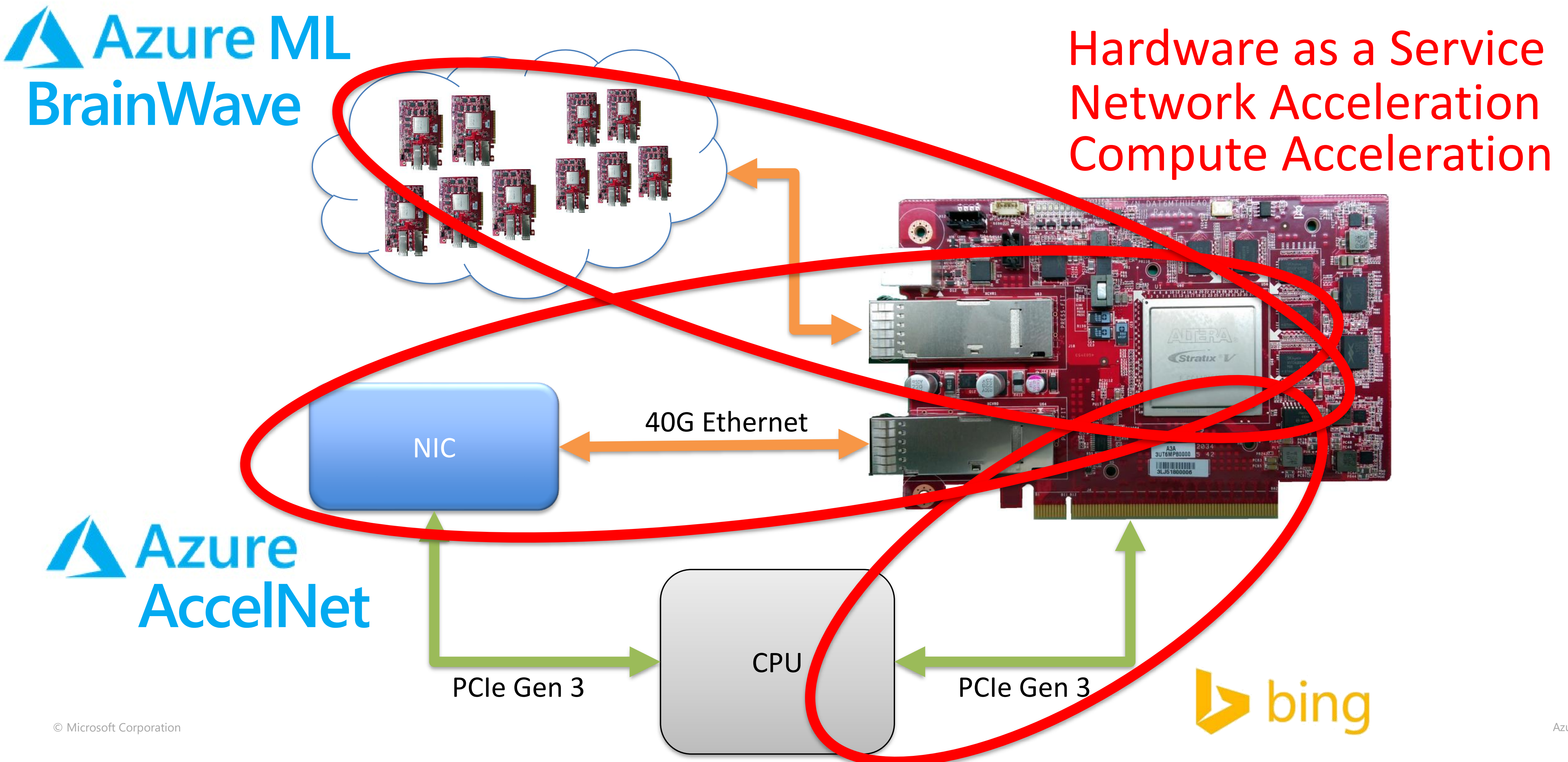



Traditional Accelerator Integration

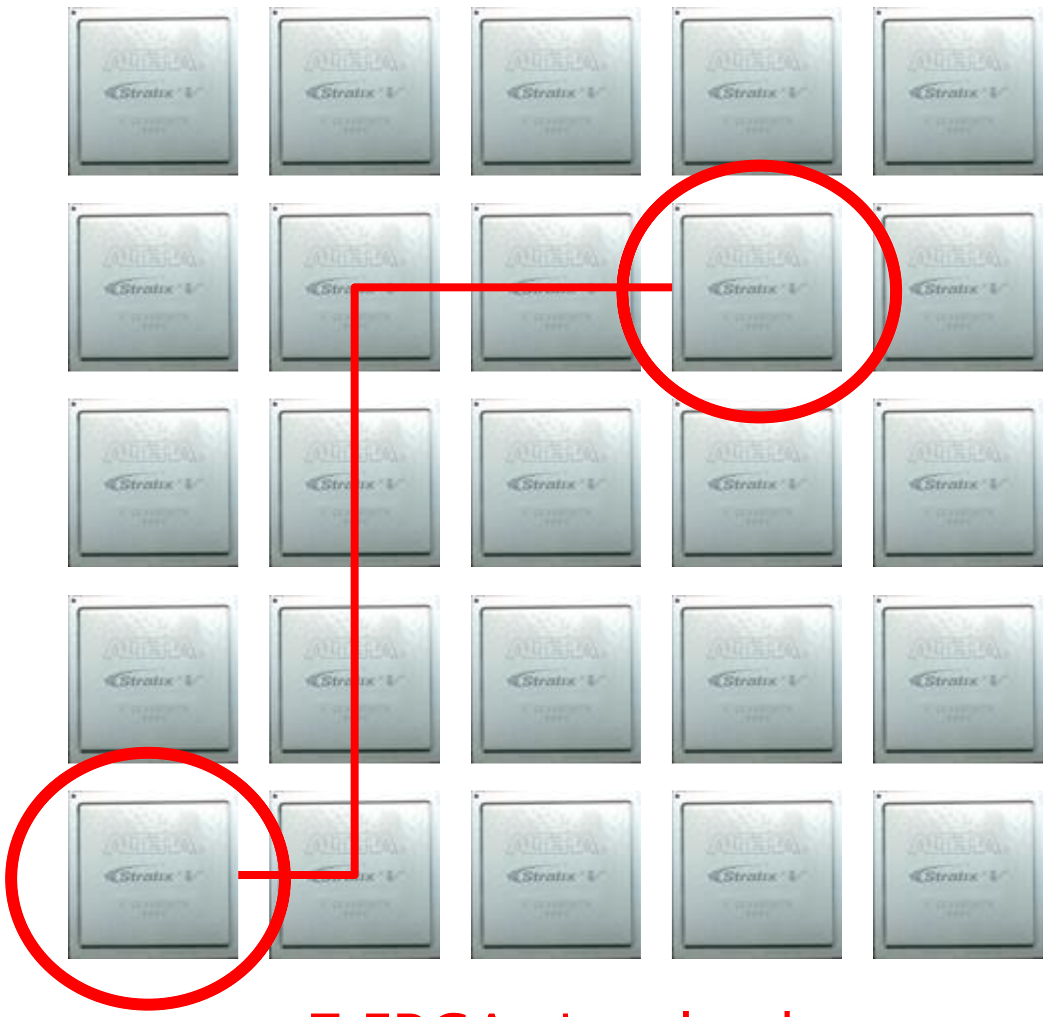



Bump in the Wire -- In-Network Acceleration

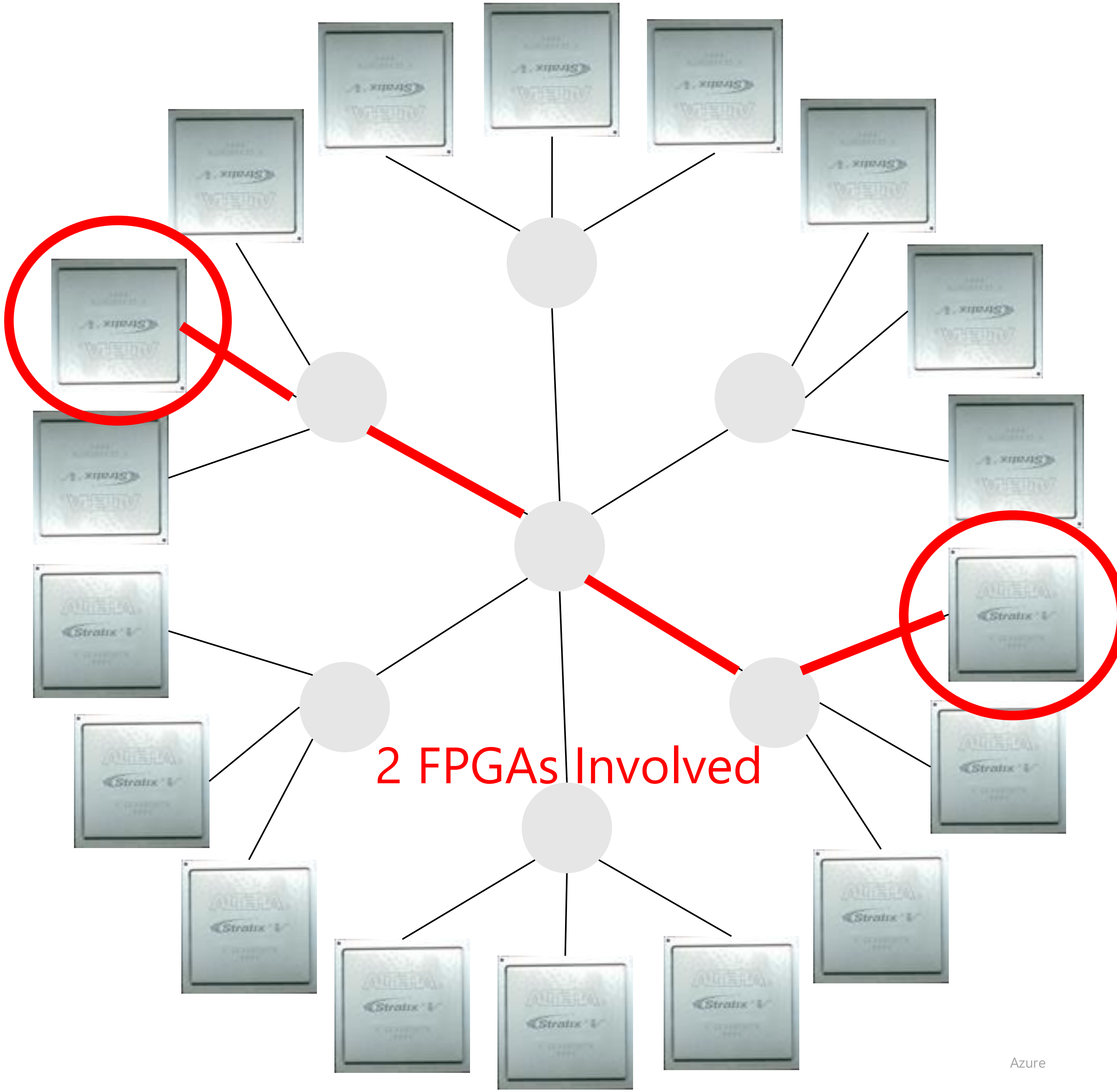
# Bump-in-the-wire Architecture



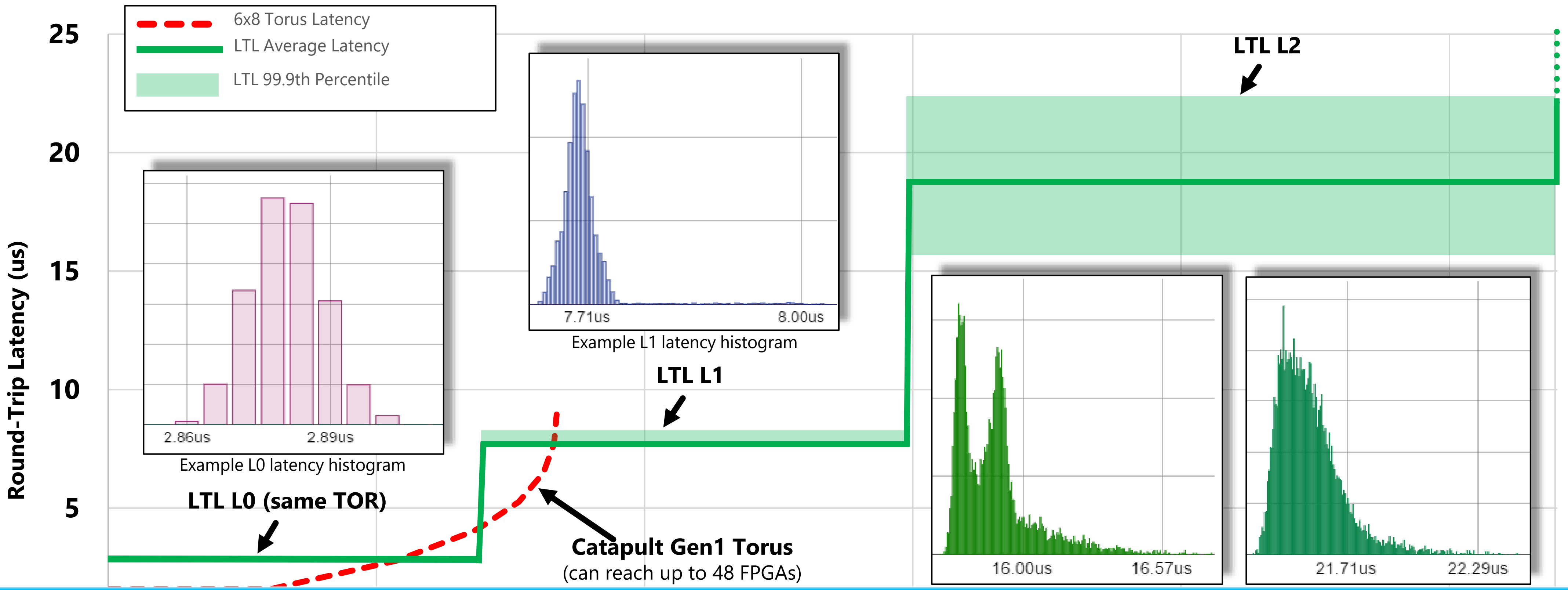
# Global-Scale FPGA



7 FPGAs Involved



# Network Latencies



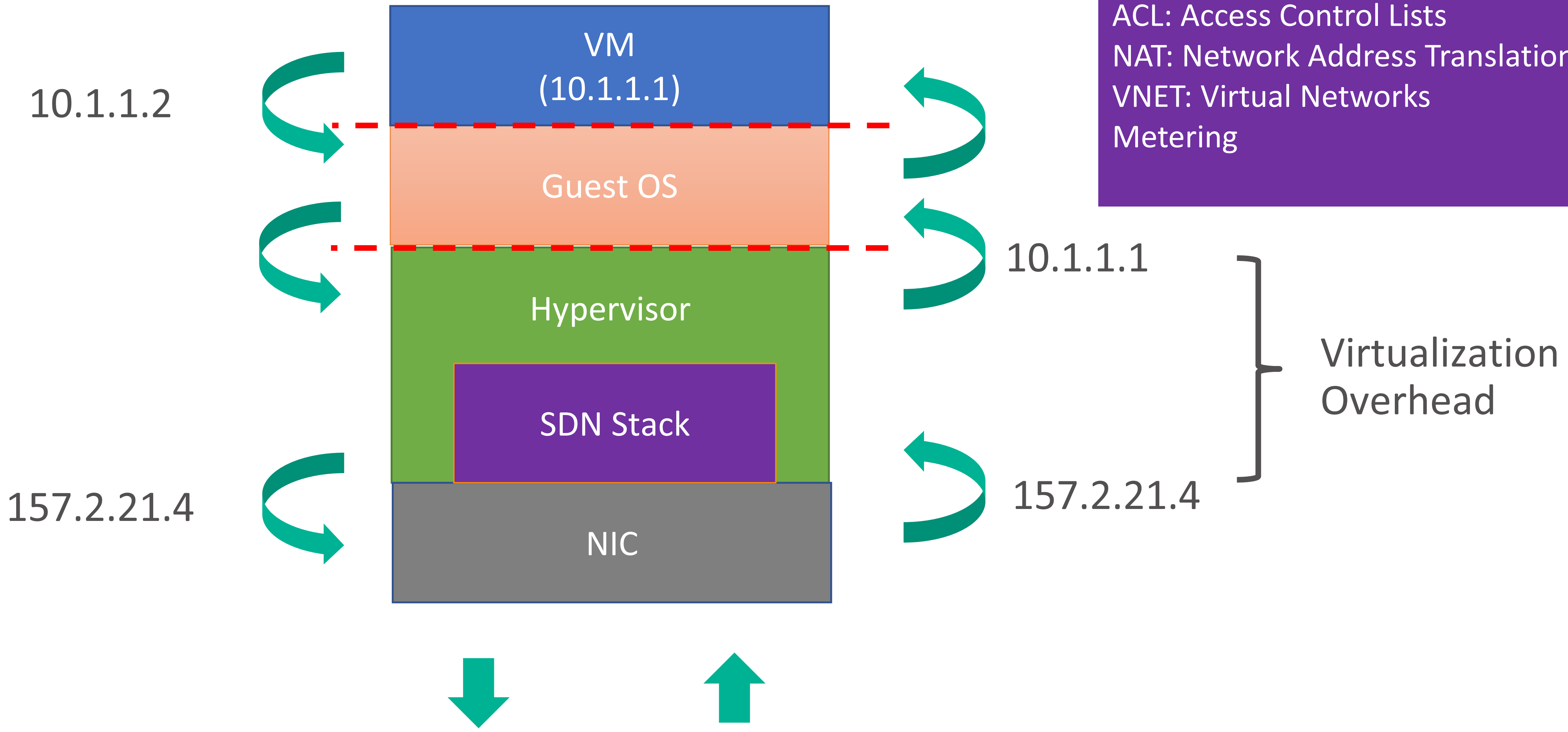

- Extremely low latency (Similar to Infiniband)
- Global-scale FPGA

# Why hasn't Supercomputing moved to the Cloud?

CPUs look largely the same, but…

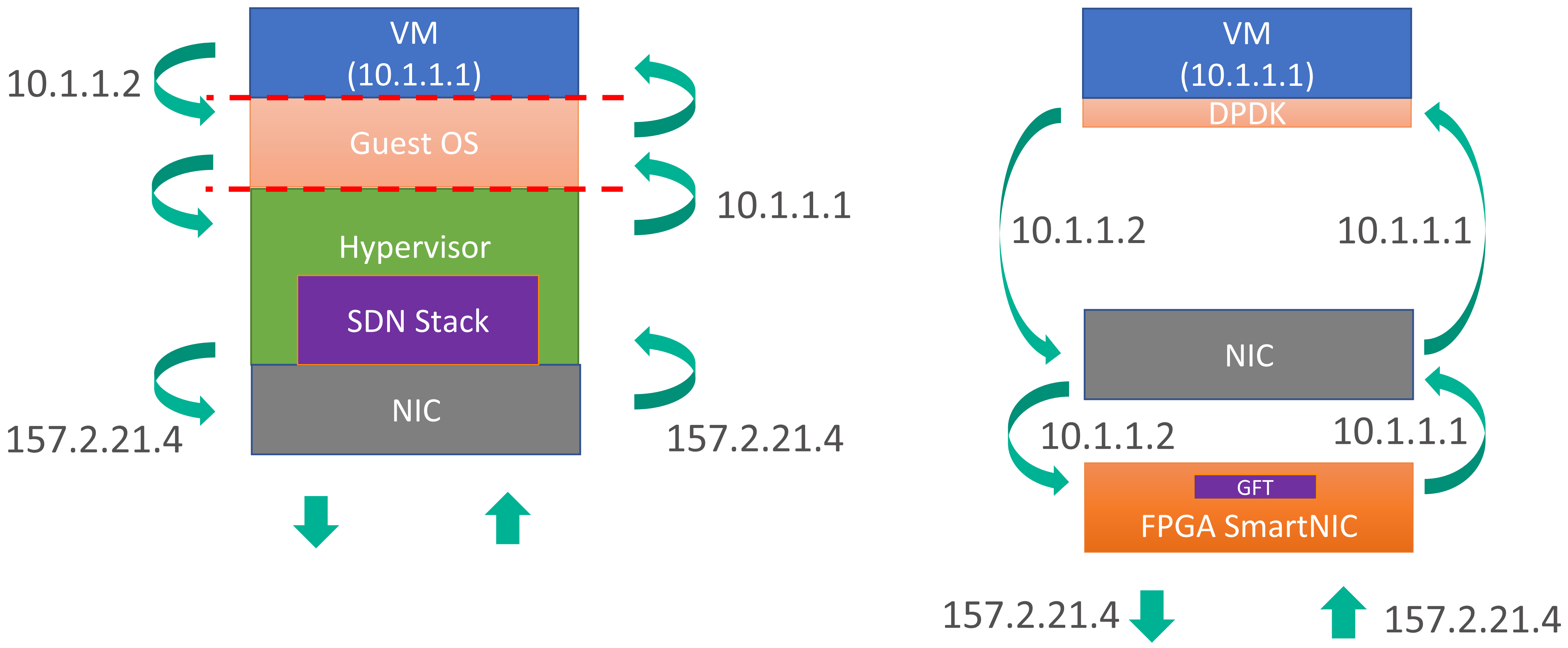
- [x] Top 500 often include specialized accelerators (especially GPUs)
- [x] Networks are highly specialized, tuned for low-latency, high bandwidth
- [ ] Won't running virtual machines kill performance?

# Virtualization Overhead – Standard Virtual Machines




<u>SDN Stack Functions</u>
SLB: Software Load Balancers
ACL: Access Control Lists
NAT: Network Address Translation
VNET: Virtual Networks
Metering

# Virtualization Overhead – SmartNICs & Bump-in-the-Wire



# 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?

What will *really* make 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*

# Conclusion

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

Microsoft