Microsoft

Hardware Acceleration in Hyperscale Cloud Infrastructures

Doug Burger

Technical Fellow, Advanced Intelligent Architectures,
Microsoft Azure

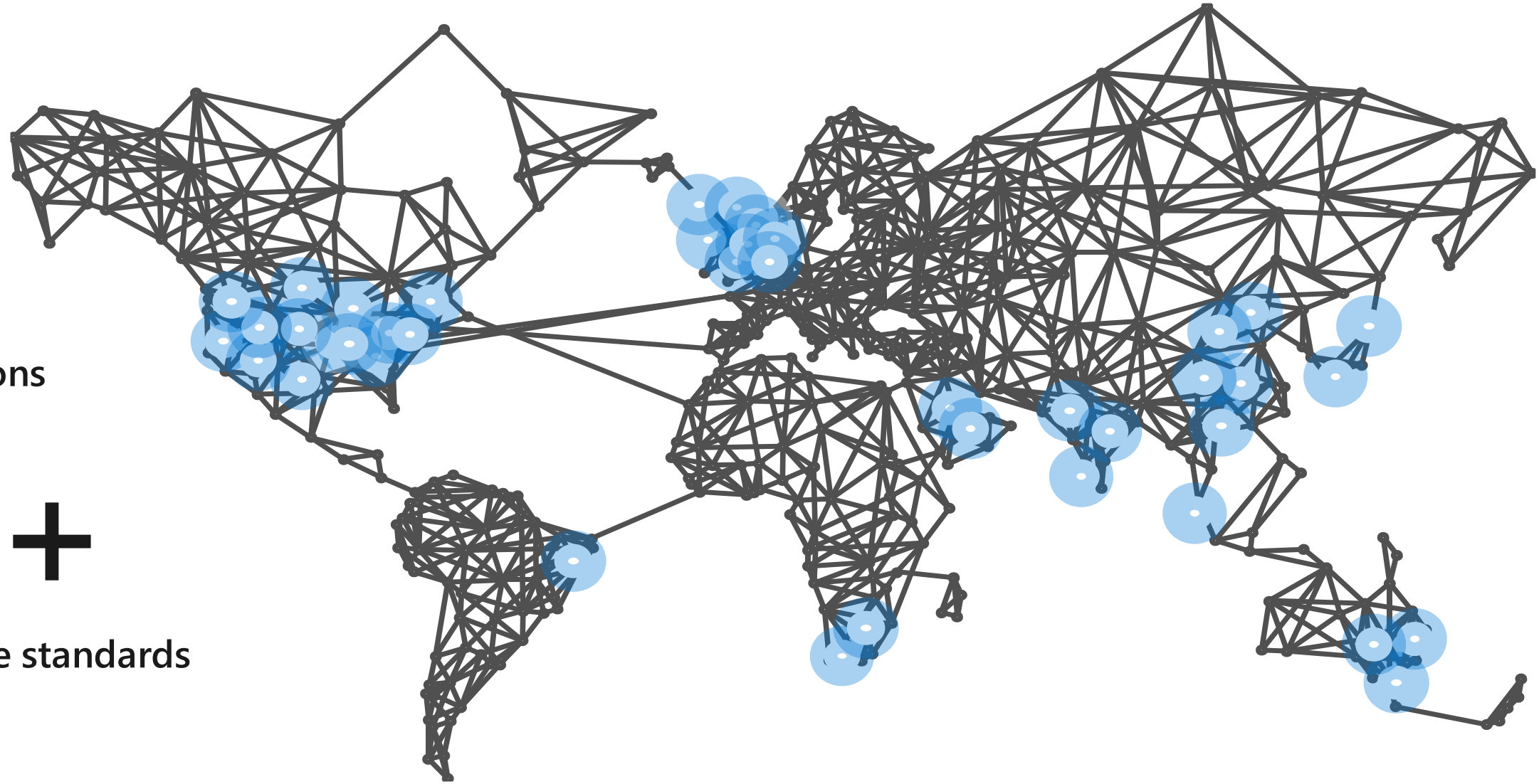
Build, manage and deploy applications on a massive global network

54

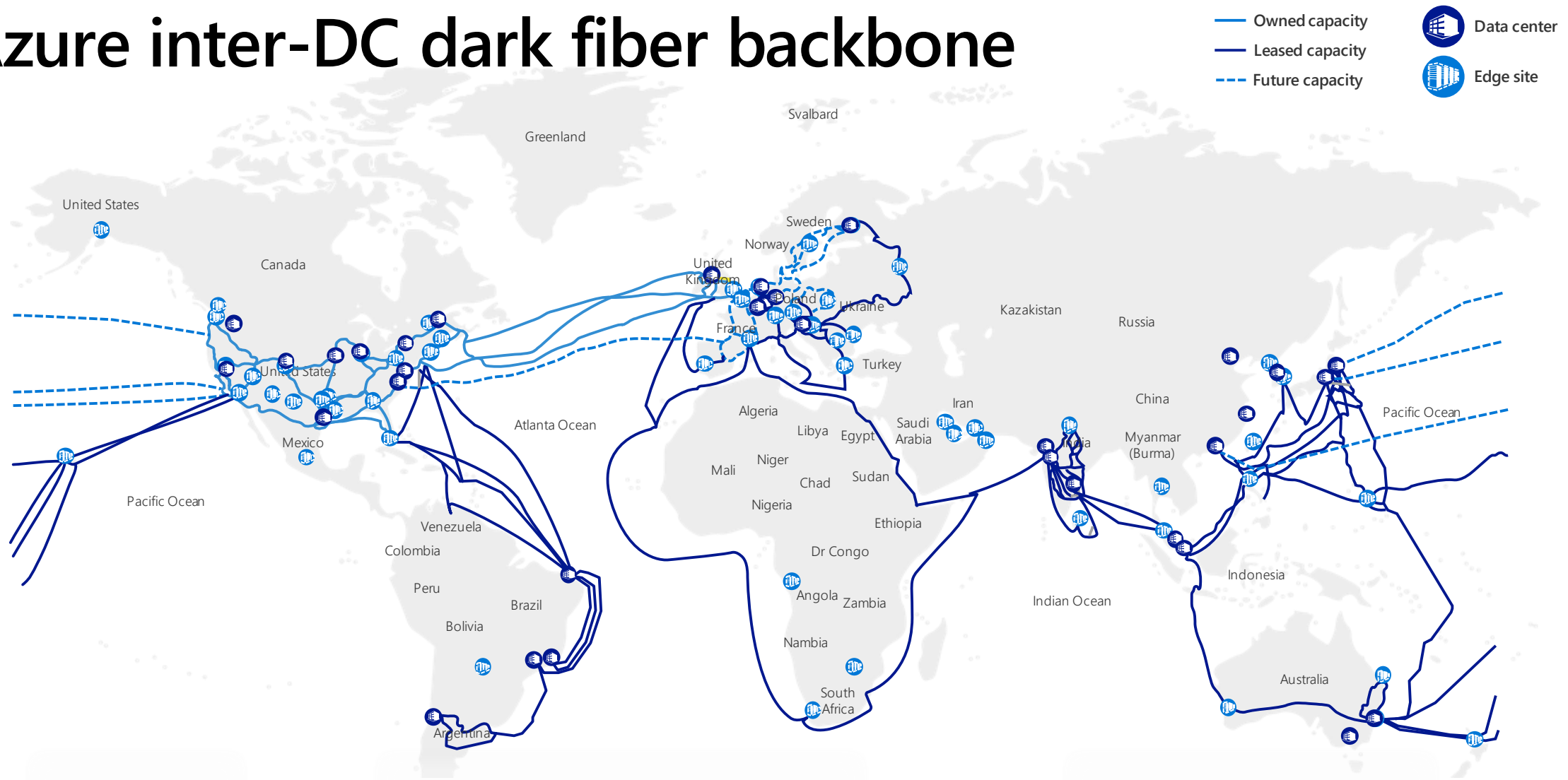
Azure regions

90+

Compliance standards



Azure inter-DC dark fiber backbone



54 Azure regions

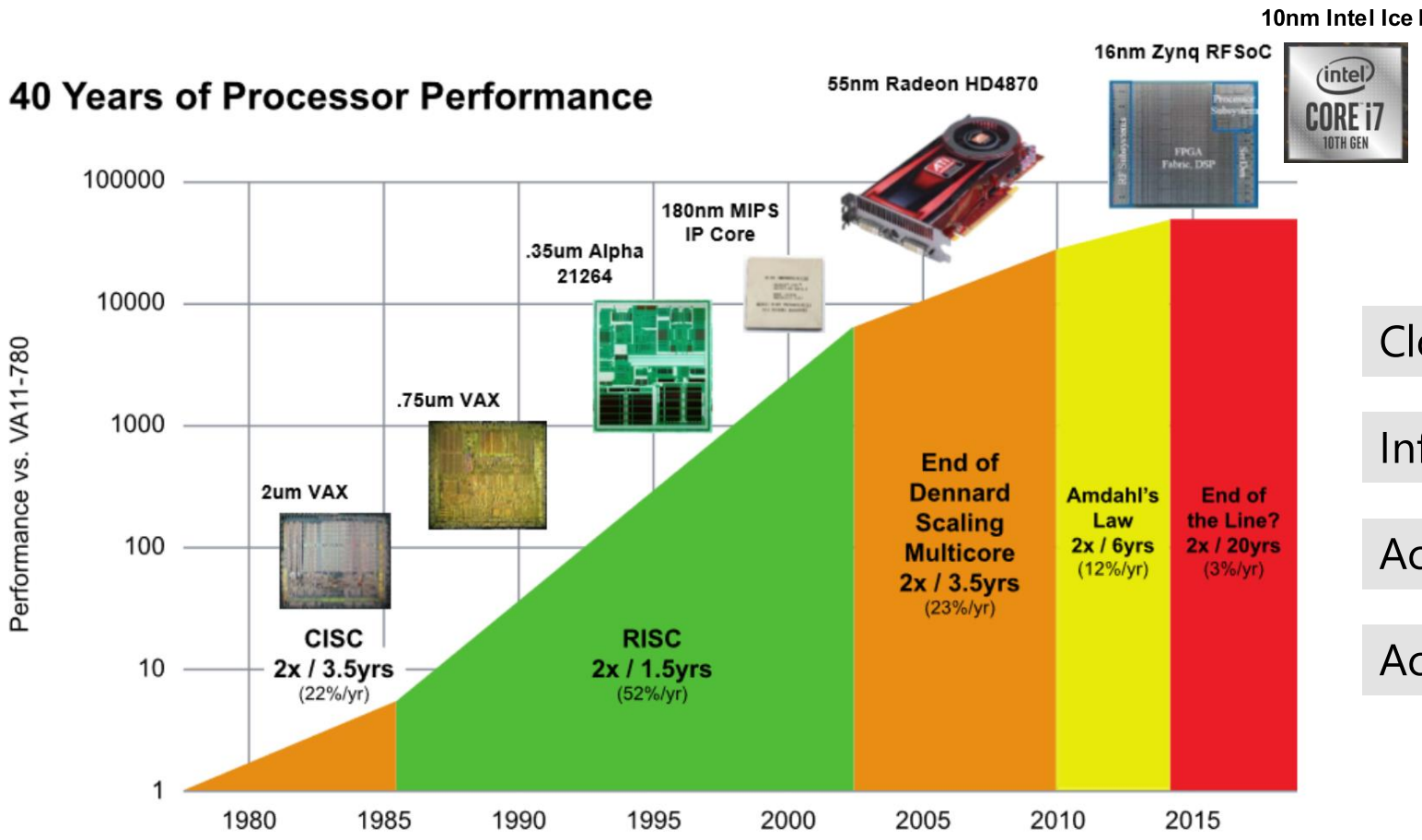
100k+ miles of fiber and subsea cable

150+ edge sites

Operating at this scale is hard,
but there's a problem …

Moore's Law & Compute Paradigm

"Moore's law states the observation that number of transistors in a dense integrated circuit doubles every 2 years."



Cloud software optimization

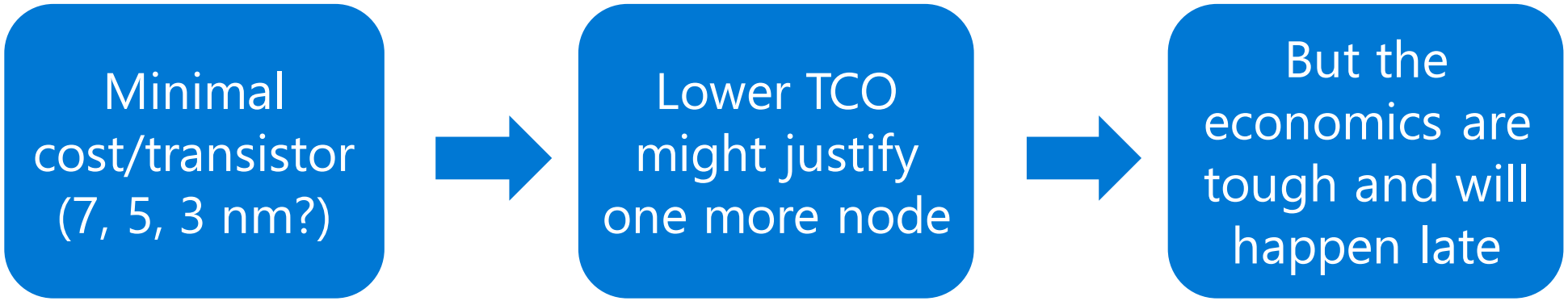
Infrastructure acceleration

Accelerators for new domains (AI)

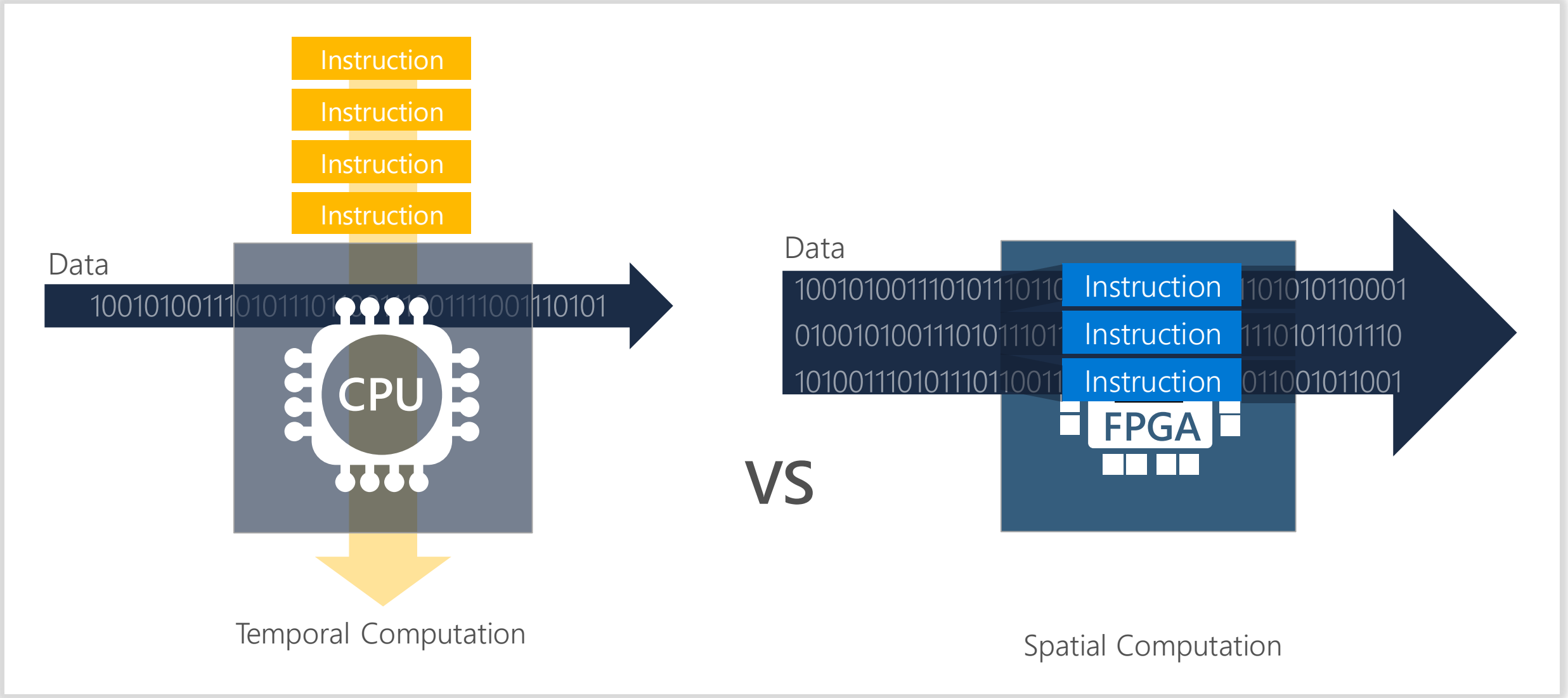
Accelerators for data processing?

Source: John Hennessy and David Patterson, Computer Architecture: A Quantitative Approach, 6/e 2018

Silicon scaling will end because of economics, not physics

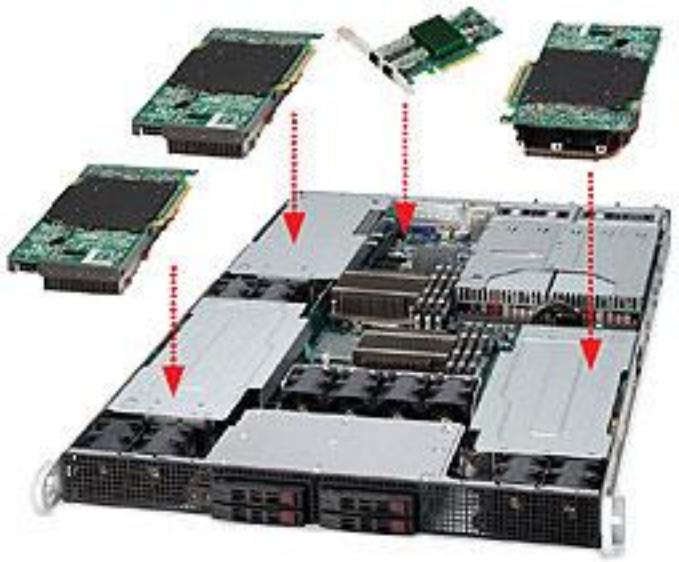


Temporal vs. spatial computing

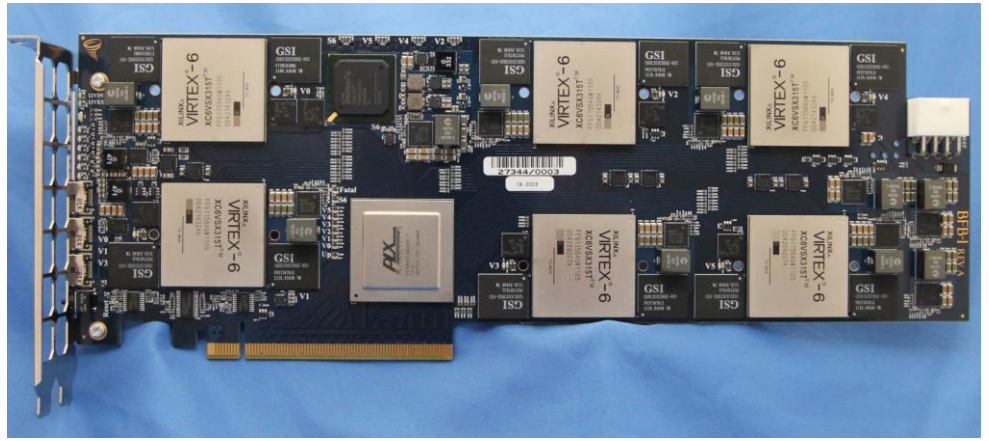


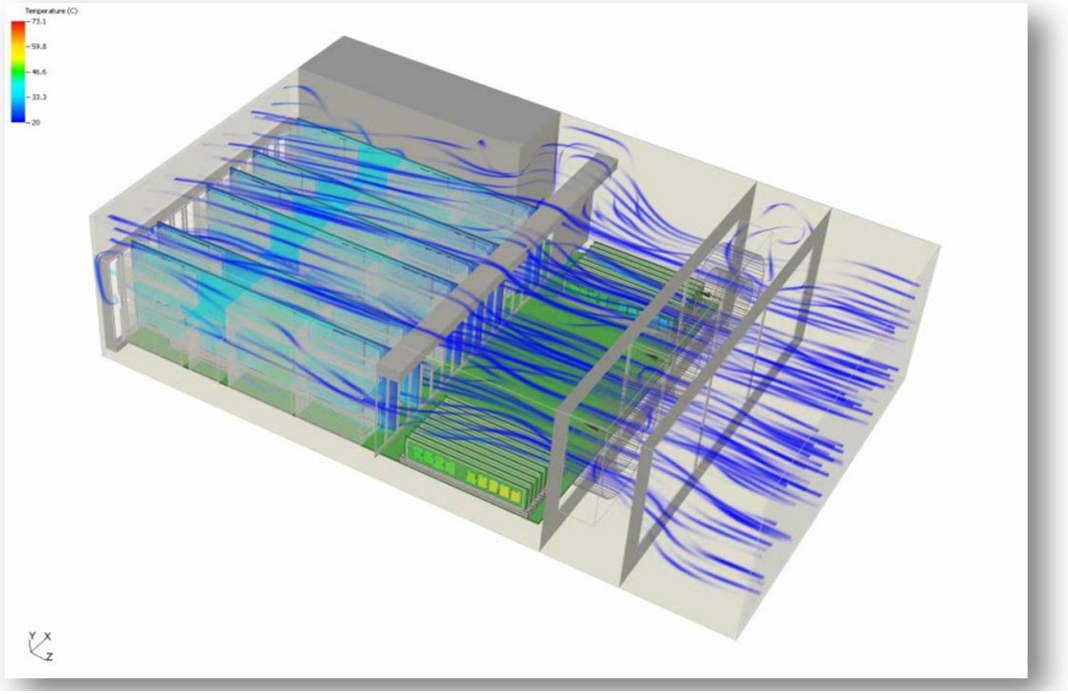
Catapult V0

- Use commodity SuperMicro servers
- 6 Xilinx LX240T FPGAs
- One appliance per rack
- All rack machines communicate over 1Gb Ethernet

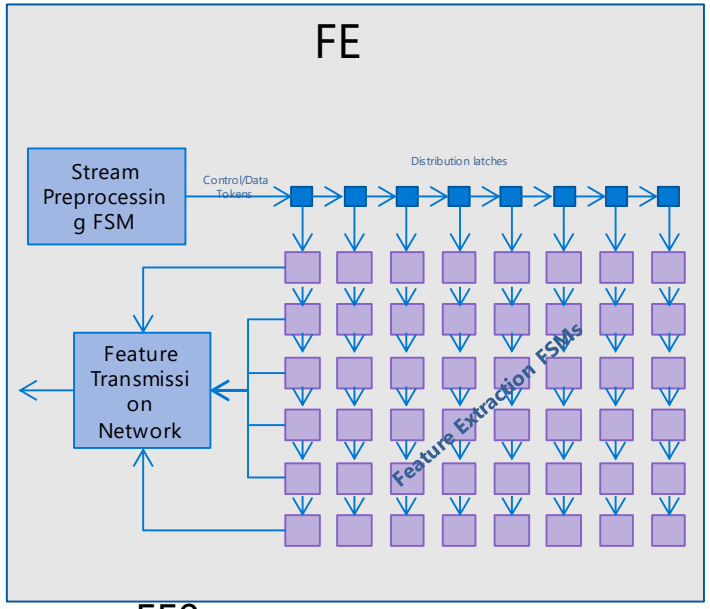


- 1U rack-mounted
- 2 x 10Ge ports
- 3 x16 PCIe slots
- 12 Intel Westmere cores (2 sockets)





Bing Ranking Implementation Details

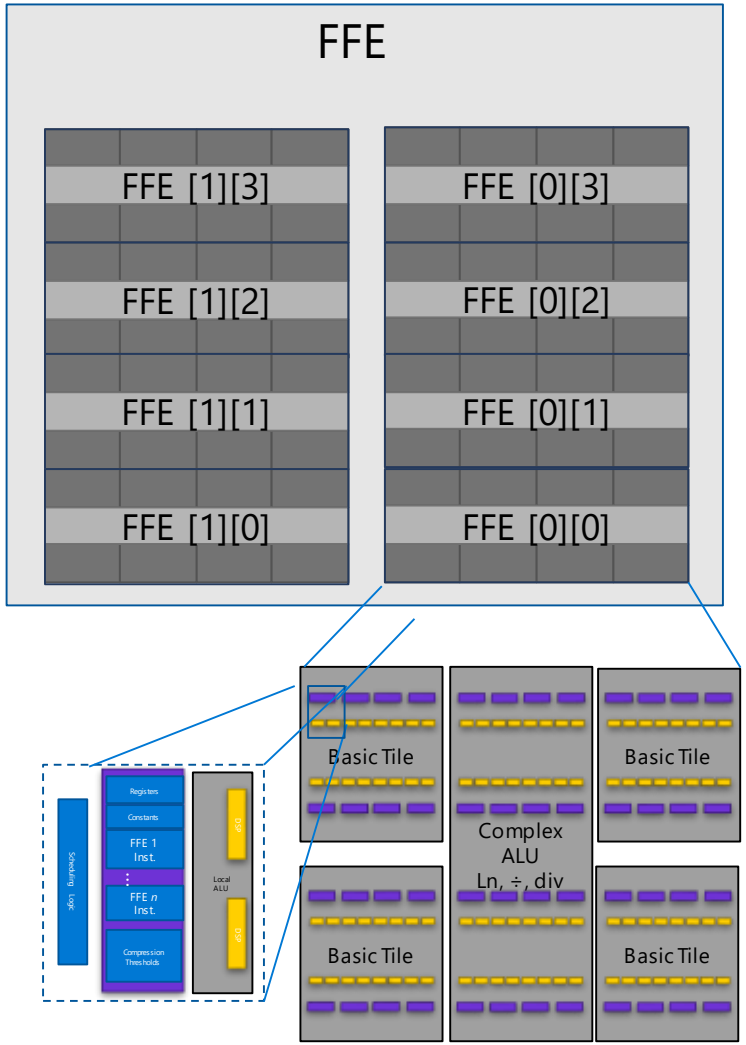


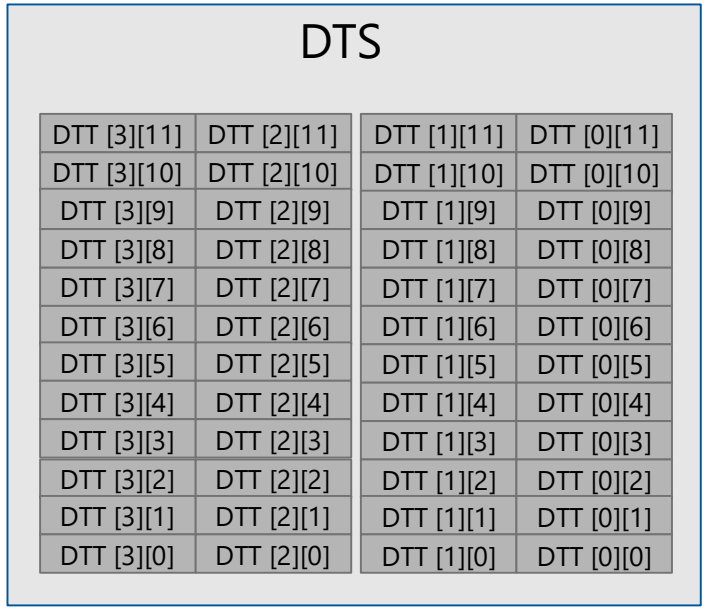
FE0

89 Non-BodyBlock Features
34 State Machines
55 % Utilization

FE1

55 BodyBlock Features
20 State Machines
45 % Utilization



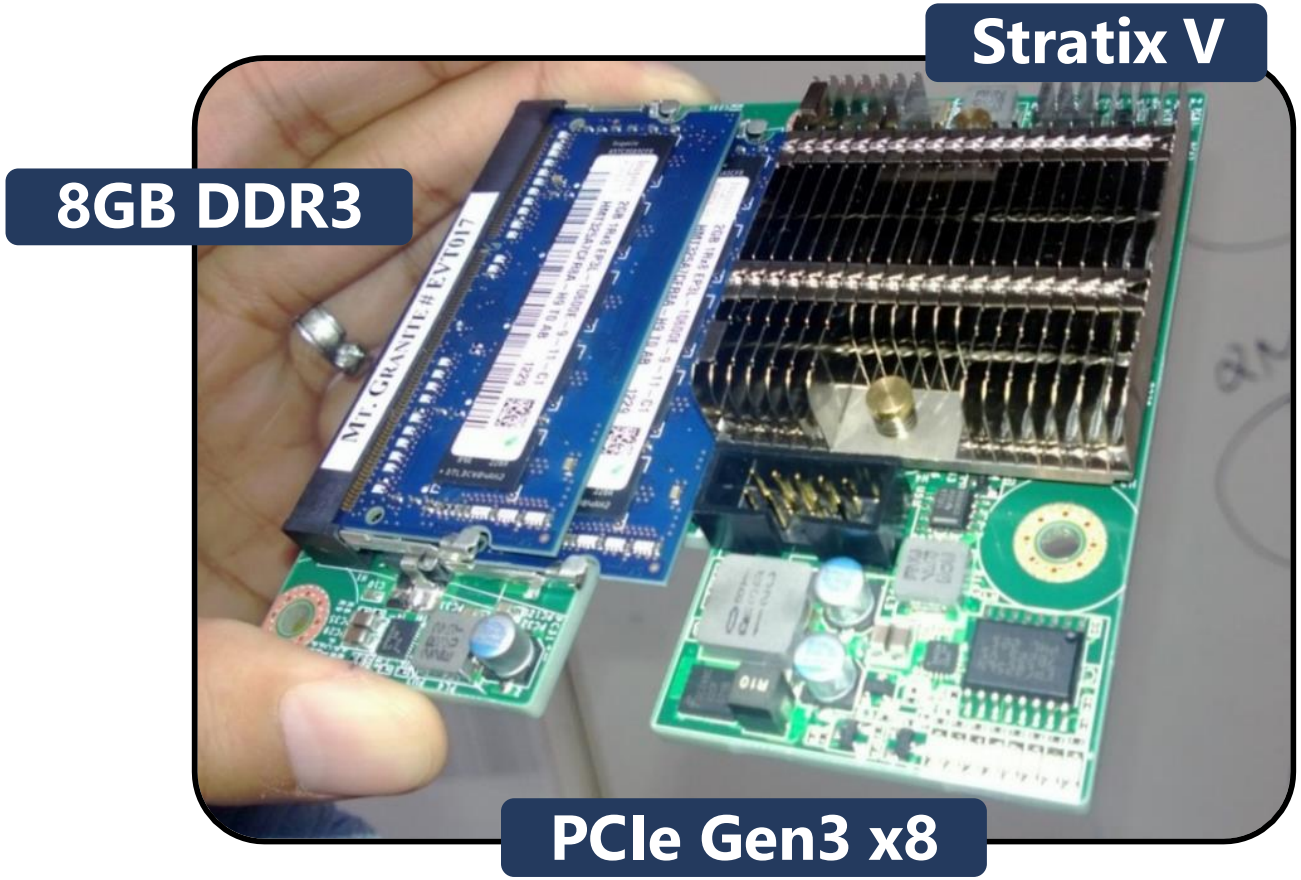


FFE: 64 cores / chip
256-512 threads
DTT: 48 DTT tiles/chip
240 tree processors
2880 trees/chip

- Fundamental flaws:
  - Additional single point of failure
  - Additional SKU to maintain
  - Too much load on the network
  - Inelastic FPGA scaling or stranded capacity

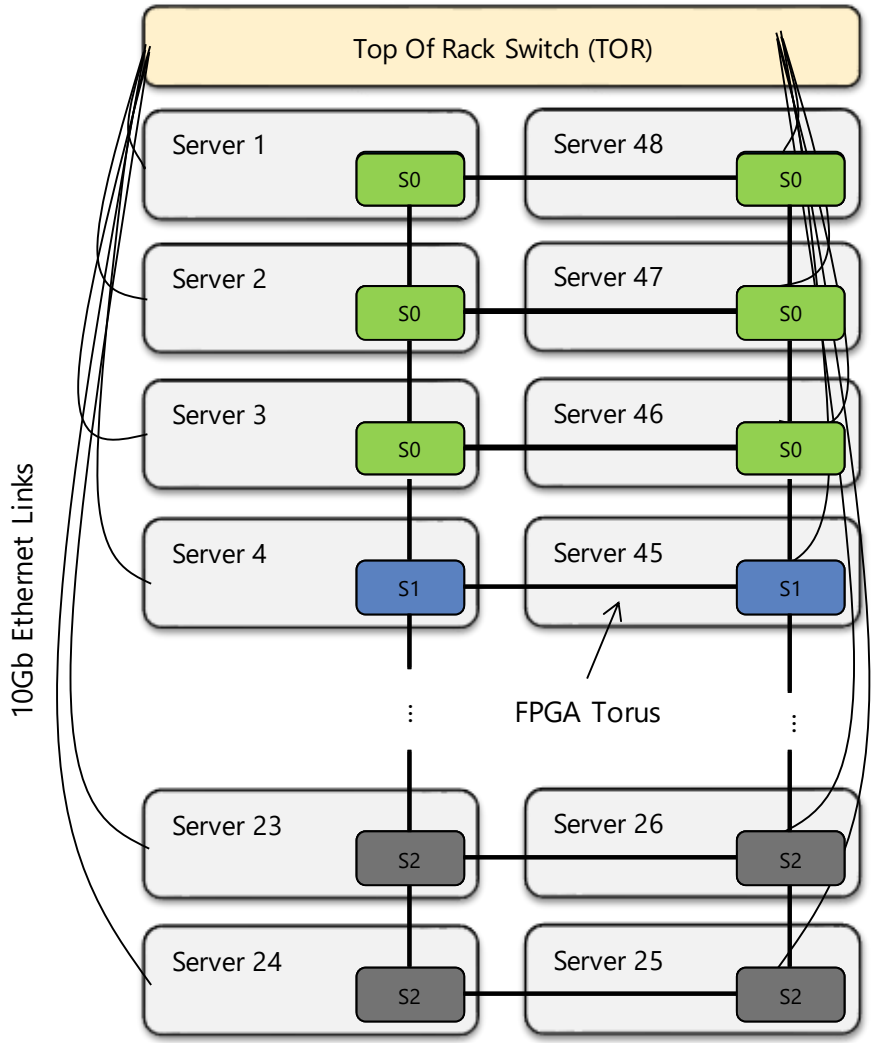
Catapult V1 Accelerator Card

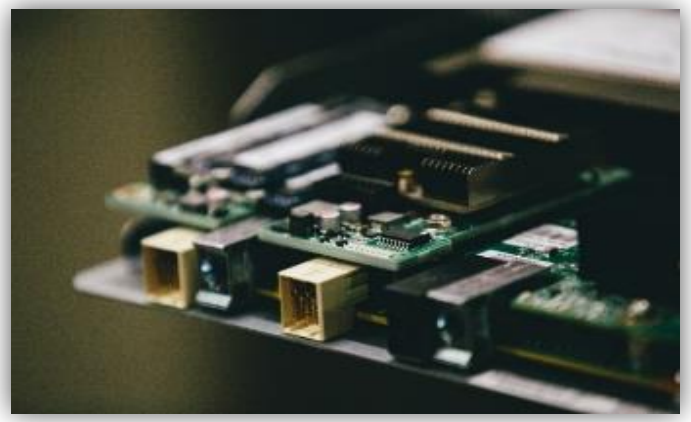
- Altera Stratix V D5
- 172.6K ALMs, 2014 M20Ks
  - 457KLEs
  - 1 KLE == ~12K gates
  - M20K is a 2.5KB SRAM
- PCIe Gen 2 x8, 8GB DDR3
- 20 Gb network among FPGAs

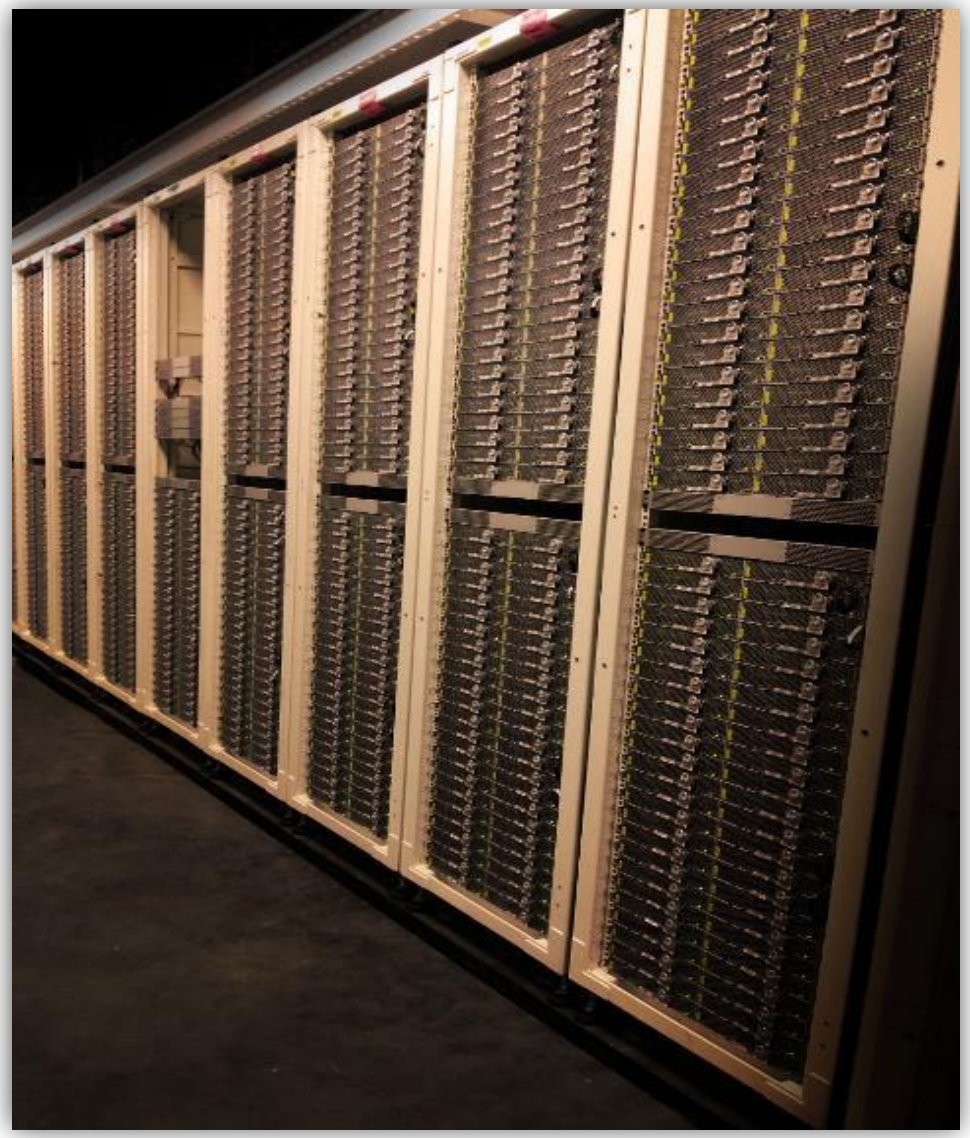


Mapped Fabric into a Pod

- **1 Pod = 48 servers**
  - Occupies 1 half-rack
  - 48-port 10G TOR switch
  - 1 server = 2 sockets, 64GB RAM, 2TB SSD storage
  - Deployed 34 pods, 1632 servers
- **FPGA Network**
  - 20Gb (2x10Gb) links to N/S/E/W neighbors
  - 2-D Torus Topology (6x8 torus)
- **Offered Capabilities**
  - Low-latency access to a local FPGA
  - Compose multiple FPGAs to accelerate large workloads
  - Low-latency, high-bandwidth sharing of storage and memory across server boundaries



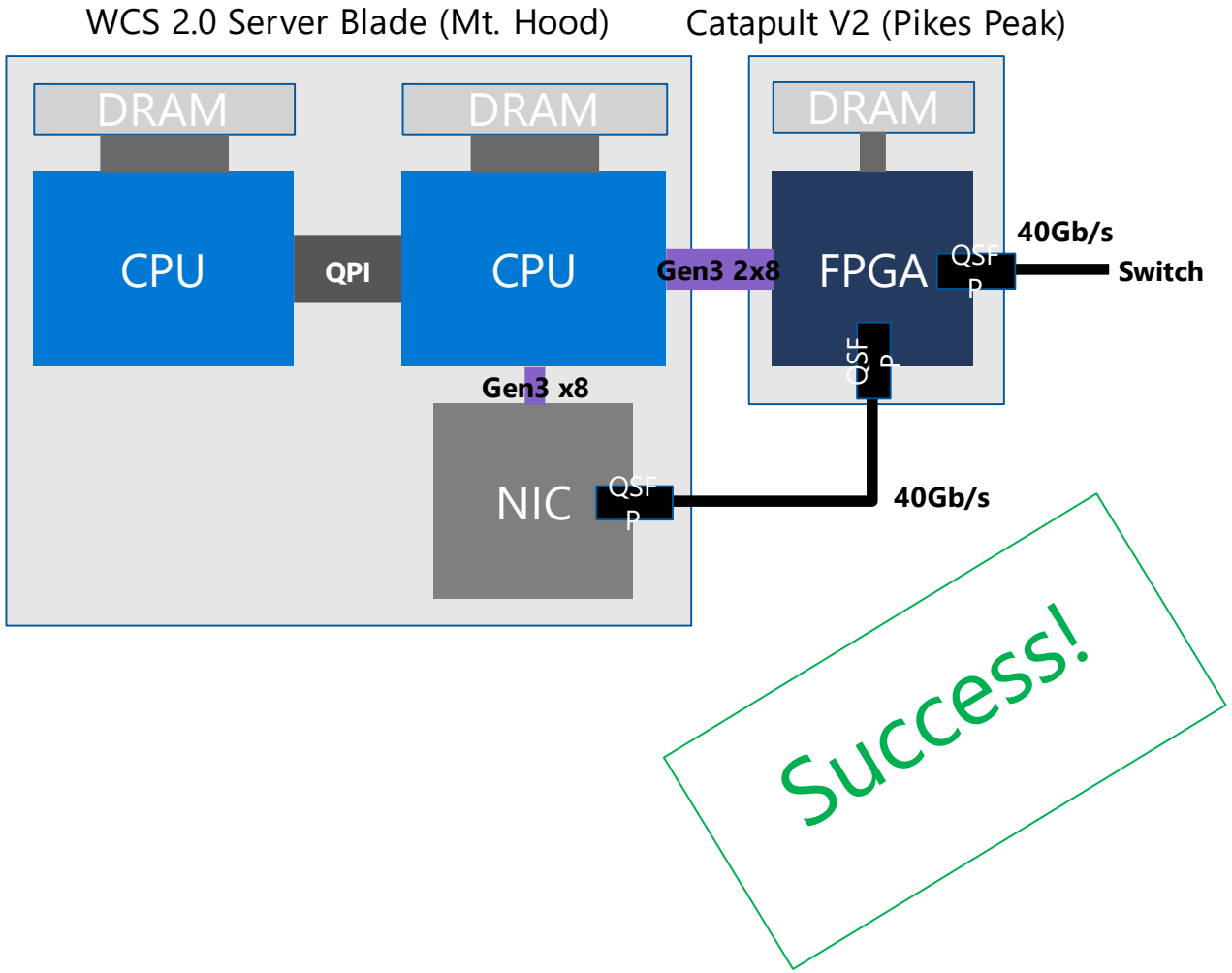




1,632 server pilot deployed in production datacenter

- Fundamental flaws:
  - Microsoft was converging on a single SKU
  - No one else wanted the secondary network
    - Complex, difficult to handle failures
    - Difficult to service boxes
  - No killer infrastructure accelerator
    - Application presence is too small

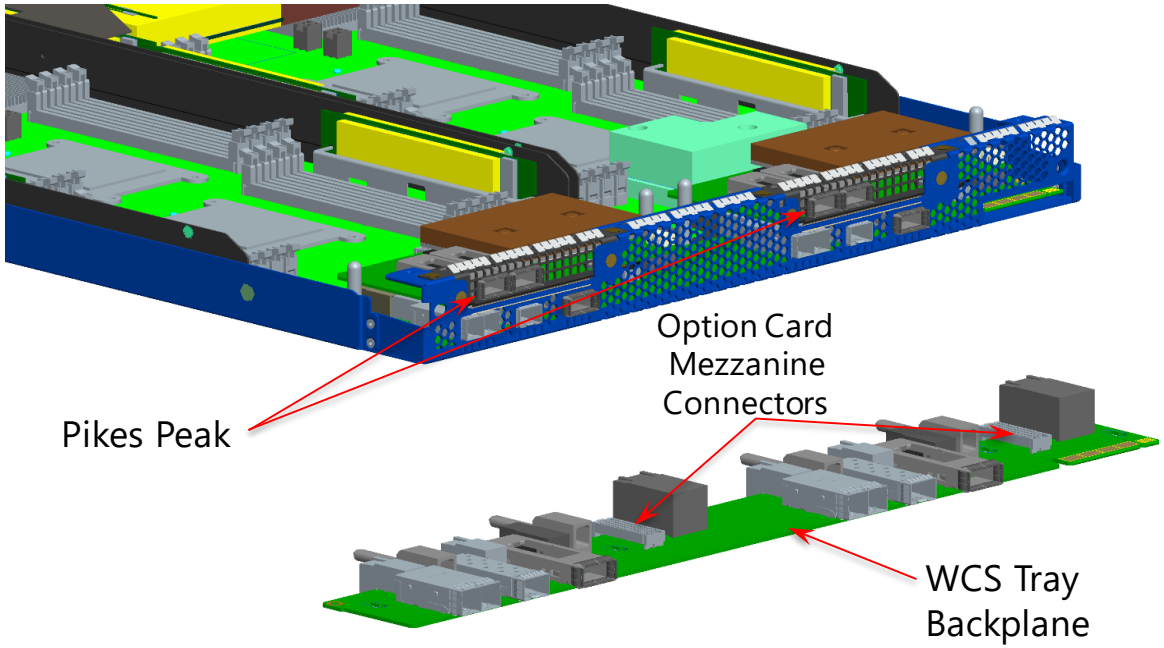
Catapult V2 Architecture



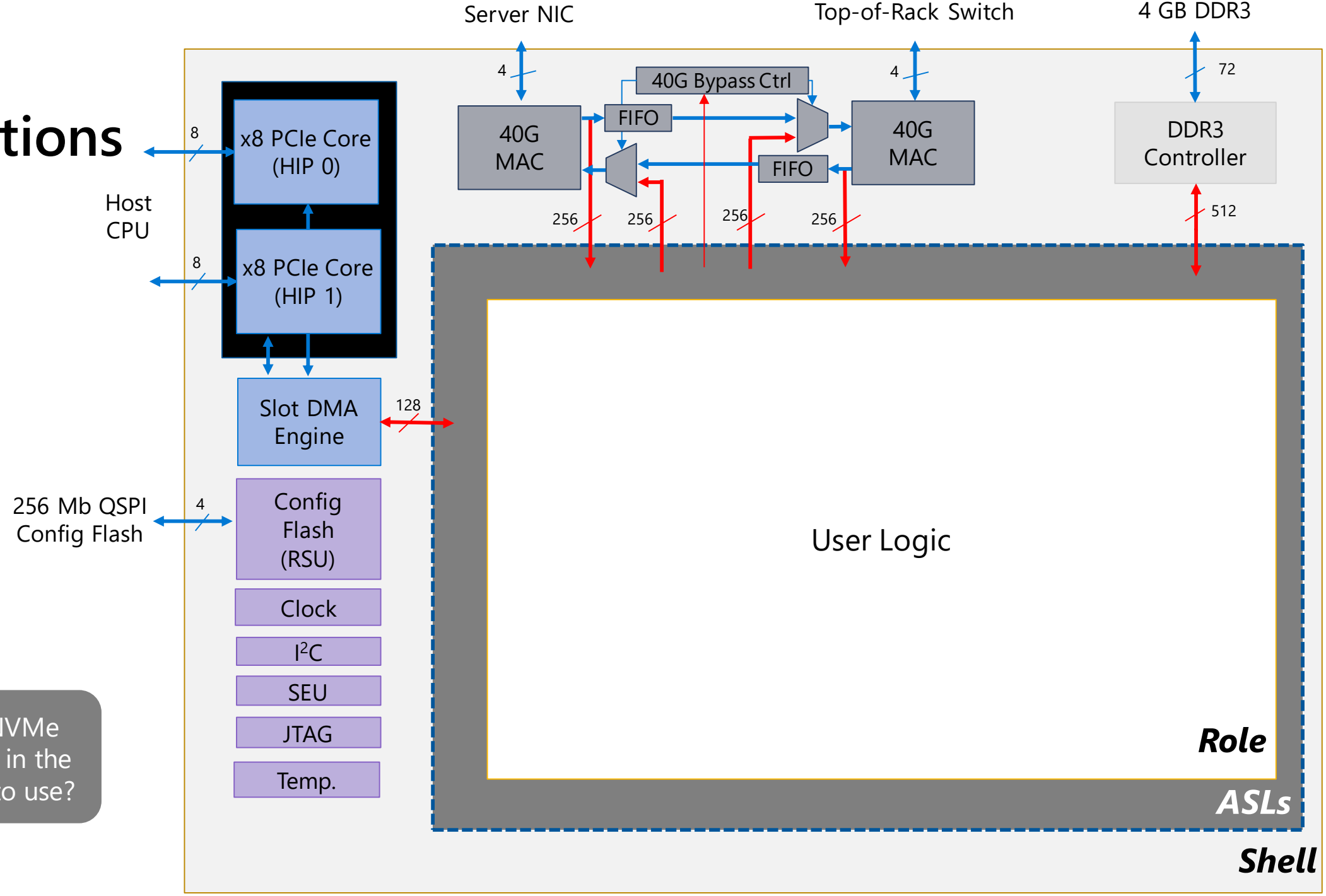
Success!

Catapult WCS Mezz card (Pike's Peak)

WCS Gen4.1 Blade with Mellanox NIC and Catapult FPGA

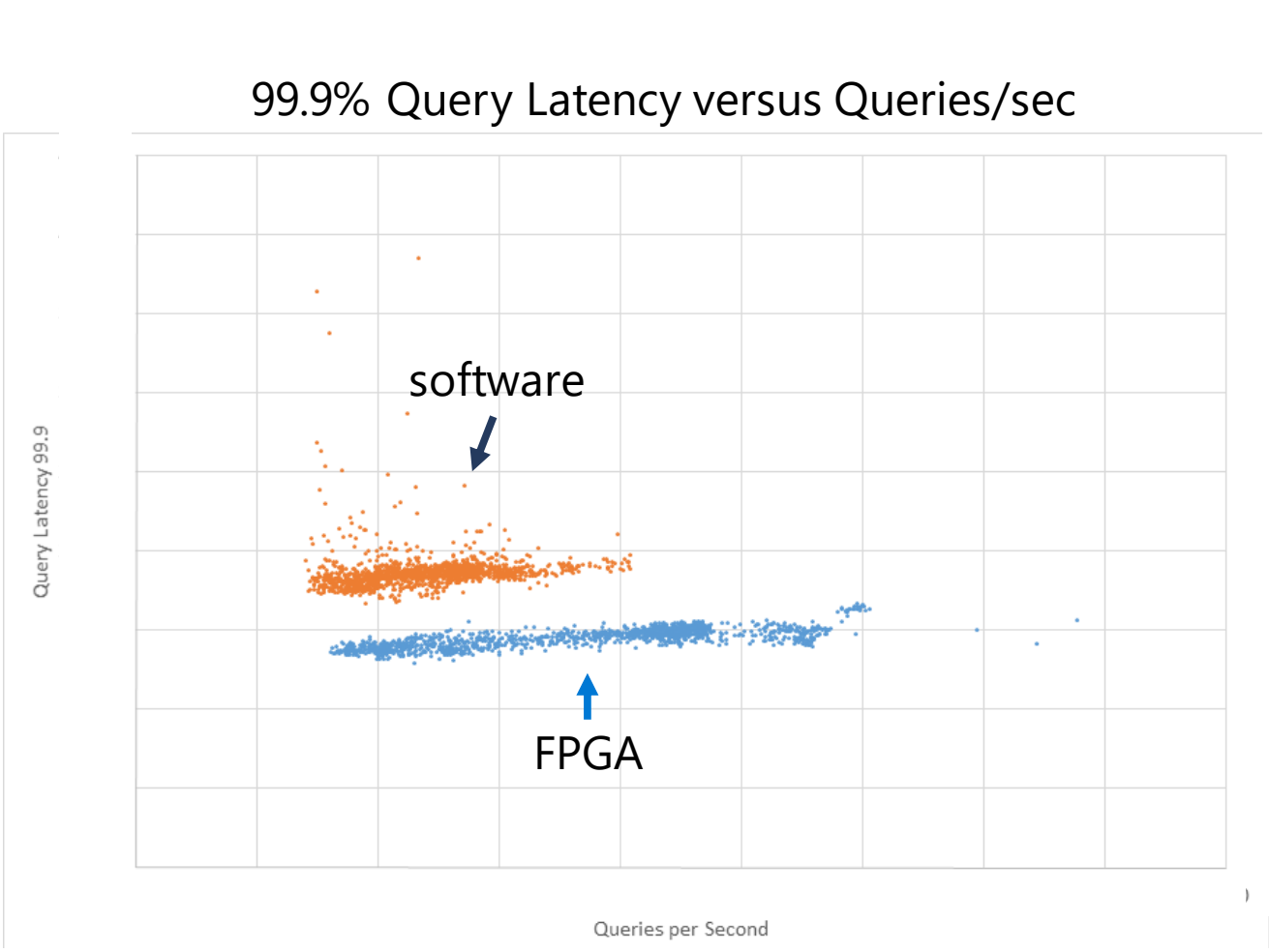


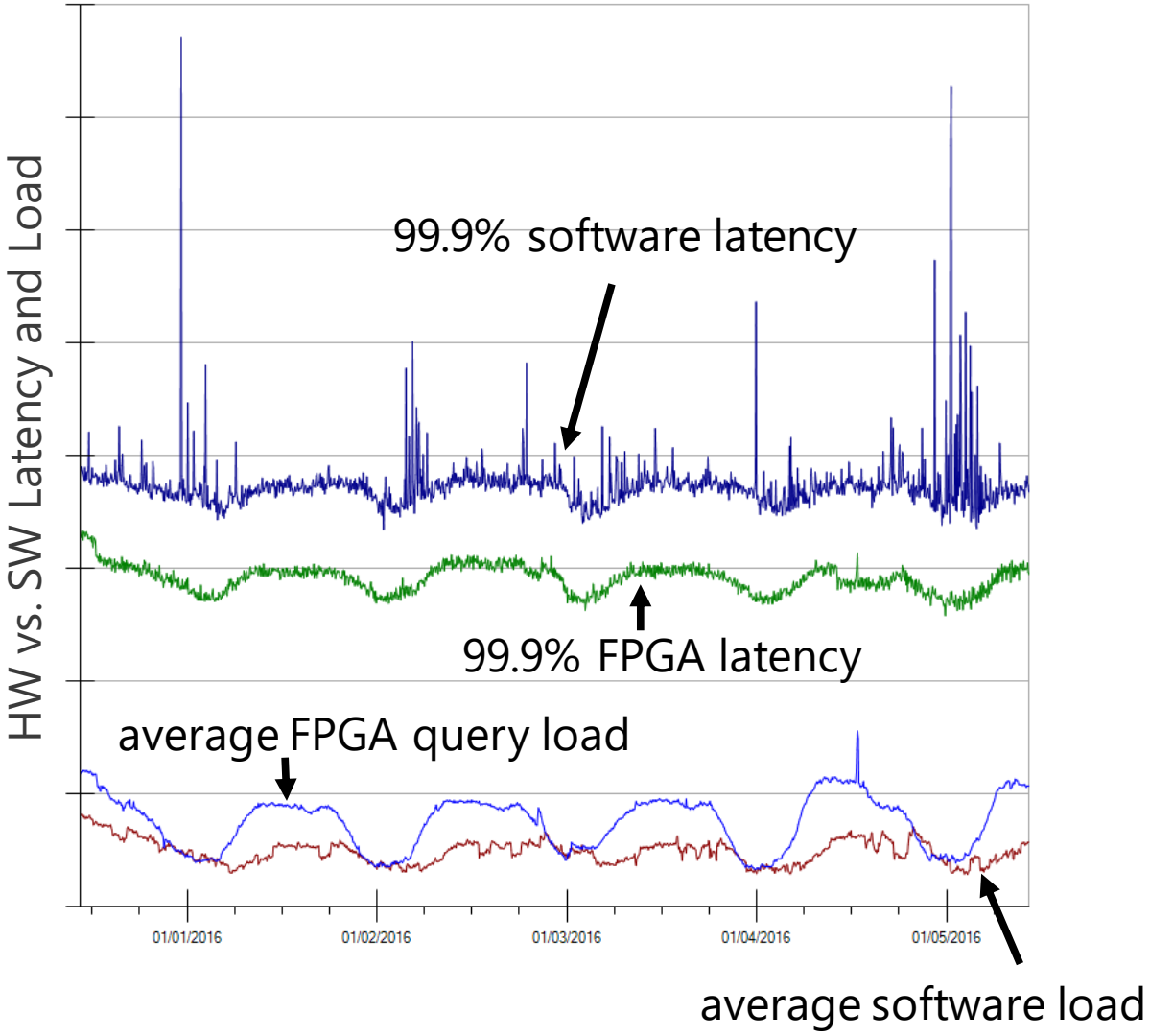
Shell Abstractions



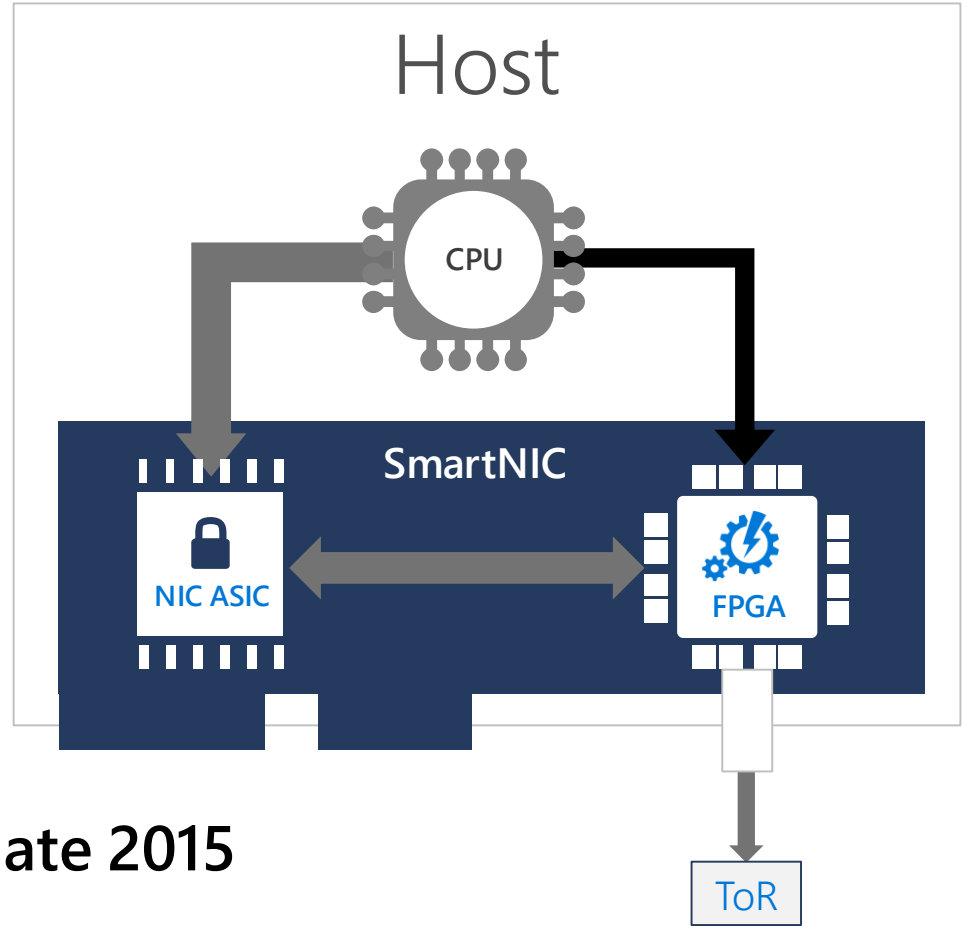
Can put NVMe controllers in the shell, how to use?

Bing Production Results





Gone to Scale: Azure SmartNIC



Use the FPGA for network processing

Roll out hardware as we do software, monthly feature updates

Programmed using Generic Flow Tables (GFT)

Language for programming SDN to hardware

Uses connections and structured actions as primitives

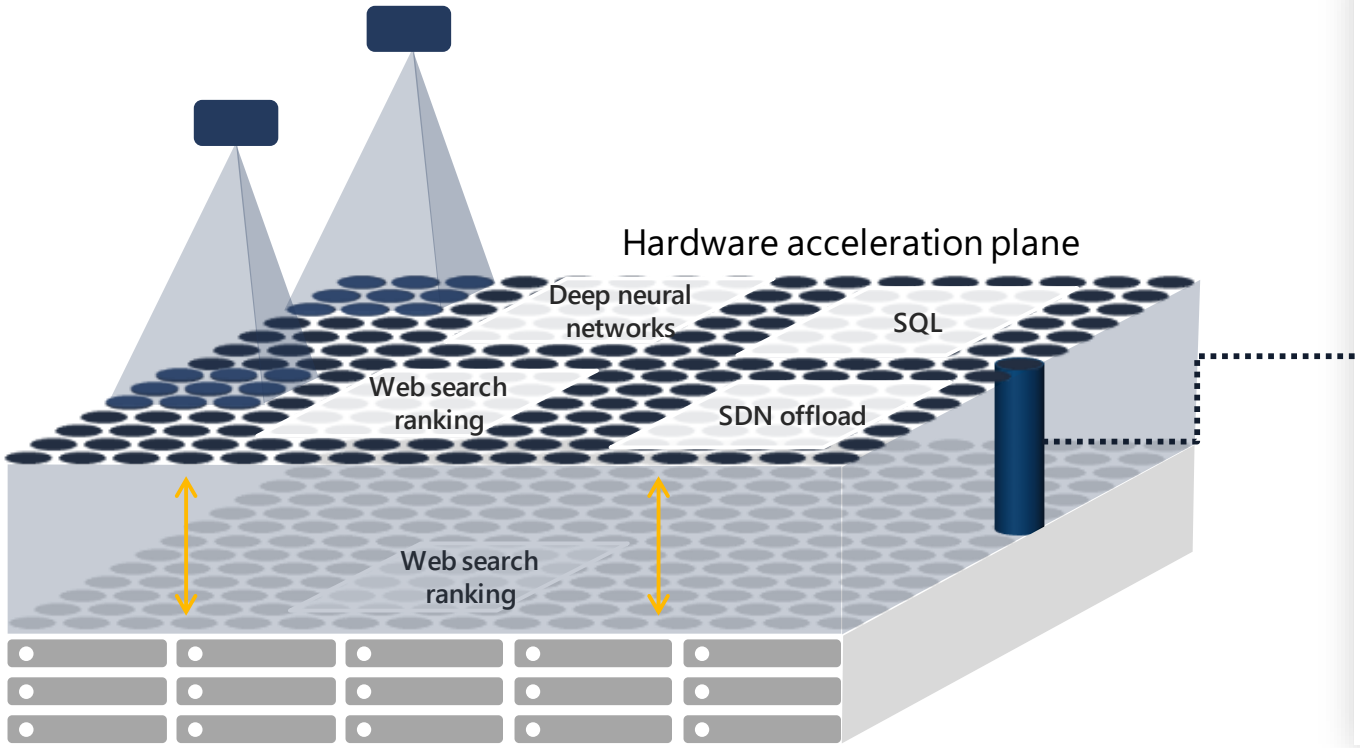
Deployed on all new Azure compute servers since late 2015

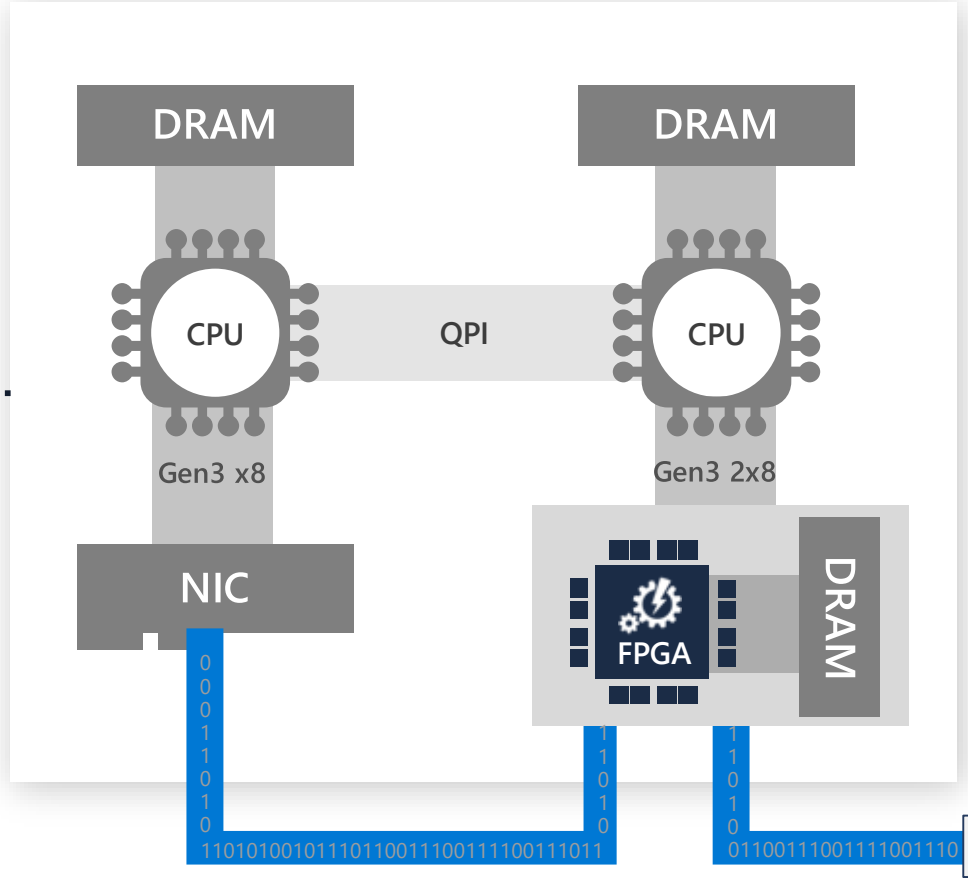
SmartNIC can also do Crypto, QoS, storage acceleration, and more…

Scalable Hardware Microservice fabric

Interconnected FPGAs form a separate plane of computation

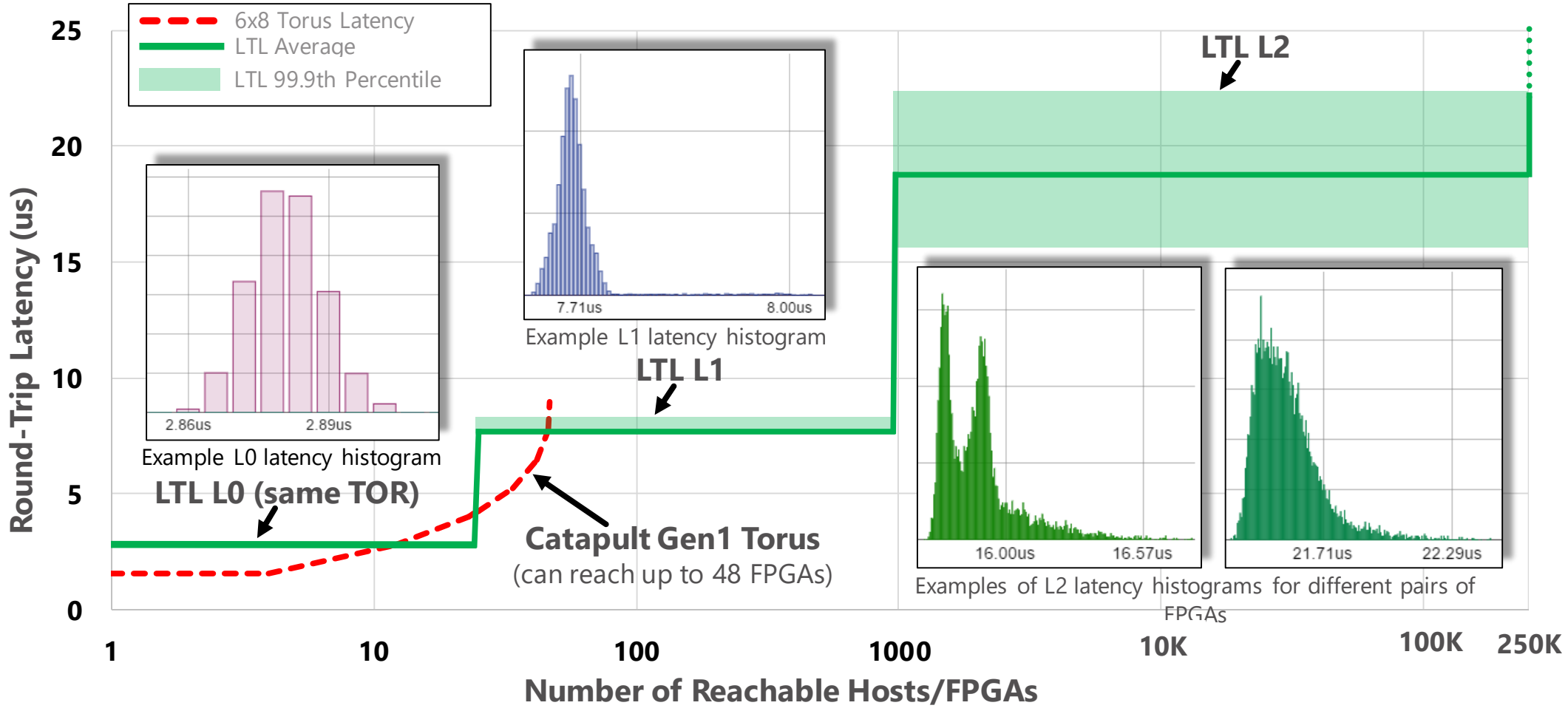
Can be managed and used independently from the CPU



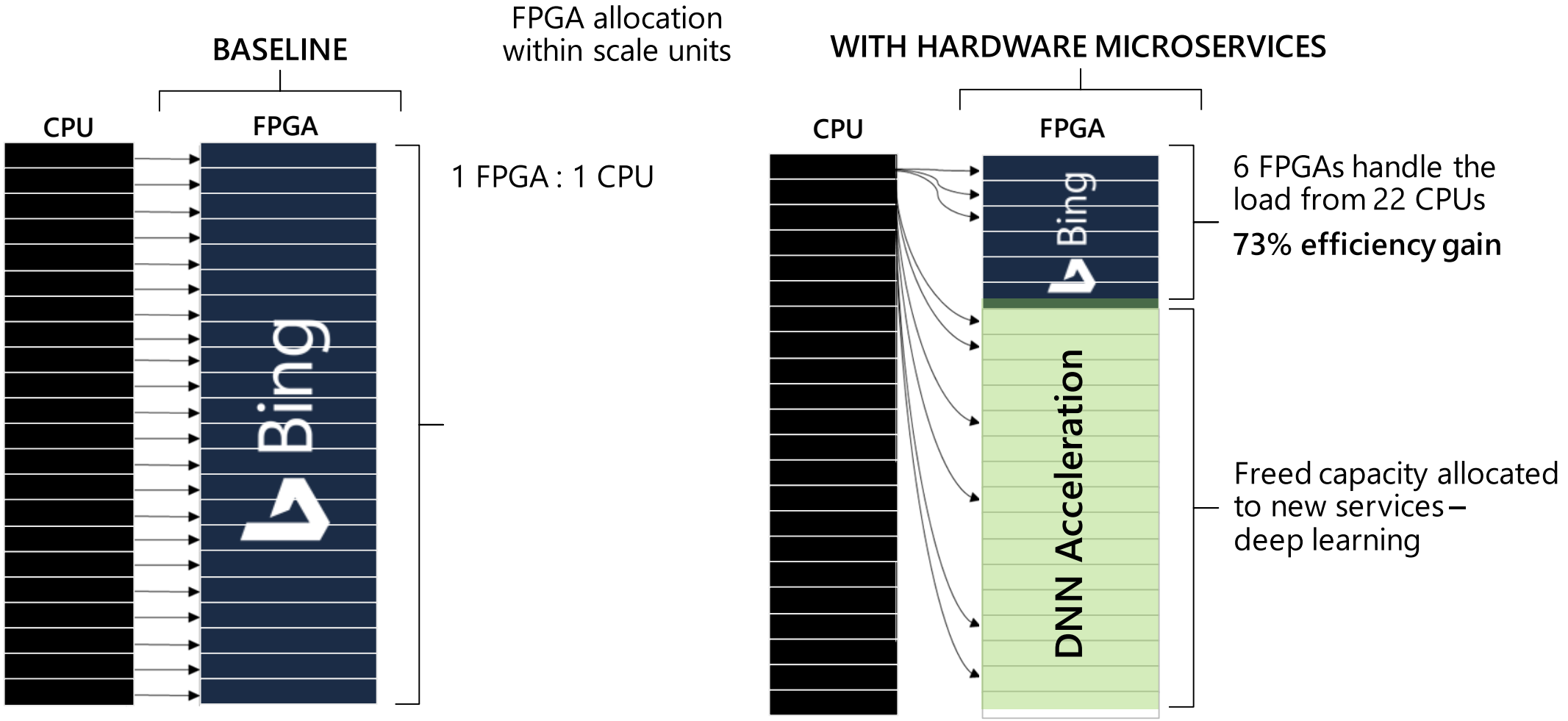


Accelerators as a First-Class Network Citizen

Network Reach and Latencies

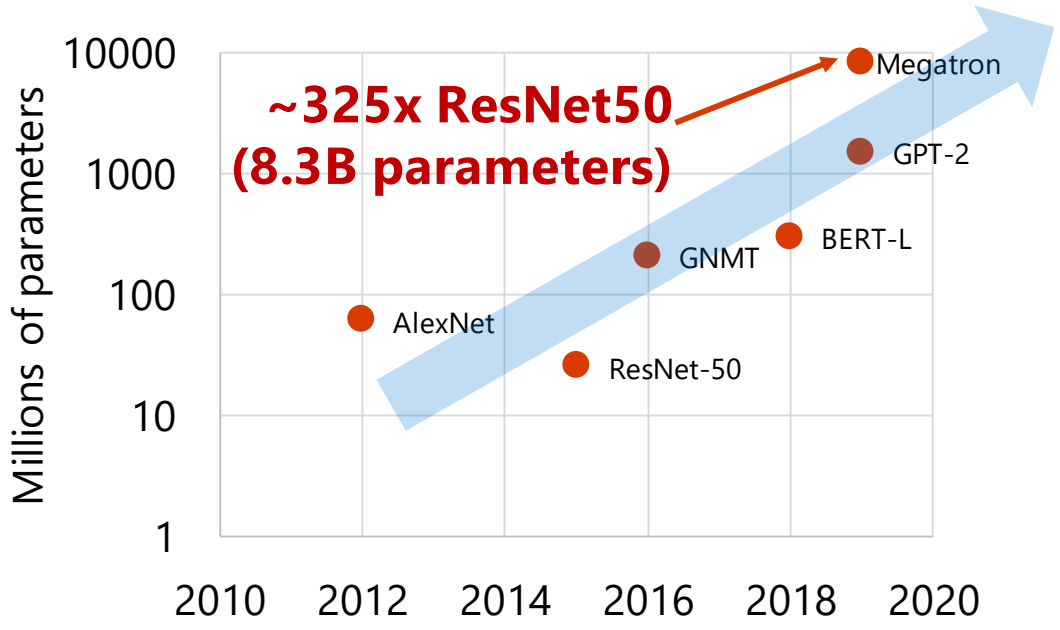


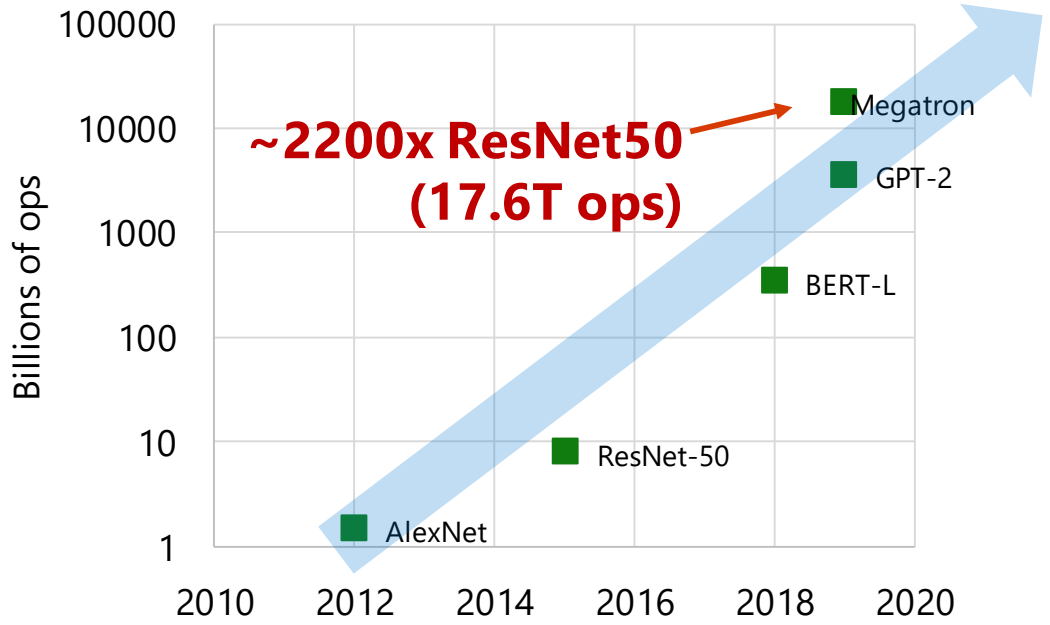
Example: Bing Hardware Microservices



HW increases FPGA utilization, enables new acceleration possibilities

AI compute: model sizes growing exponentially, 10X per year





Project Brainwave Engine

Fully scaled MSFP8 on Stratix 10

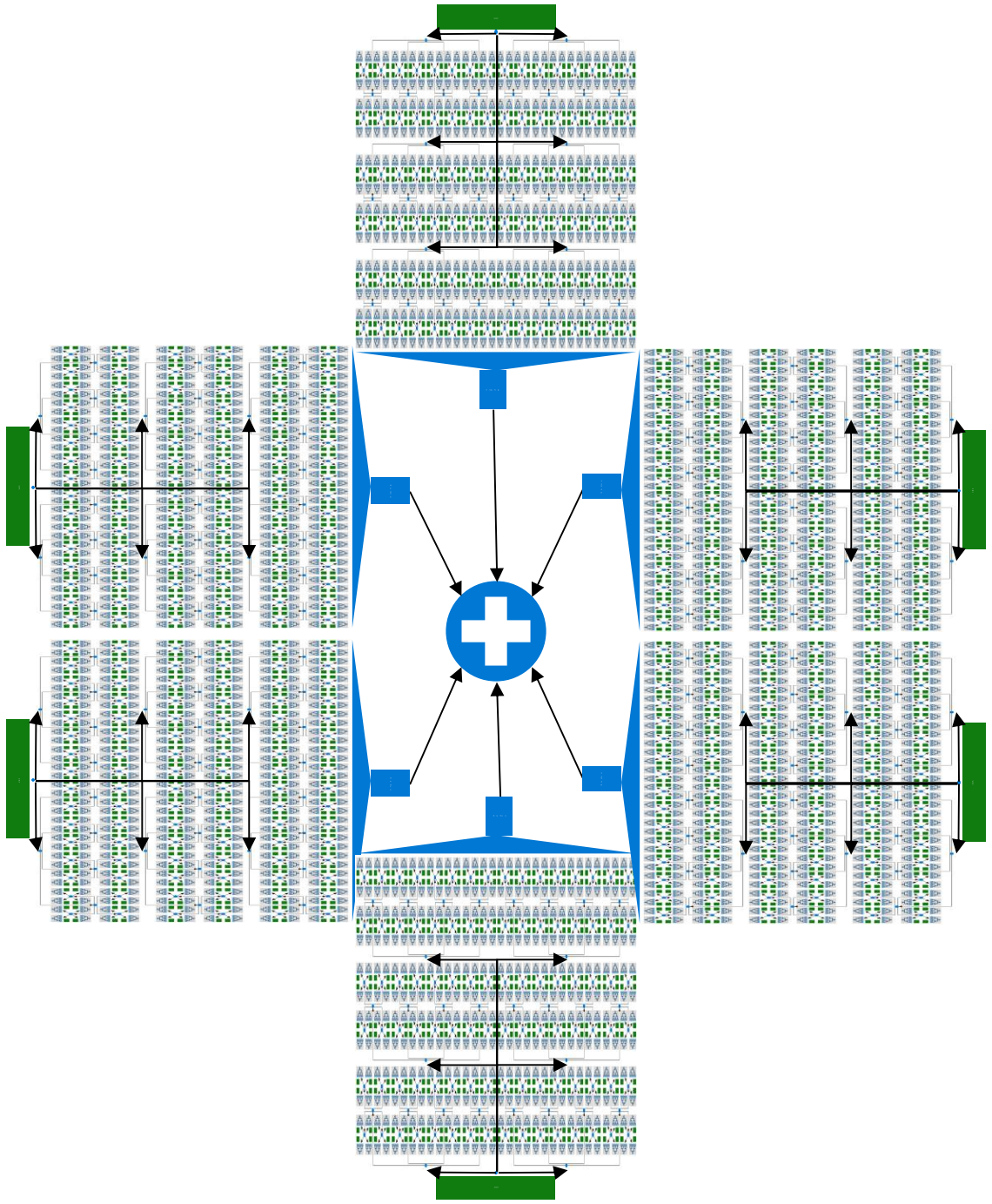
- 6 tensor cores
- 400 independent dots per core
- 40 vector lanes per dot product
- **Total =** 6 * 400 * 40 = **96,000 MACs**

96,000 compressed weights per cycle

48T peak @ 250MHz (engineering Si)

96T peak @ 500Mhz (production Si)

0.64 TOPS/watt @ 150W (Intel 14nm)



Using Project Brainwave for land use mapping

Using FPGAs for ultra-fast inferencing different types of land use for the entire United States (ESRI NAIP Data, 20+ TB), ~10 minutes, $42

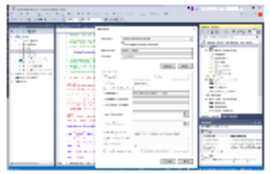
Data | Build | Train | Deploy | Intelligent Apps

esri Satellite Images for the Entire US

NAIP Data
20TB, 200M images

Stored on Azure Premium Storage

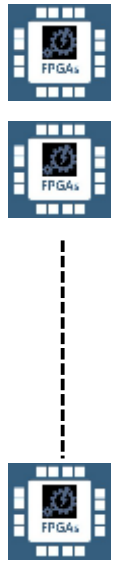
Geo AI Data Science Virtual Machine

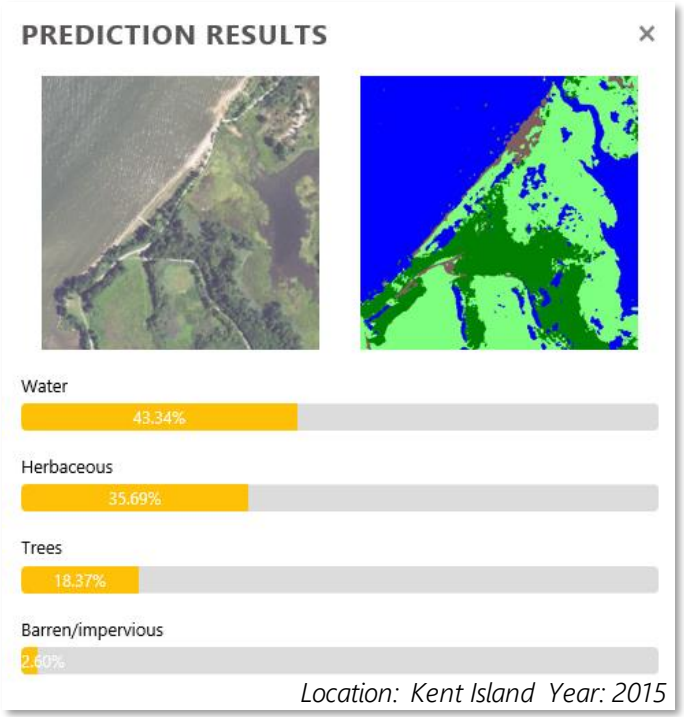


Visual Studio Tools for AI

Azure Machine Learning

Azure Batch AI



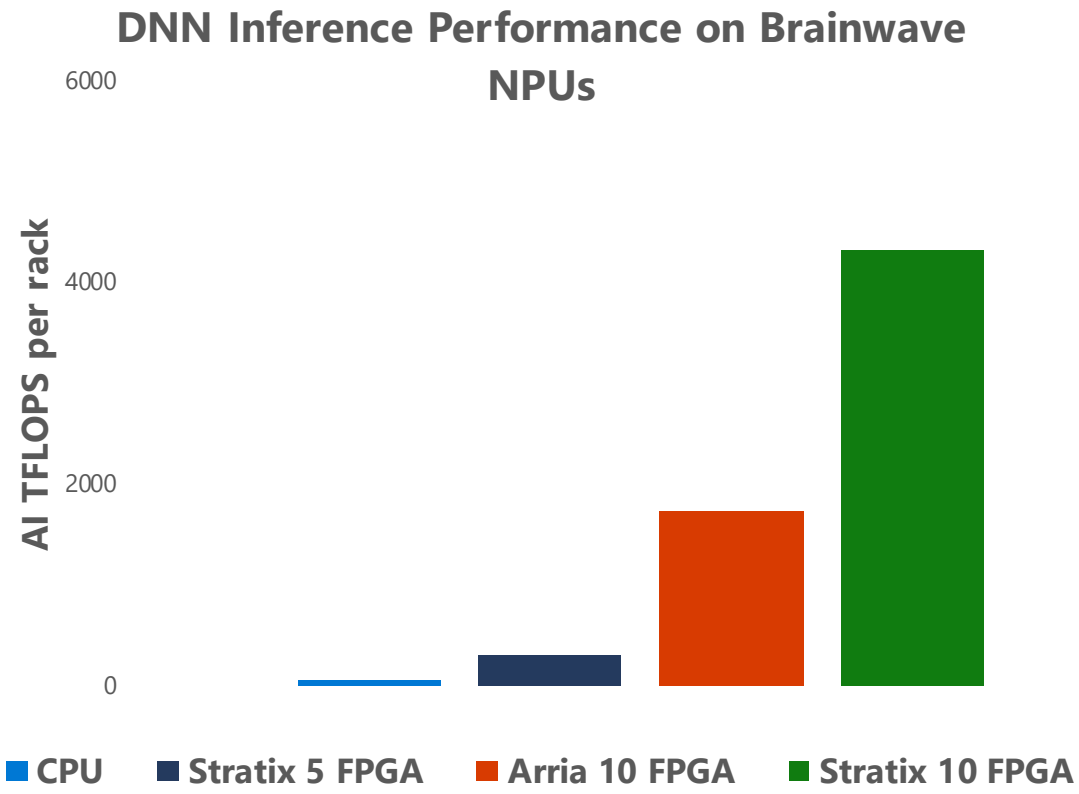


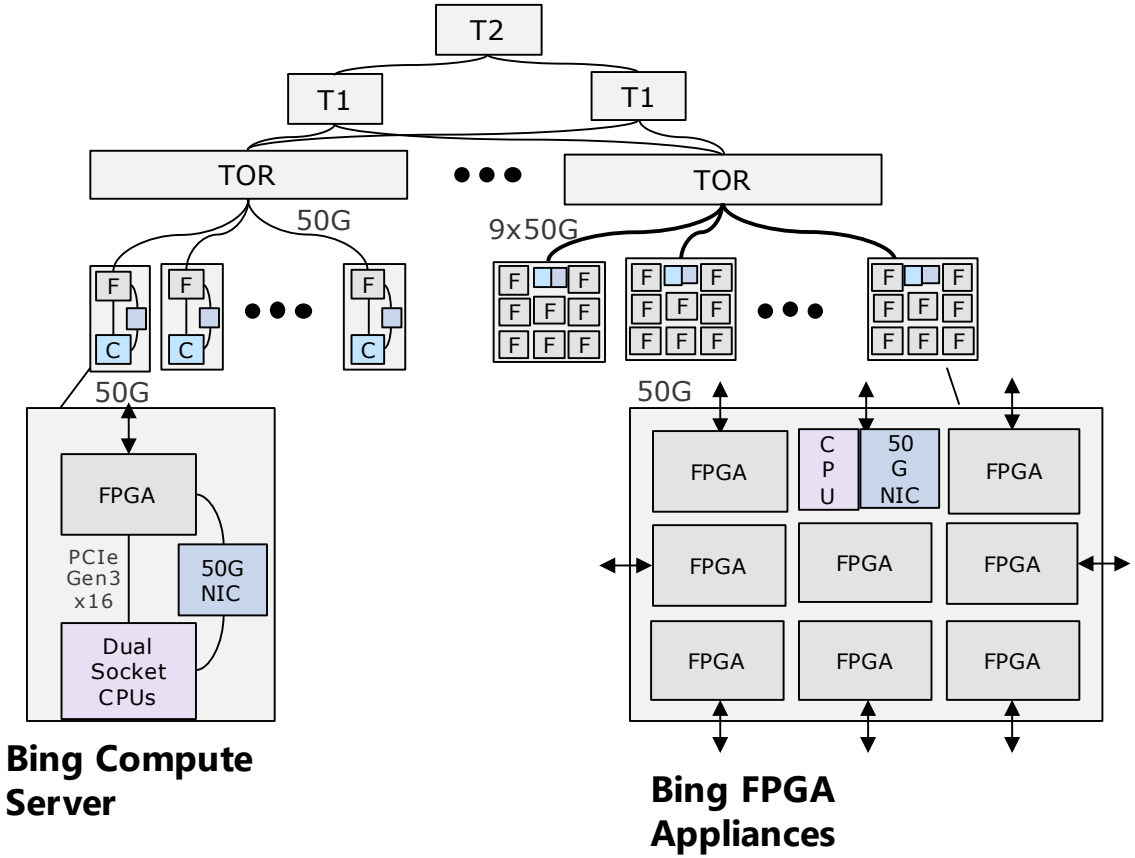
Land Classification Model ResNet-50

Ultra-fast Inferencing using FPGAs

Data ----------> Intelligence ----------> Action

Bing's 500 Petaflops FPGA Supercomputer (2018-now)





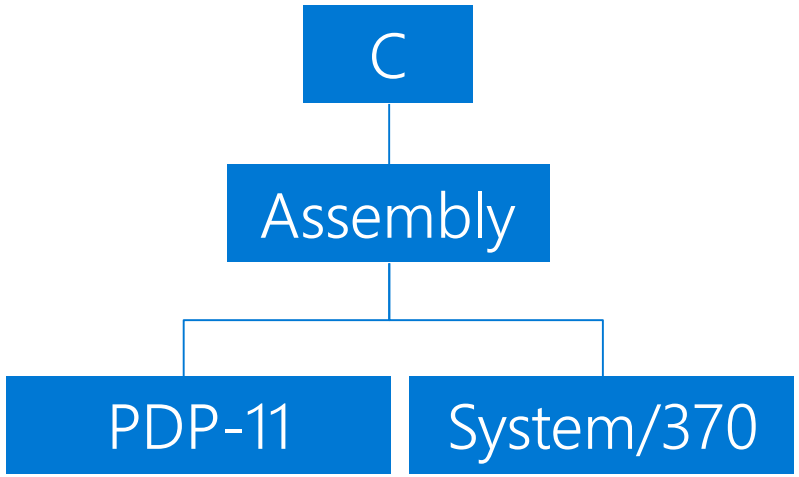
Brainwave NPUs are on steep growth trajectory for low-batch DNN inference

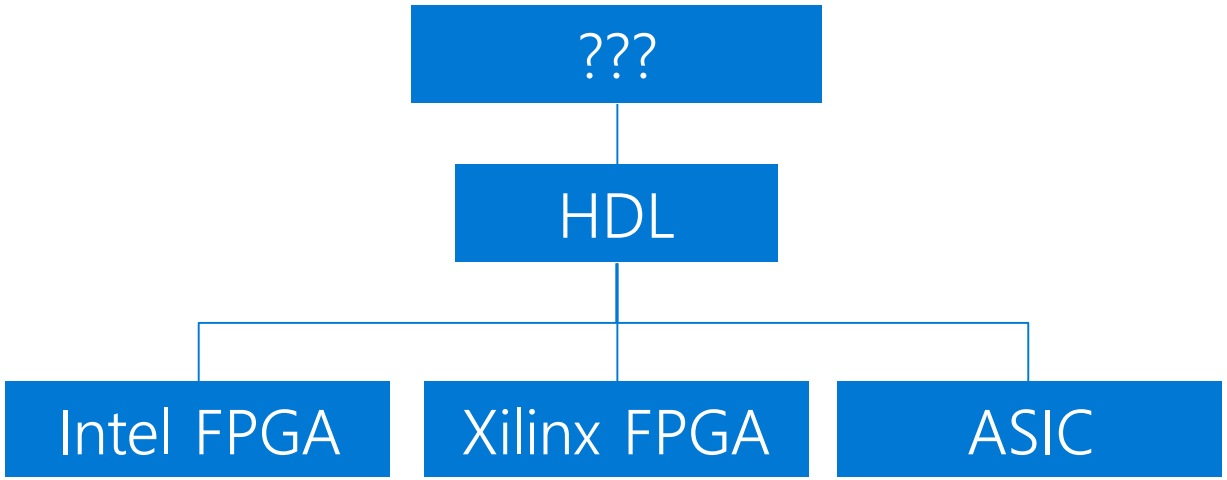
Transformers (Bert, GPT-2) have turned NLP from a data-bound problem into a compute-bound problem

This shift is transformative and is leading to monthly capability leaps

Software Developers Should Program Hardware

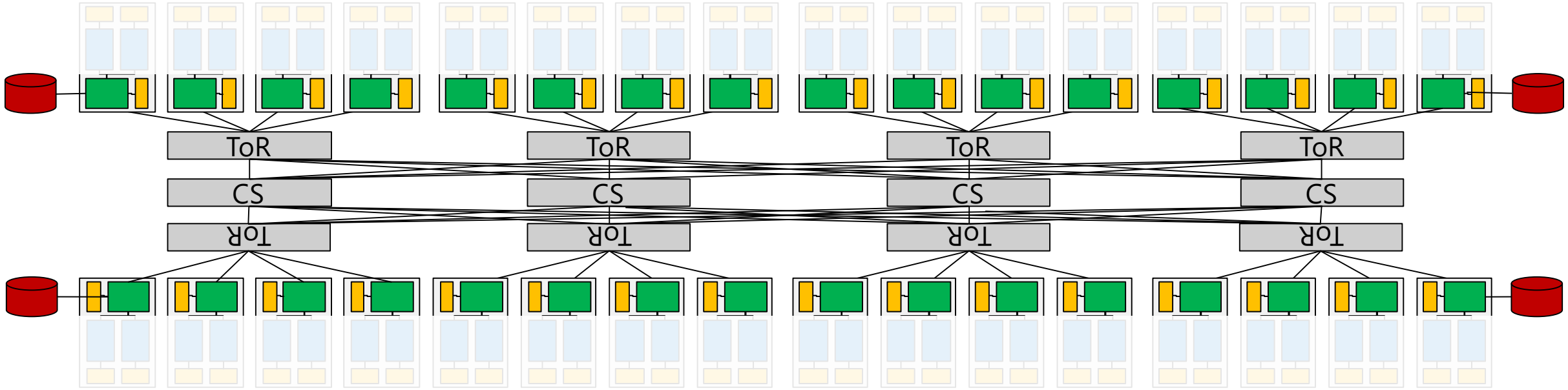
- Hide low-level details
- Expose fundamental HW characteristics
- Run cross-platform
- Low-friction testing, verification, performance tuning, debugging





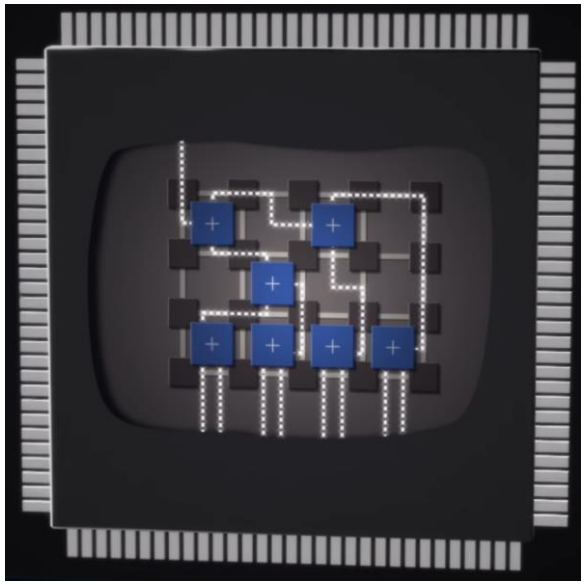
Coming Back to Post-Moore

- ✓ Cloud software optimization
- ✓ Infrastructure acceleration
- ✓ Accelerators for new domains (AI)
- Accelerators for data processing?



Project Catapult + Brainwave History

Field Programmable Gate Arrays



2011: Project Catapult Launched

2013: Bing pilot runs decision trees 40X faster

2015: Bing ranking throughput increased 2X

2016: Azure Accelerated Networking delivers industry-leading cloud performance

2017: Over 1M servers deployed with FPGAs at hyperscale

2017: Hardware Microservices harness FPGAs for distributed computing

2017: FPGAs enable real-time AI, ultra-low latency inferencing without batching; Bing launches first FPGA-accelerated Deep Neural Network

2018: Project Brainwave launched in Azure Machine Learning

Thank you!

Microsoft

Backup slides

Demo | AI for Earth

http://www.microsoft.com/aiforearth

References

How to Use FPGAs for Deep Learning Inference to Perform Land Cover Mapping on Terabytes of Aerial Images

https://blogs.technet.microsoft.com/machinelearning/2018/05/29/how-to-use-fpgas-for-deep-learning-inference-to-perform-land-cover-mapping-on-terabytes-of-aerial-images/

Pixel-Level Land Cover Classification Using the Geo AI Data Science Virtual Machine and Batch AI

https://blogs.technet.microsoft.com/machinelearning/2018/03/12/pixel-level-land-cover-classification-using-the-geo-ai-data-science-virtual-machine-and-batch-ai/

Rough: Storyboard Notes/Flow of Deck

Overall story: new architecture that's emerging; paradigm shifting; global computing and mass platform; security, AI, edge acceleration;

Straw-man flow:

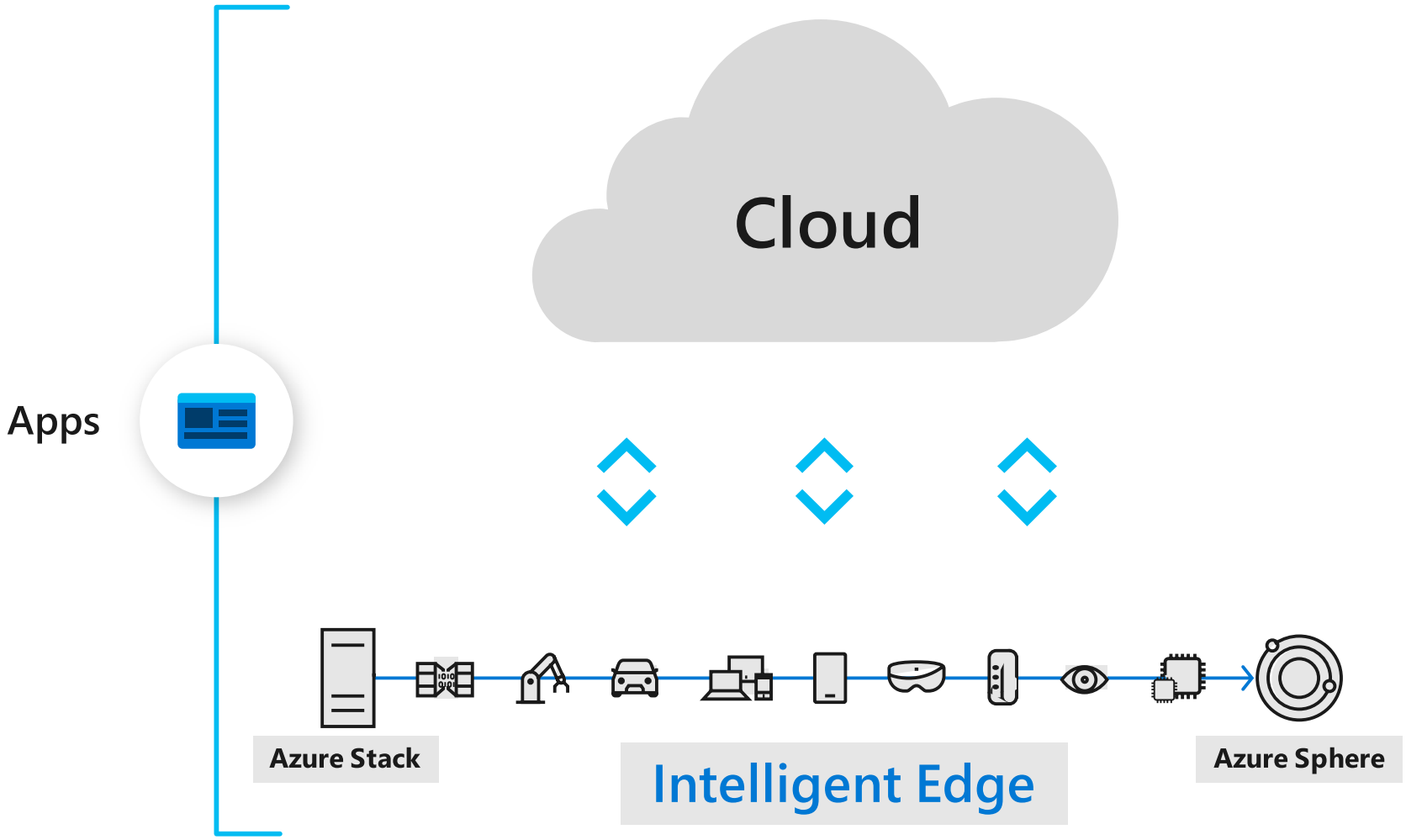
- Azure all up slide
- Cloud inflection point
- Azure global deployment and scale
- Digital transformation and what it means to companies
- Cloud and edge story. Supercomputer, Sphere (security) Starbucks example
- Data centers. Dublin DC, underwater DC
- Palix, Re-thinking traditional storage system design, co-designing future hardware & software infrastructure for the cloud and building a completely new storage system
- AI: OpenAI deal; general trend toward super AI
- End on an aspirational note…i.e. quantum

Links to other deck:

1. [Link to top level folder]

2. Mark Russinovich Presentations including
   a) Mark LinkedIn deck
   b) Inside Azure Architecture
   c) Mark's Build deck

Tomorrow



Azure Datacenters & Architecture

Azure global infrastructure

54 total Azure regions: 44 generally available + 10 coming soon

Azure servers: General purpose

RAM ⟷ Cores

3x Beast

0.00000000533 Beast

Gen 2	
Processor	2 x 6 Core 2.1 GHz
Memory	32 GiB
Hard Drive	6 x 500 GB
SSD	None
NIC	1 Gb/s

Godzilla	
Processor	2 x 16 Core 2.0 GHz
Memory	512 GiB
Hard Drive	None
SSD	9 x 800 GB
NIC	40 Gb/s

Gen 6	
Processor	2 x Skylake 24 Core 2.7GHz
Memory	768GiB DDR4
Hard Drive	None
SSD	4 x 960 GB M.2 SSDs and 1 x 960 GB SATA
NIC	40 Gb/s + FPGA

Optimized Gen 6	
Processor	2 x 24 core Skylake Lake
Memory	192 GB DDR4
Hard Drive	None
SSD	4 x 960 GB M.2 NVMe
NIC	40 Gb/s + FPGA

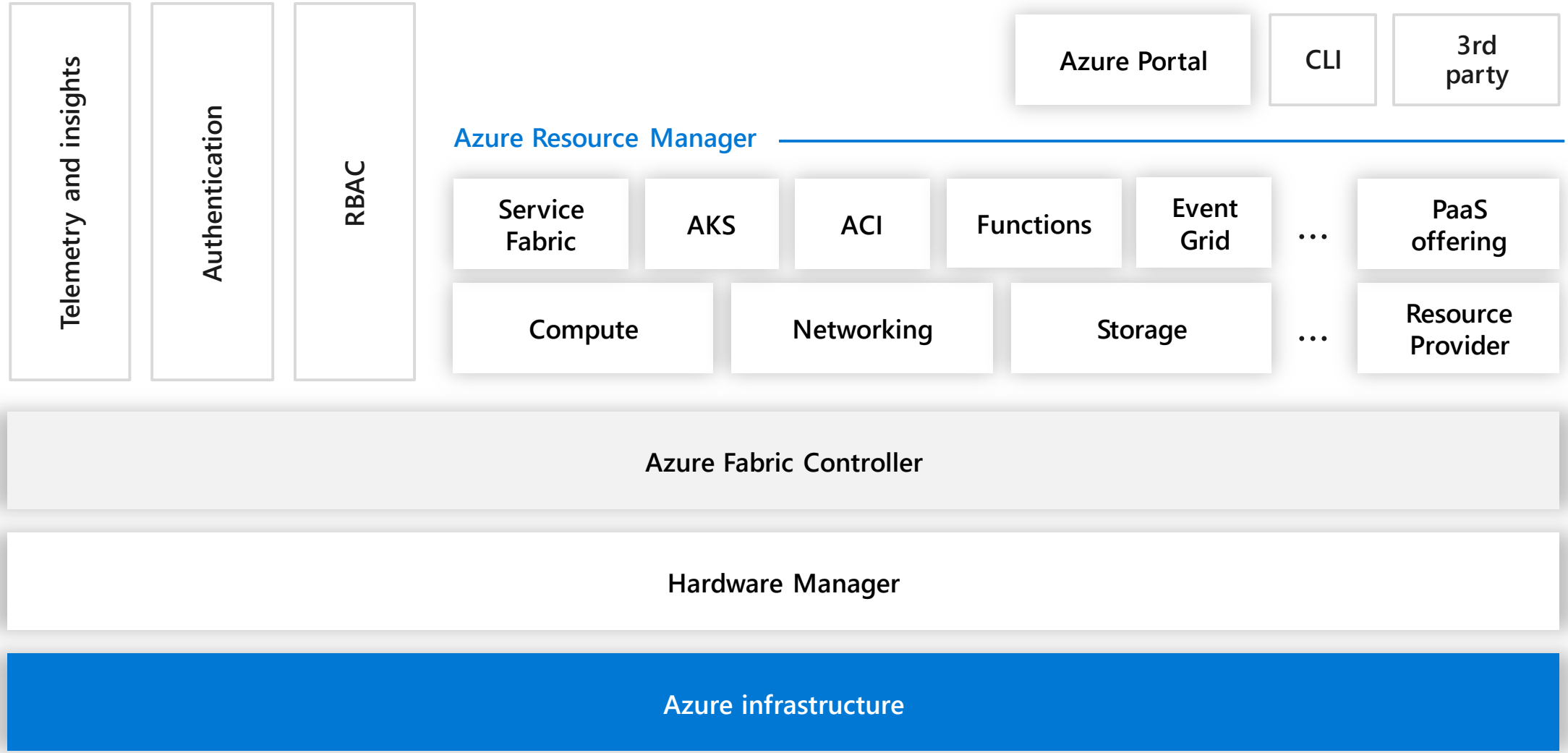
Optimized Gen 7	
Processor	2 x 26 core Cascade Lake
Memory	192 GB DDR4
Hard Drive	None
SSD	5 x 960 GB M.2 NVMe
NIC	50 Gb/s + FPGA

Beast	
Processor	4 x 18 Core 2.5 GHz
Memory	4096 GiB
Hard Drive	None
SSD	4 x 2 TB NVMe, 1 x 960 GB SATA
NIC	40 Gb/s

Beast v2	
Processor	8 x 28 Core 2.5 GHz
Memory	12 TiB
Hard Drive	None
SSD	4 x 2 TB NVMe, 1 x 960 GB SATA
NIC	50 Gb/s

Azure Sphere	
Processor	2 x M4 Core @ 200 MHz
Memory	64KB RAM
WiFi	2.4/5.0 GHz 802.11 b/g/n

Azure architecture



Azure Front Door Service

Web Application Firewall integration at edge

Global, inline, always on

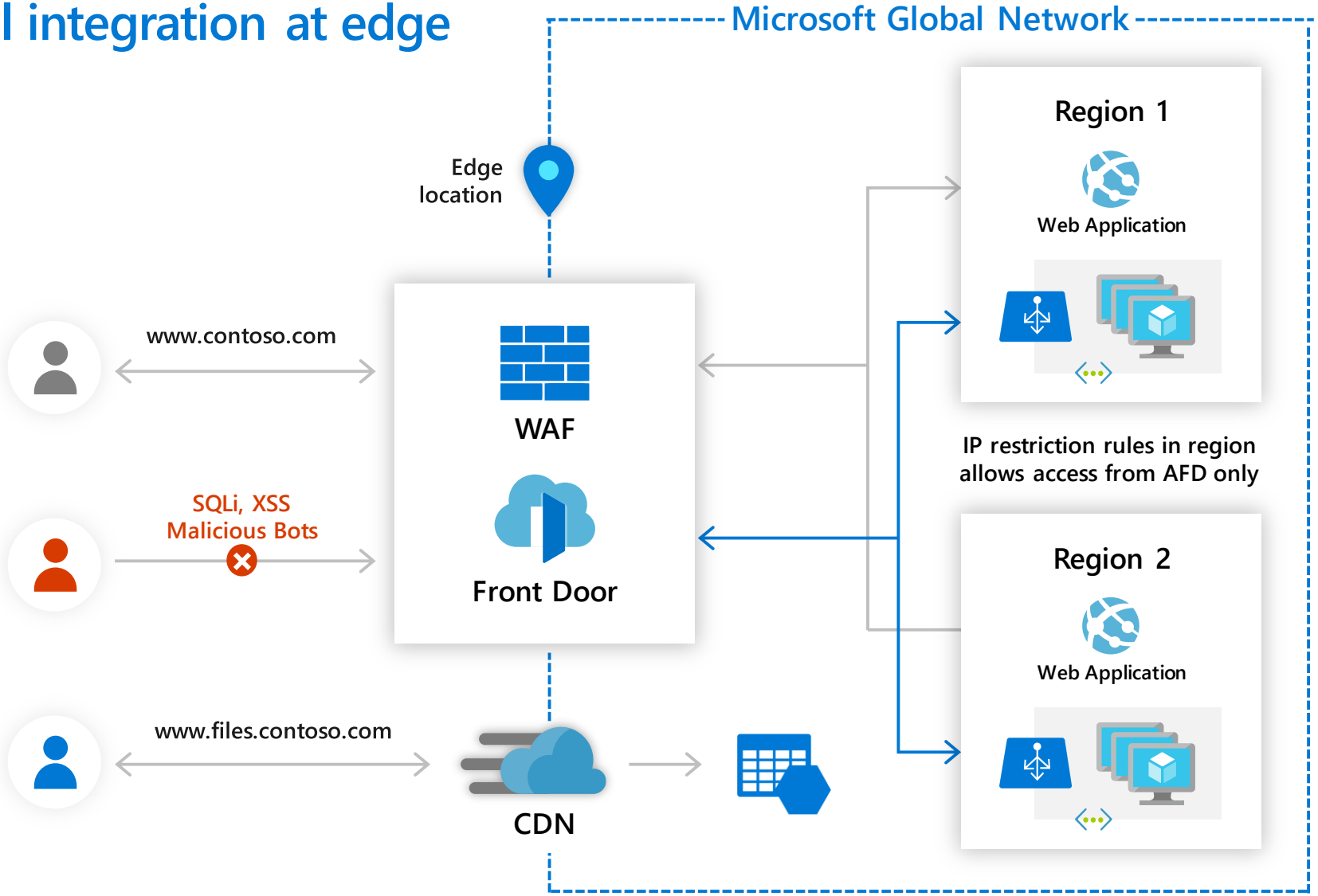
Customizable protection

- IP restriction
- Geo filtering
- http parameters
- request methods
- size restriction

Conditional rate limiting

Preconfigured managed ruleset against OWASP TOP 10 common vulnerabilities

Bot detection and mitigation based on IP reputation



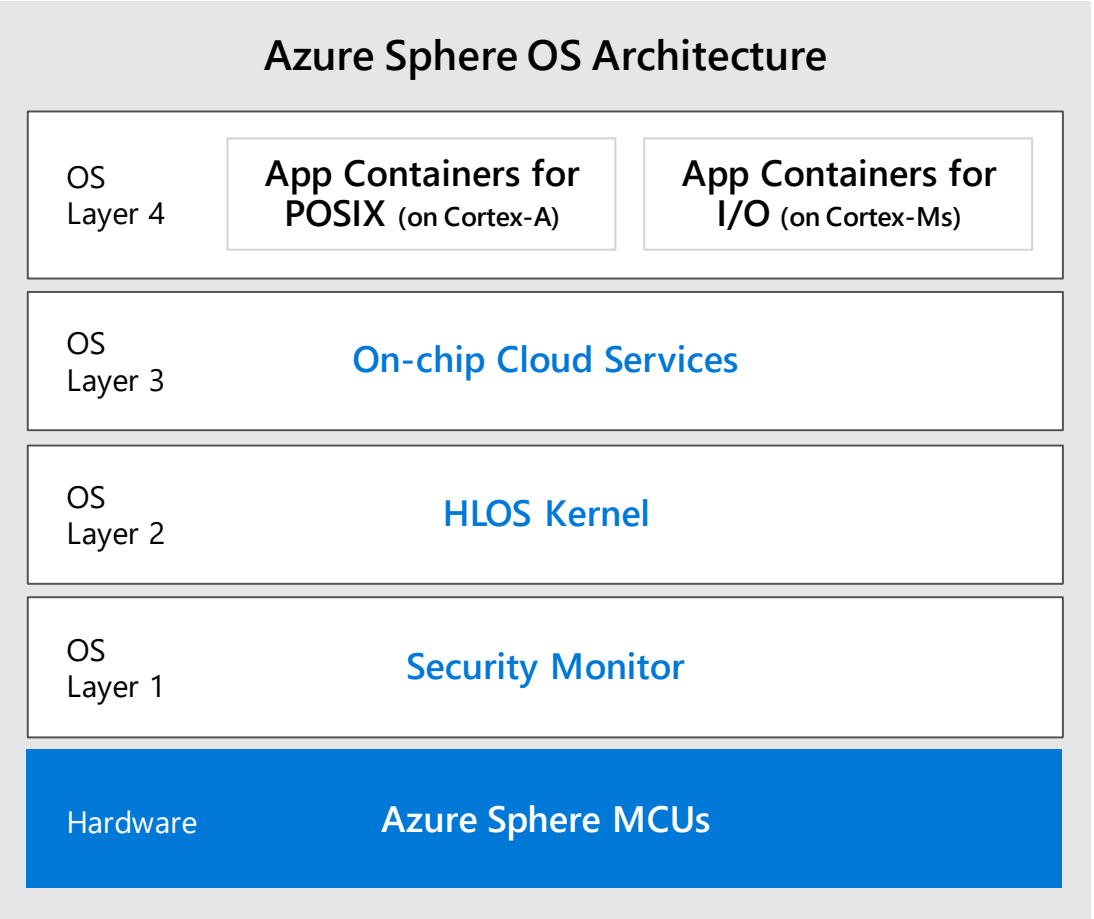
The Azure Sphere OS is optimized for IoT, security, and agility

Secure Application Containers
Compartmentalize code for agility, robustness & security

On-chip Cloud Services
Provide update, authentication, and connectivity

Custom Linux kernel
Empowers agile silicon evolution and reuse of code

Security Monitor
Guards integrity and access to critical resources

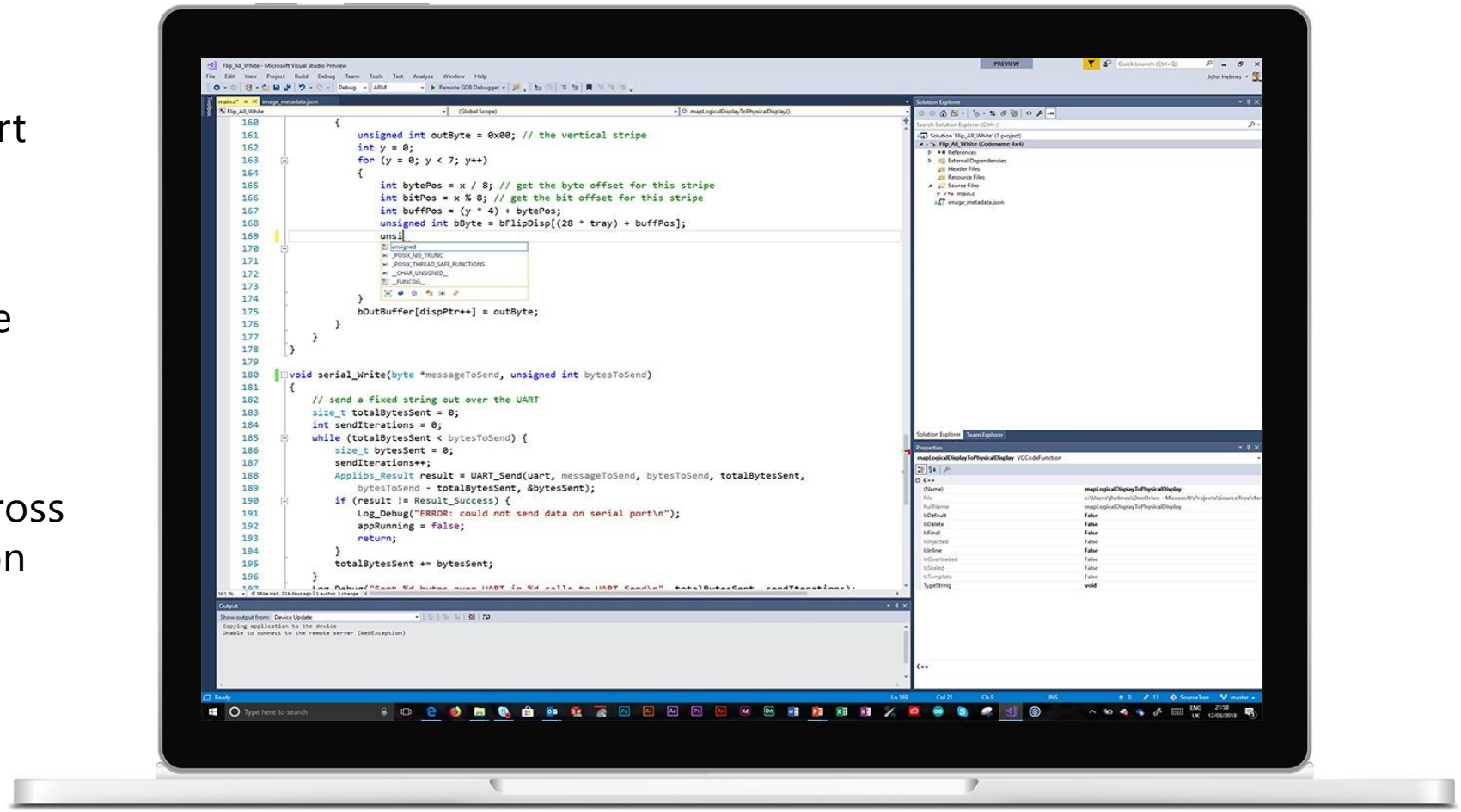


Modernize MCU development with Azure Sphere and Visual Studio

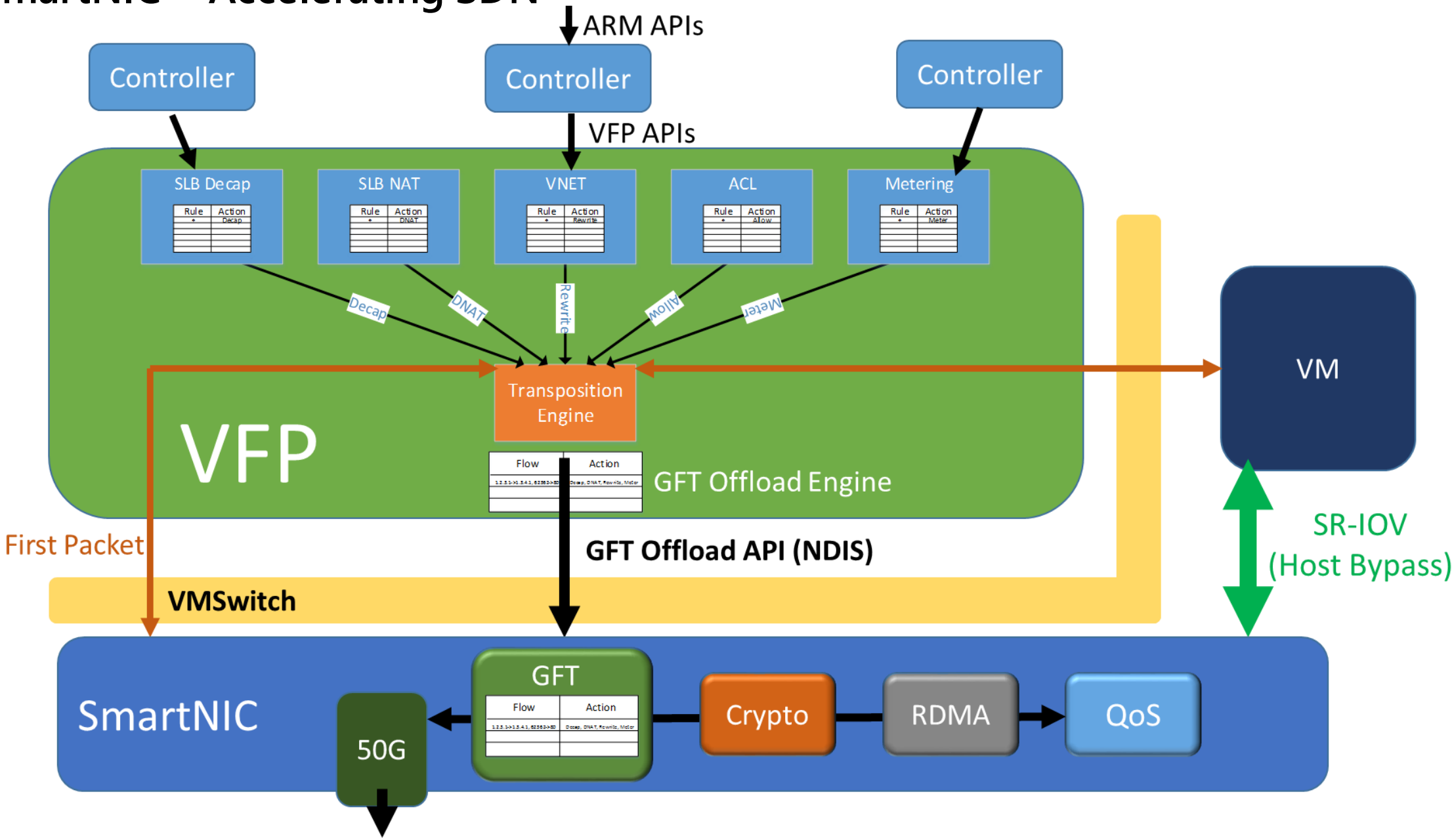
Simplify development
Focus your device development effort on the value you want to create

Streamline debugging
Experience interactive, context-aware debugging across device and cloud

Collaborate across your team
Apply tool-assisted collaboration across your entire development organization

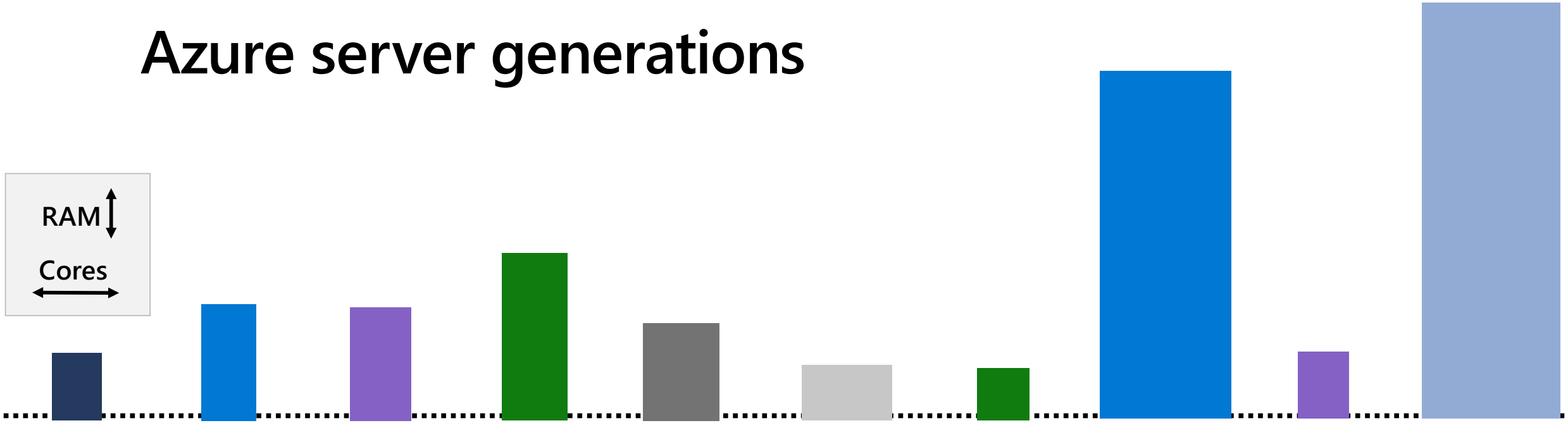


Azure SmartNIC – Accelerating SDN



Inside Azure Servers

Azure server generations



Gen 2	
Processor	2 x 6 Core 2.1 GHz
Memory	32 GiB
Hard Drive	6 x 500 GB
SSD	None
NIC	1 Gb/s

Gen 3	
Processor	2 x 8 Core 2.1 GHz
Memory	128 GiB
Hard Drive	1 x 4 TB
SSD	5 x 480 GB
NIC	10 Gb/s

HPC	
Processor	2 x 12 Core 2.4 GHz
Memory	128 GiB
Hard Drive	5 x 1 TB
SSD	None
NIC	10 Gb/s IP, 40 Gb/s IB

Gen 4	
Processor	2 x 12 Core 2.4 GHz
Memory	192 GiB
Hard Drive	4 x 2 TB
SSD	4 x 480 GB
NIC	40 Gb/s

Godzilla	
Processor	2 x 16 Core 2.0 GHz
Memory	512 GiB
Hard Drive	None
SSD	9 x 800 GB
NIC	40 Gb/s

Gen 5.1	
Processor	2 x 20 Core 2.3 GHz
Memory	256 GiB
Hard Drive	None
SSD	6 x 960 GB PCIe Flash and 1 x 960 GB SATA
NIC	40 Gb/s + FPGA

GPU Gen 5	
Processor	2 x 8 Core 2.6 GHz
Memory	256 GiB
Hard Drive	1 x 2 TB
SSD	1 x 960 GB SATA
NIC	40 Gb/s
GPU	2 x 2 Compute GPU

Beast	
Processor	4 x 18 Core 2.5 GHz
Memory	4096 GiB
Hard Drive	None
SSD	4 x 1920 GB NVMe and 1 x 960 GB SATA
NIC	40 Gb/s

Gen 6	
Processor	2 x Skylake 24 Core 2.7GHz
Memory	768GiB DDR4
Hard Drive	None
SSD	4 x 960 GB M.2 SSDs and 1 x 960 GB SATA
NIC	40 Gb/s
FPGA	Yes

Mega Beast	
Processor	4 x 18 Core 2.5 GHz
Memory	12,288 GiB
Hard Drive	None
SSD	4 x 1920 GB NVMe and 1 x 960 GB SATA
NIC	40 Gb/s

Project Olympus

Flexible and Modular design fo
Open Compute Project open s

Compute
Intel, AMD, ARM64 CPUs
High density GPU expansion fo
NVM (DRAM+battery) and 3D

Storage
High density HDD and Flash e
Microsoft custom designed SS

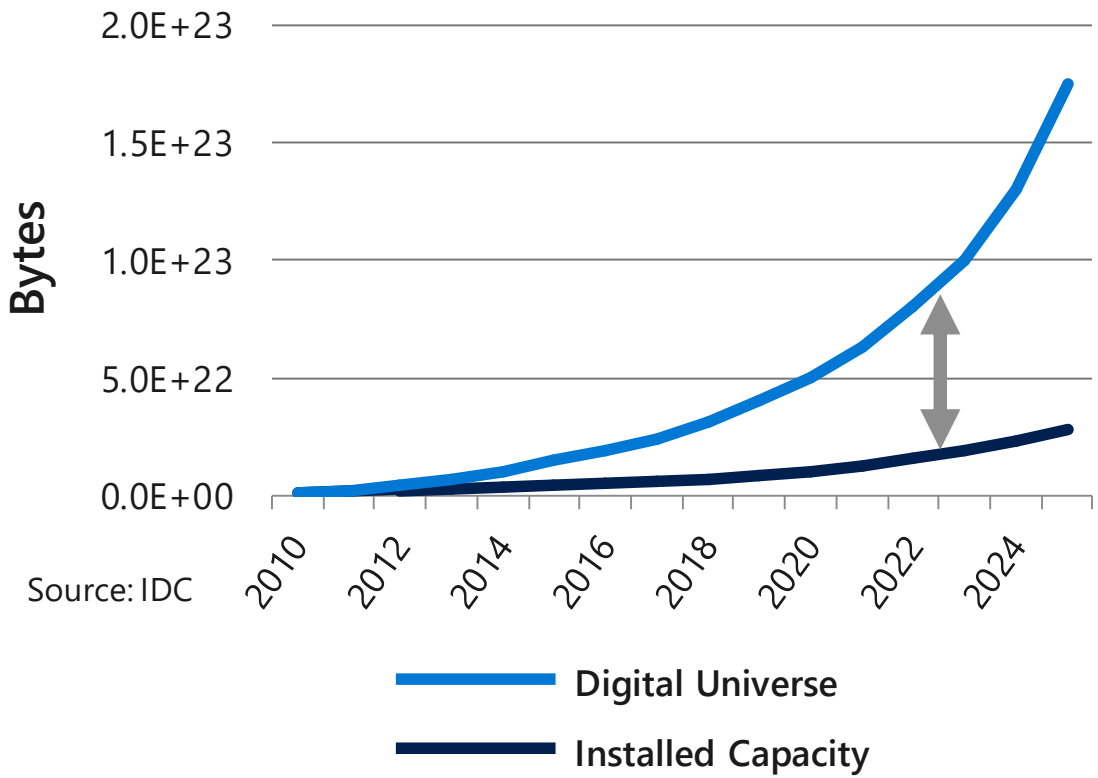
Networking
40 going up to 50 Gbps netwo
Accelerated VMs using FPGAs

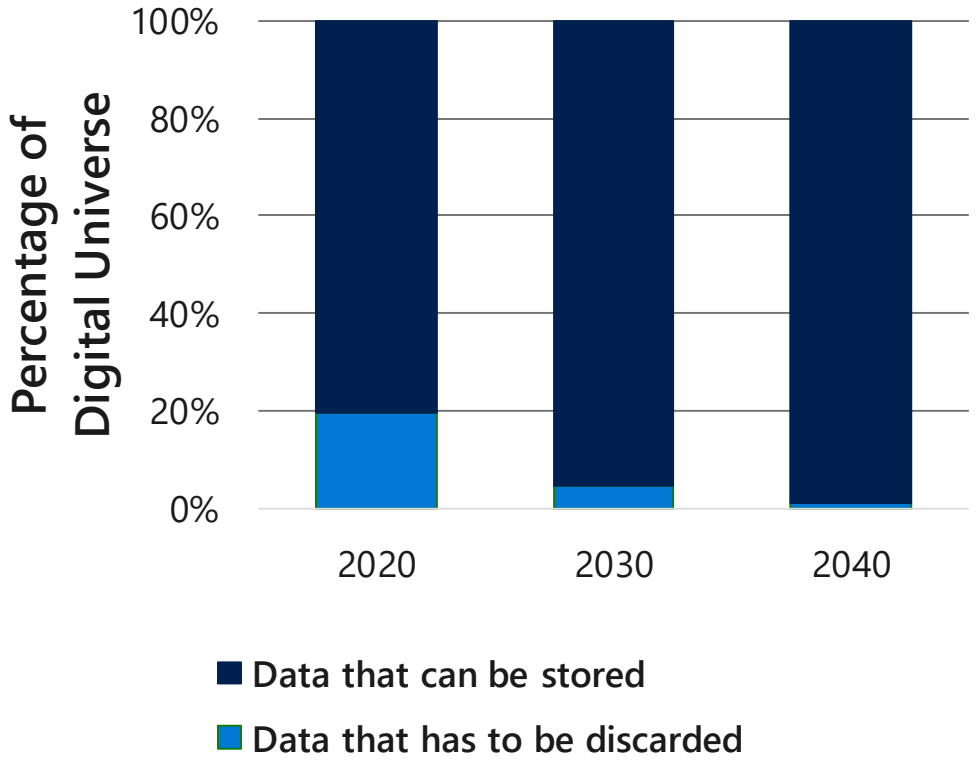
See larger image

IDG

25. Microsoft Project Olympus

Storage capacity is growing too slowly

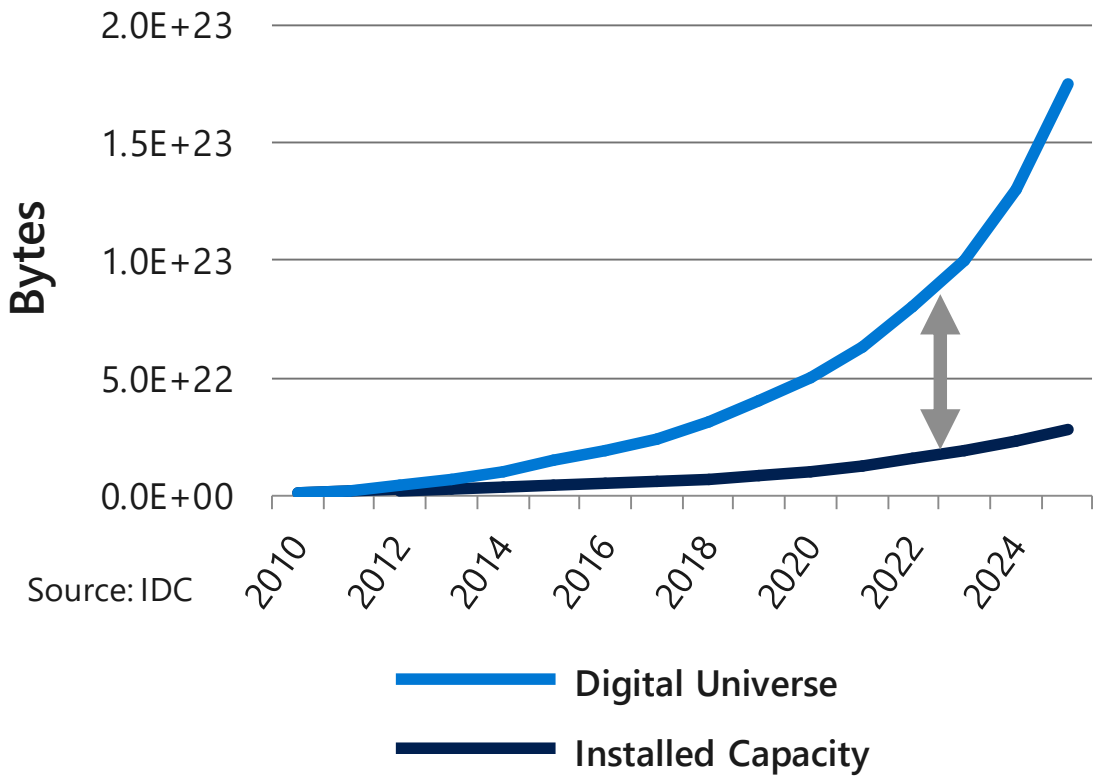


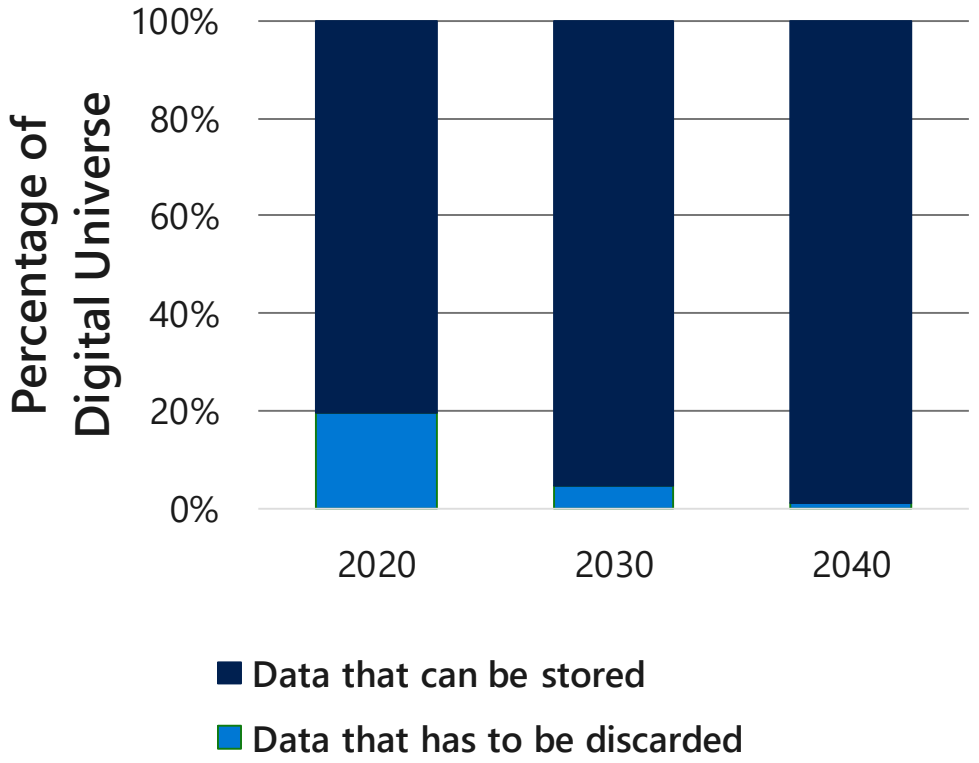


Storing Digital Data in Synthetic DNA

Oct 31, 2019

Storage capacity is growing too slowly





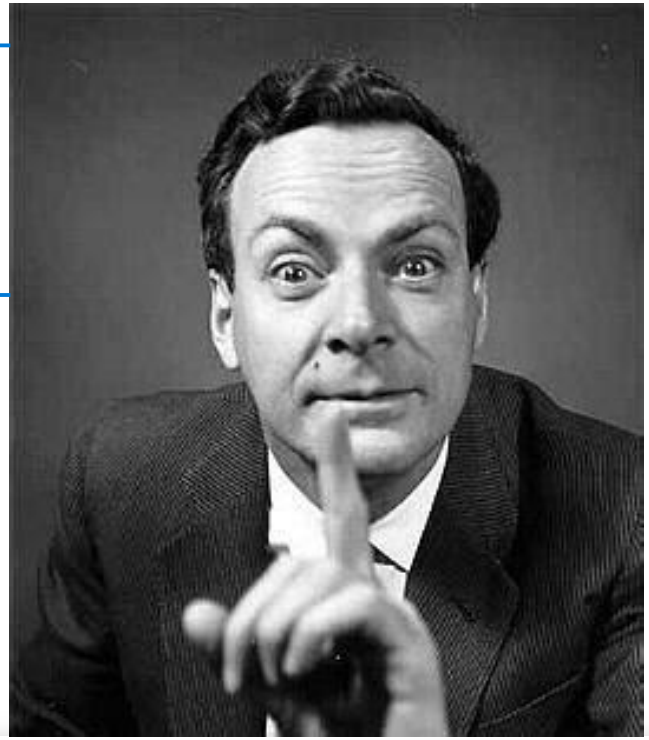
When Moore met Feynman

The number of transistors doubles every 18-24 months

The industry roadmaps are based on that continued rate of improvement

Arrange the atoms the way we want

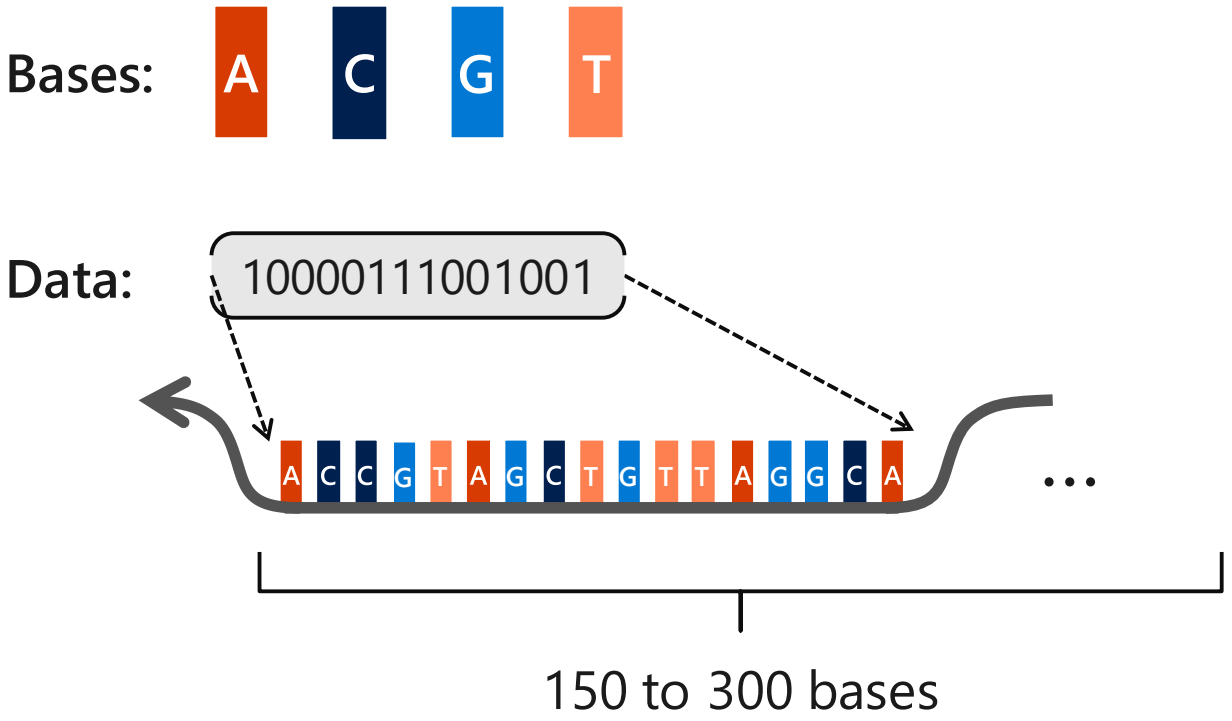
DNA molecules use approximately 50 atoms for one bit



Let’s store data in DNA!

Norbert Wiener, U.S. News and World Rep, 1964; Mikhail Neiman, Radiotechnika, 1965; Church et al, Science, 2012; Goldman et al, Nature, 2013

DNA data storage basics



Simple mapping:

Bits	Base
00	A
01	C
10	G
11	T

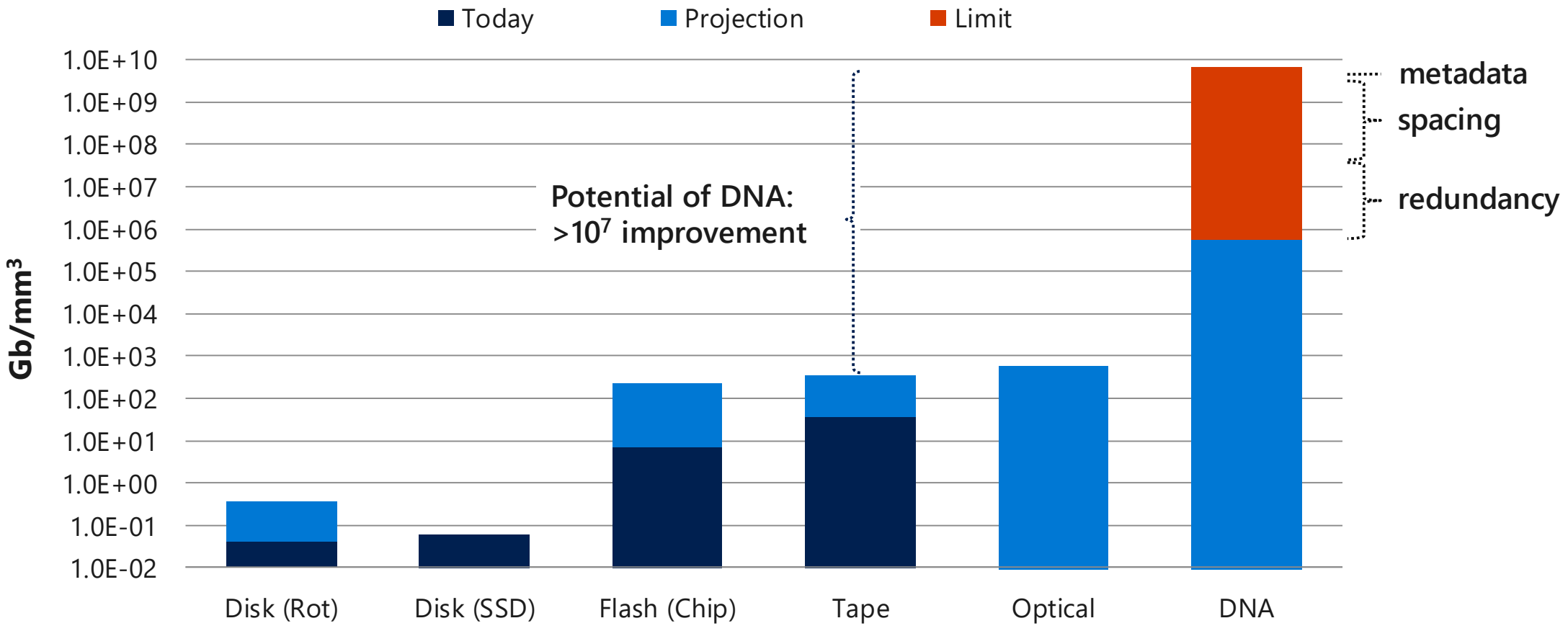
Store data in synthetic DNA strands

Dense, really dense

Cold Storage: 1EB
Size: Two Walmart Supercenters

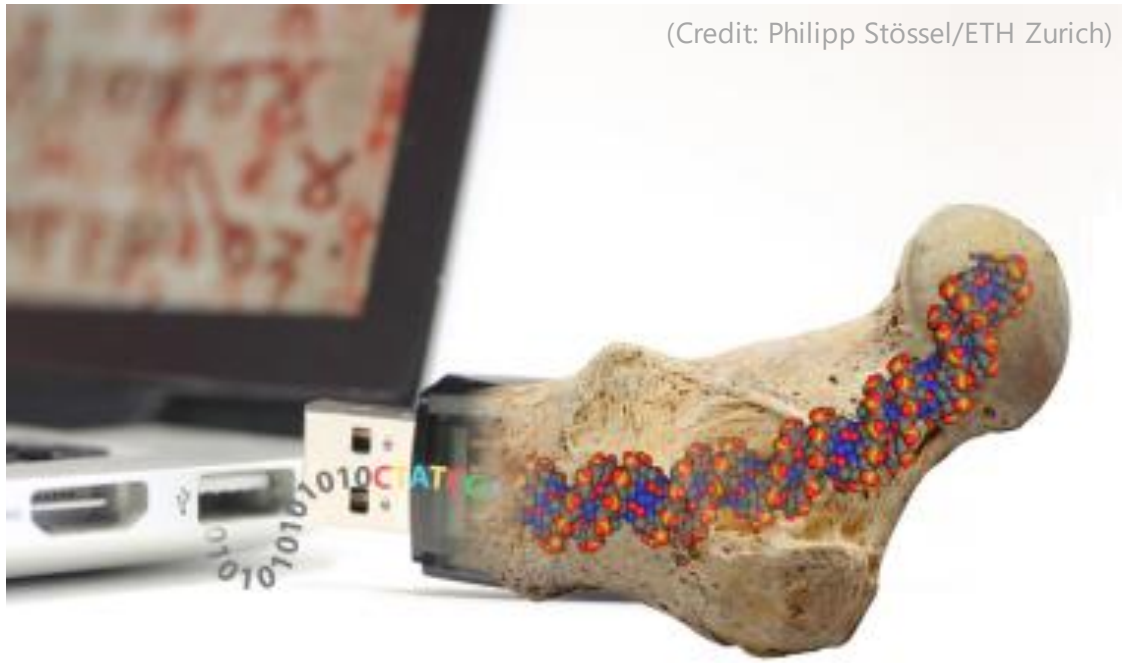
vs.

Comparison with other media



Information durability

DNA “synthetic fossils” survive 2,000 – 2,000,000 years



Extreme density makes these conditions cheap and easy to keep

No obsolescence issue, DNA will always be relevant

Size of a mainframe

800 Kbases/day

Size of a workstation

80 Gbases/day

Size of a portable SSD

10 Gbases/day

Same medium as read technology improves:

No obsolescence issue, DNA will always be relevant

Same medium as read technology improves:

Medium changes as read and write technology improves:




Brainwave

Problem

How to improve performance for batch-of-1 inferences at scale with low cost?

Goals

Provide large-scale, performant, and affordable real-time AI inference services

Solution

Hardware accelerated models using FPGAs

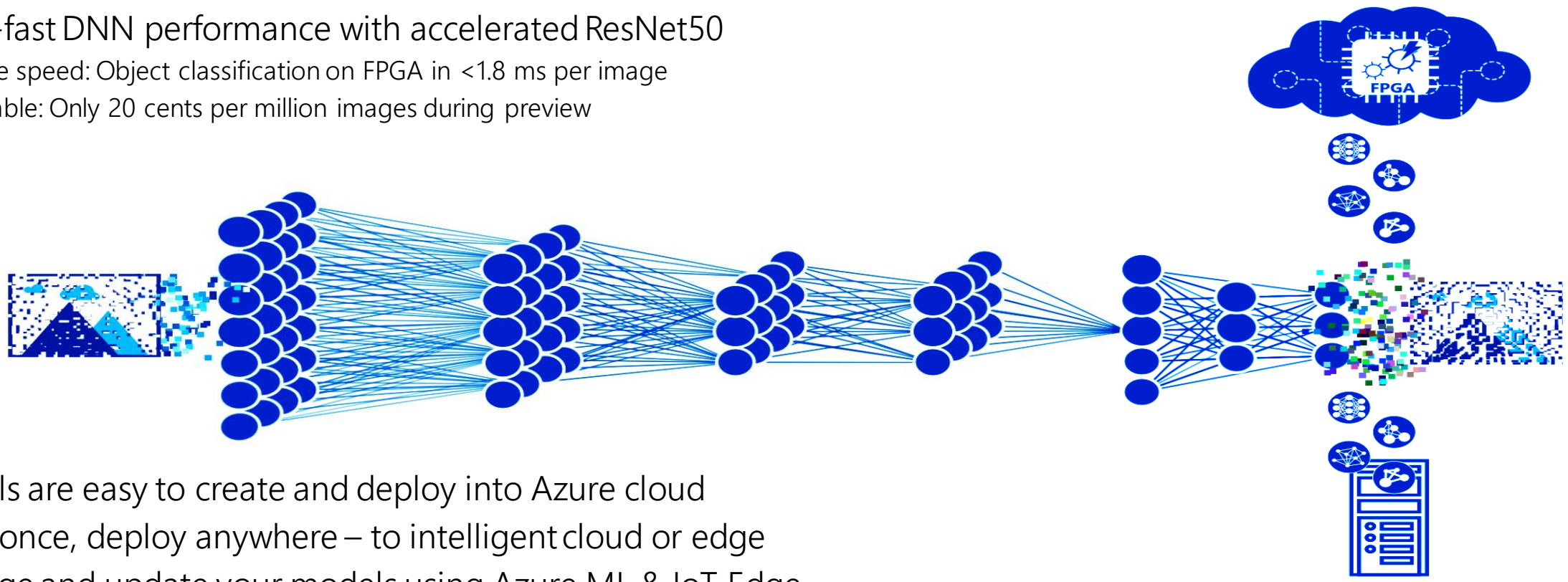
Azure Machine Learning Accelerated by Project Brainwave

Real-time AI at cloud scale with phenomenal performance and affordable cost

Ultra-fast DNN performance with accelerated ResNet50

Extreme speed: Object classification on FPGA in <1.8 ms per image

Affordable: Only 20 cents per million images during preview



Models are easy to create and deploy into Azure cloud

Write once, deploy anywhere – to intelligent cloud or edge

Manage and update your models using Azure ML & IoT Edge

http://aka.ms/aml-real-time-ai

Using Project Brainwave for land use mapping

Using FPGAs for ultra-fast inferencing different types of land use for the entire United States (ESRI NAIP Data, 20+ TB)

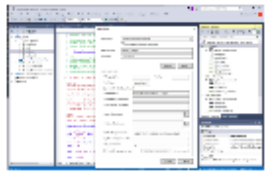
Data | Build | Train | Deploy | Intelligent Apps

esri Satellite Images for the Entire US

NAIP Data
20TB, 200M images

Stored on Azure Premium Storage

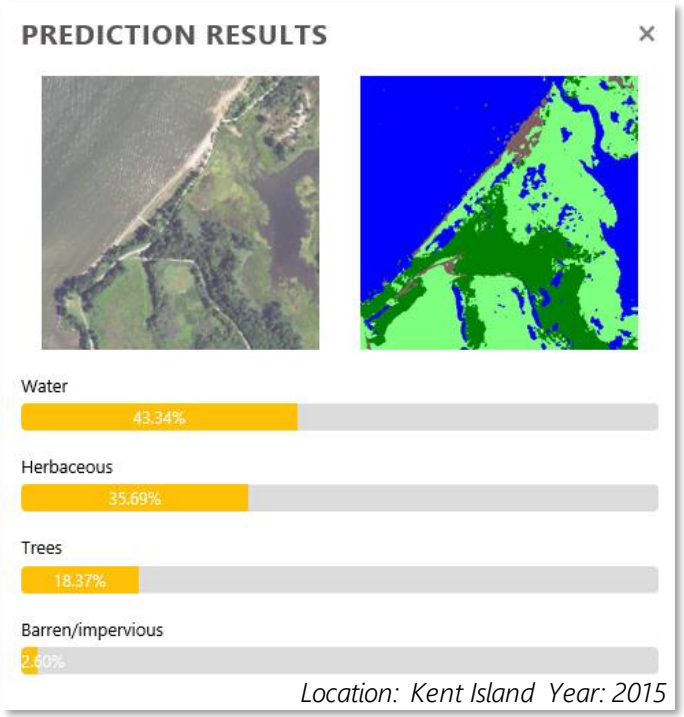
Geo AI Data Science Virtual Machine



Visual Studio Tools for AI

Azure Machine Learning

Azure Batch AI



Land Classification Model ResNet-50

Ultra-fast Inferencing using FPGAs

Data ---------- Intelligence ---------- Action

Bing models in production with MSFP on Brainwave

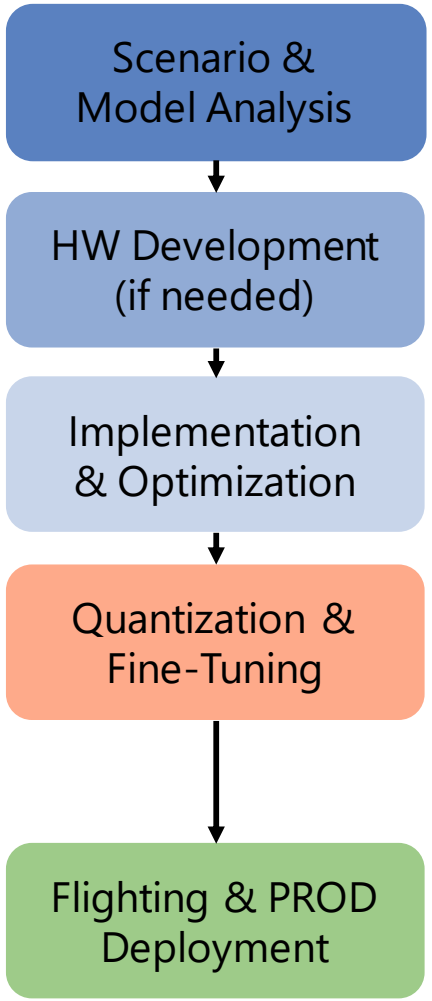
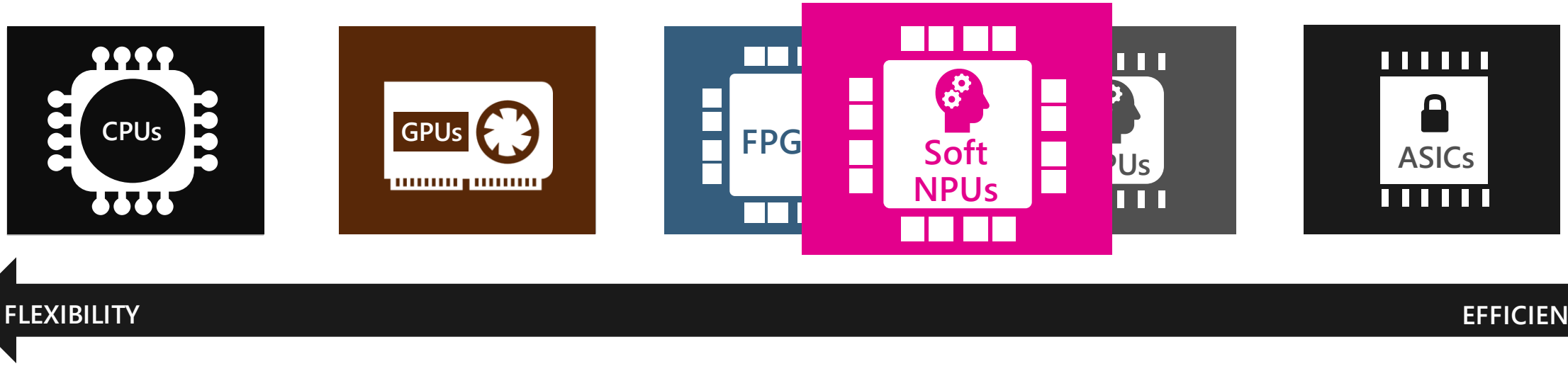
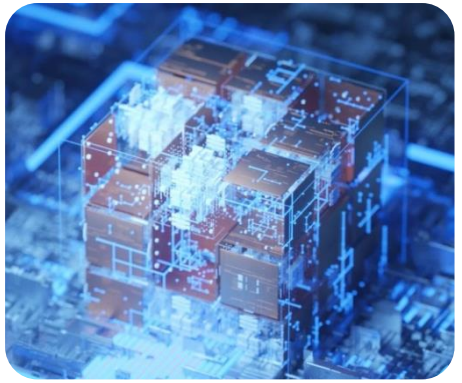


Image Answer	Answers for Multimedia Images
QP MLBERT	Quantus query passage ranking
Bling SD	Web Intelligence model to produce semantic document body
Universal Deep Think v1	Intelligent Search (QnA, Captions) & Ranking
Smart Reply, Meeting Insights, Key Points	Intelligence for M365
Vortex	Intelligent Search (QnA, Captions) & Ranking
WaveNet & QRNN	Text to speech for Bing Mobile & Microsoft Assistant Commute
ANN	Approximate nearest neighbor search
TP1, Deep Think v2, Deep Think v2.1, Deep Scan, Deep Scan v2, Deep Scan v2.1, Deep Read, Poly Math v4.1, Poly Math v5, Poly Math 5.1, Universal PM1, AGI encoder/decoder	Intelligent Search (QnA, Captions) & Ranking
LiRIC	List Reading and Inference using Comprehension
ResNet-50, ResNet-152, DenseNet-121, VGG-16, SSD-VGG	Computer vision scenarios for Azure ML/3P

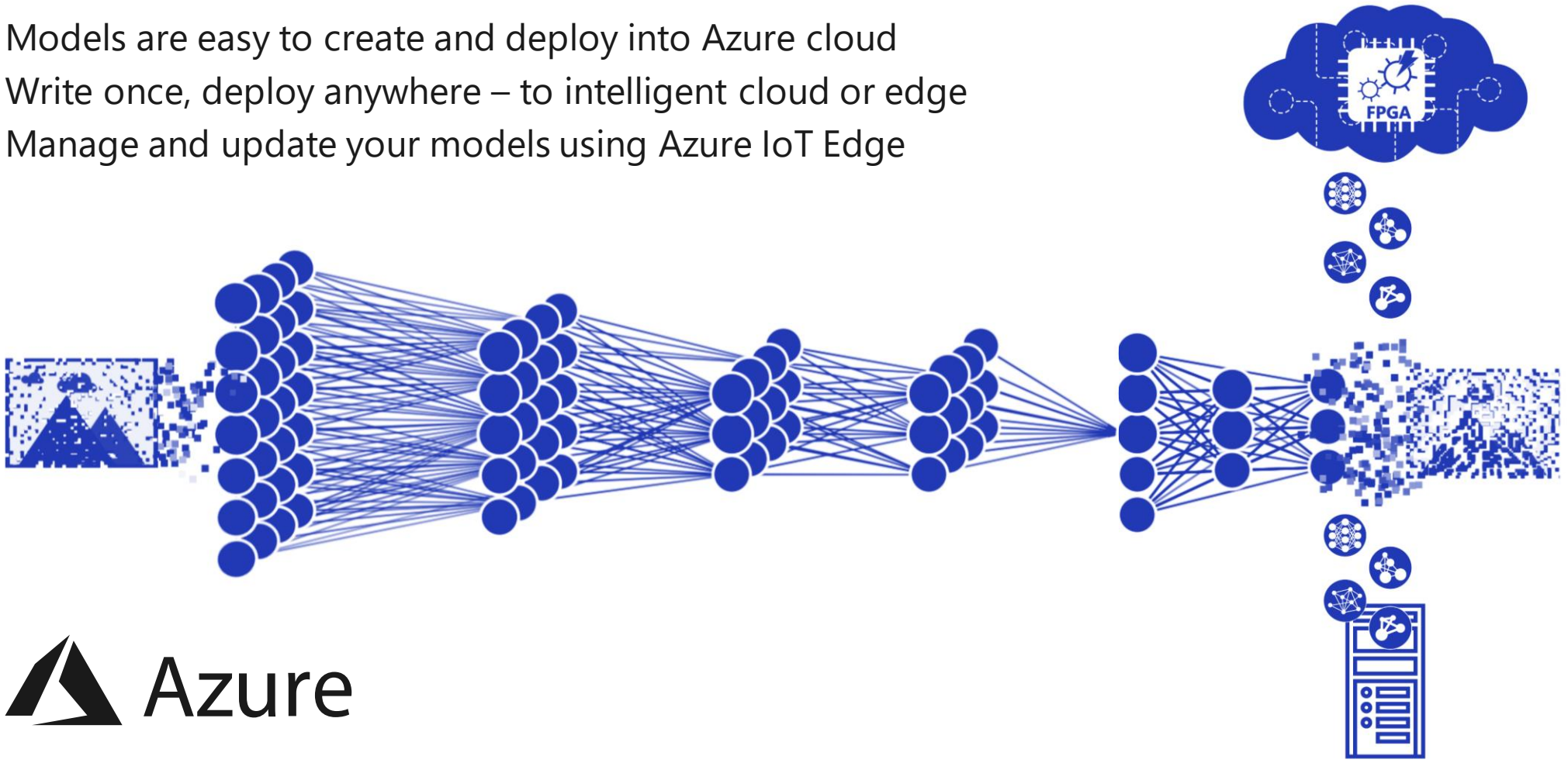
The Rise of Specialized Hardware for AI





Try Azure ML with Project Brainwave on cloud or edge

Models are easy to create and deploy into Azure cloud
Write once, deploy anywhere – to intelligent cloud or edge
Manage and update your models using Azure IoT Edge



Azure

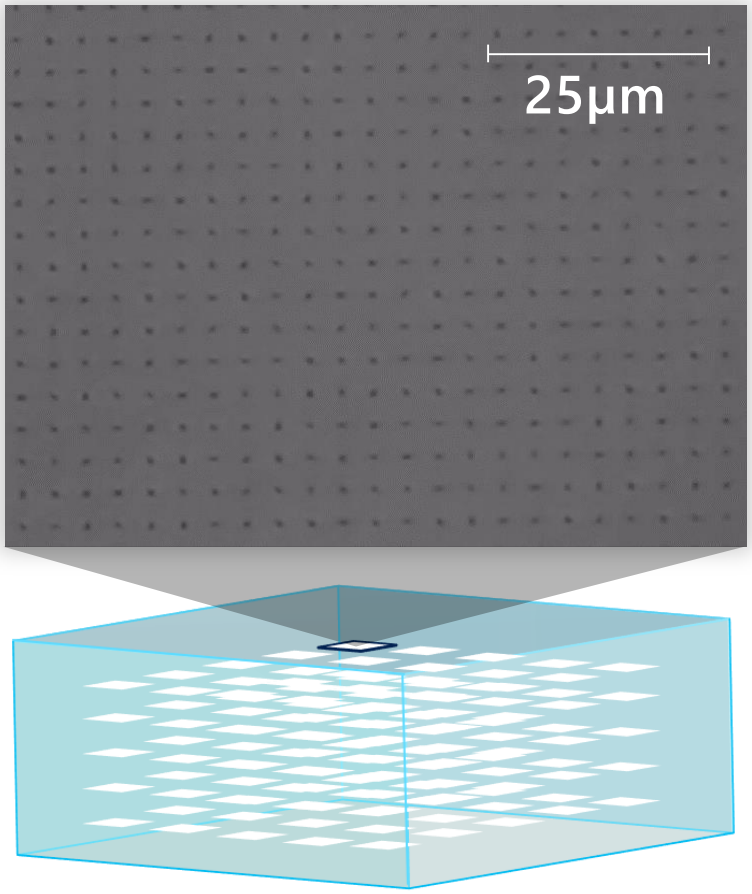
http://aka.ms/rtai
brainwave-edge@microsoft.com

Project Natick

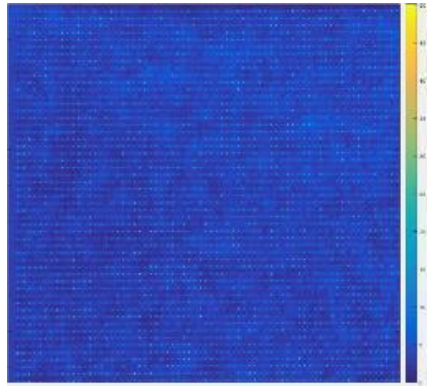
Project Silica

Data stored in quartz glass

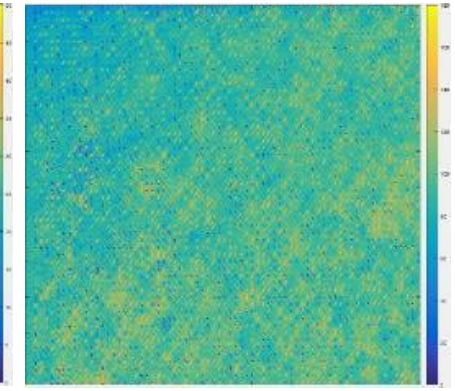
Written using femtosecond lasers

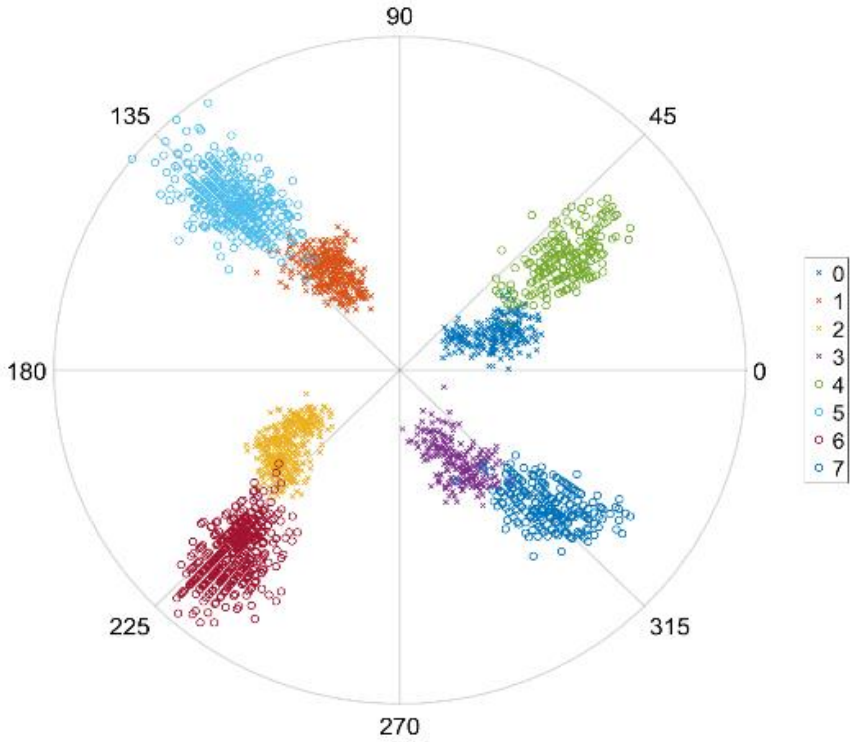


Retardance

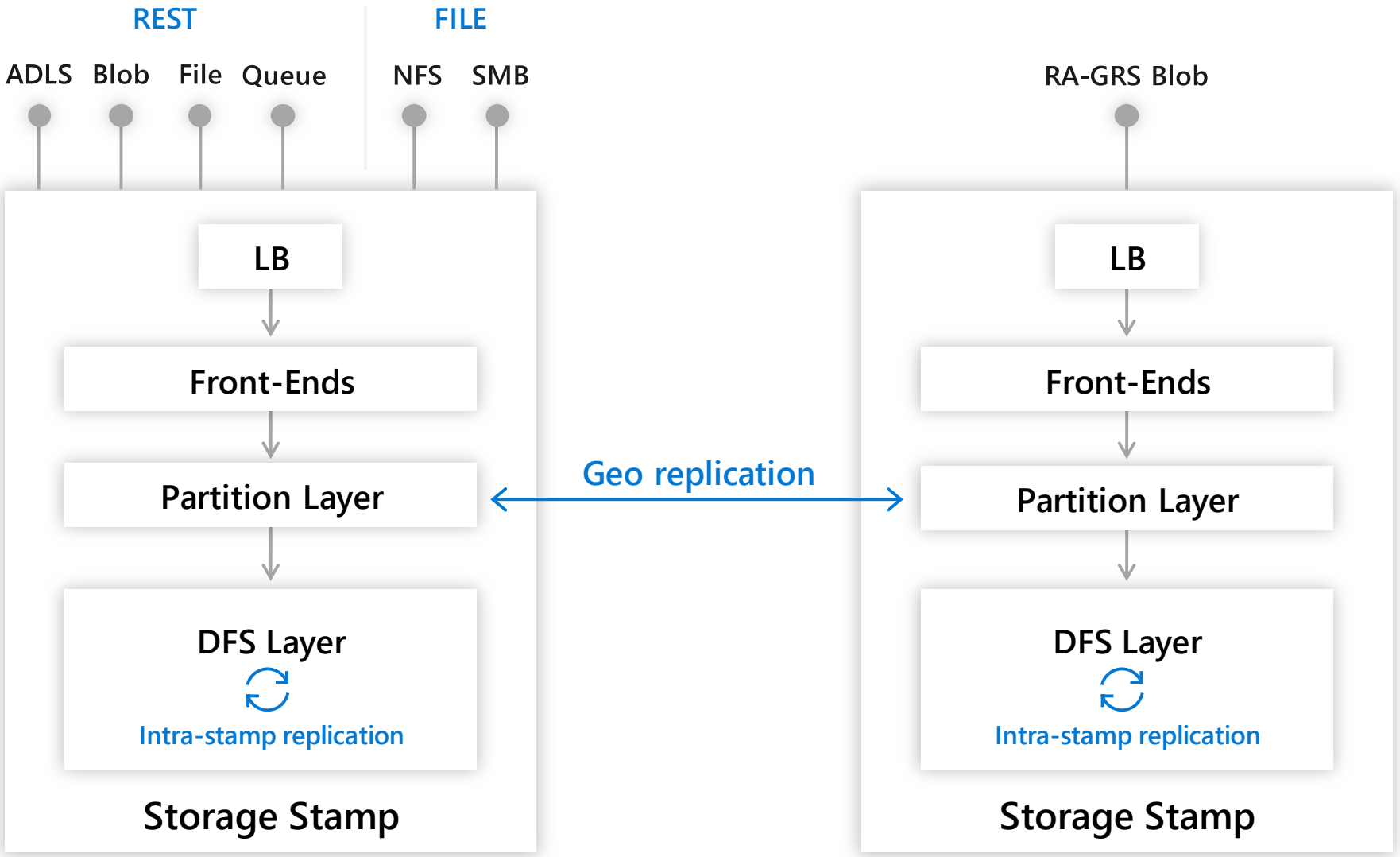


Polarization Angle





Azure Storage



Cloud Scale Analytics