

The Catapult Project - An FPGA view of the Data Center (Or) Teaching Deployed Data Center new Tricks

Dan Fay
Microsoft Research

Derek Chiou
Microsoft Azure Cloud Silicon
UT Austin

Today's Data Centers

- O(100K) servers/data center
- Tens of MegaWatts, difficult to power and cool
- Very noisy
- Security taken very seriously
- Incrementally upgraded
 - 3 year server depreciation, upgraded quarterly
- Applications change very rapid (weekly, monthly)
- Many advantages including economies of scale, data all in one place, etc.
 - Very high volumes means every efficiency improvement counts
- At data center scales, don't need to get an order of magnitude improvement to make sense
 - Positive ROI at large scale easier to achieve
- ***How can we improve efficiencies?***

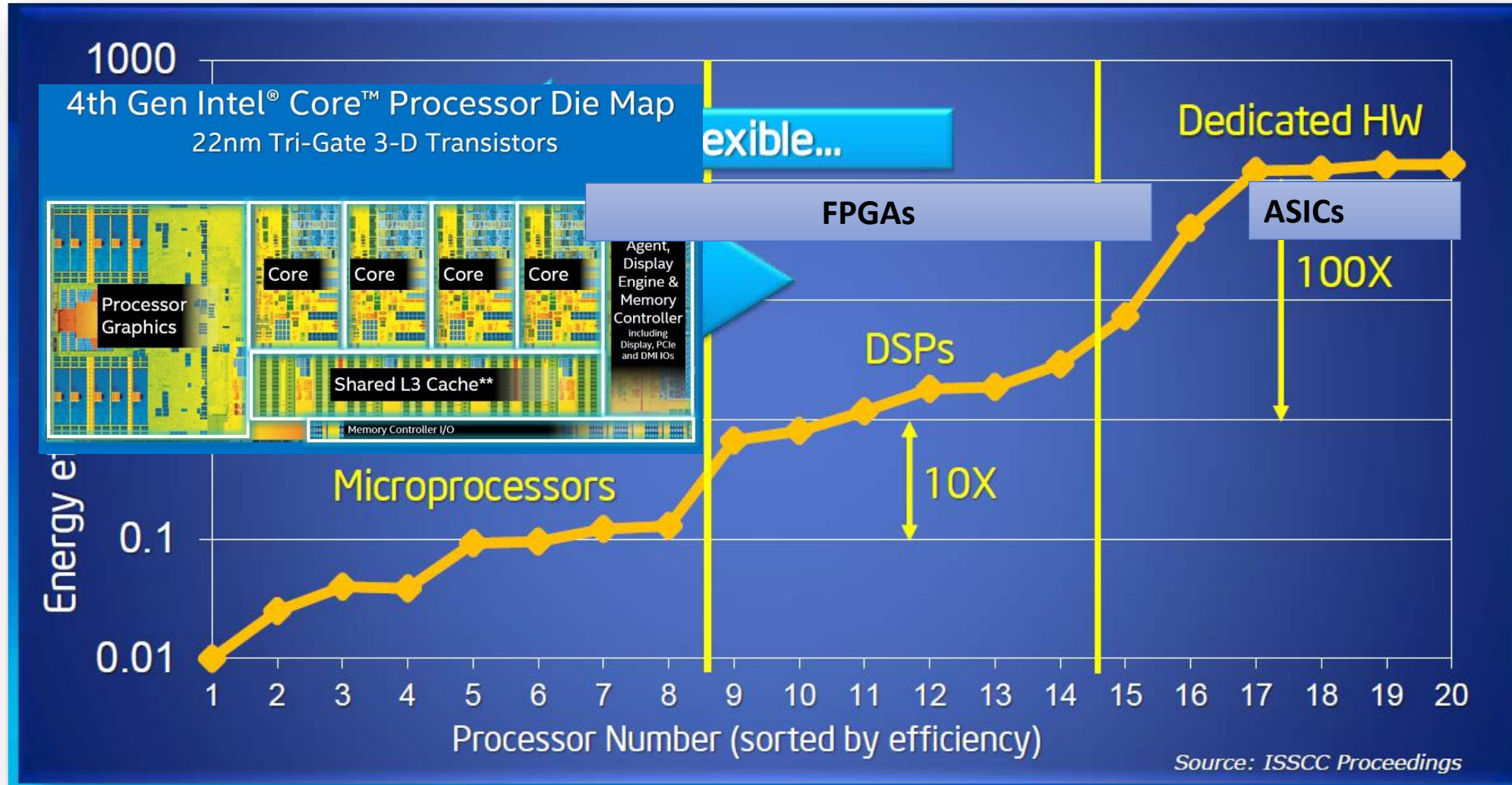


Microsoft Cloud Services

Microsoft Azure



Efficiency via Specialization



Source: Bob Broderson, Berkeley Wireless group

Original Design Requirements

Don't Cost Too Much

<30% Cost of Current Servers

1. Specialize HW with an FPGA Fabric
2. Keep Servers Homogeneous

Don't Burn Too Much Power

<10% Power Draw
(25W max, all from PCIe)

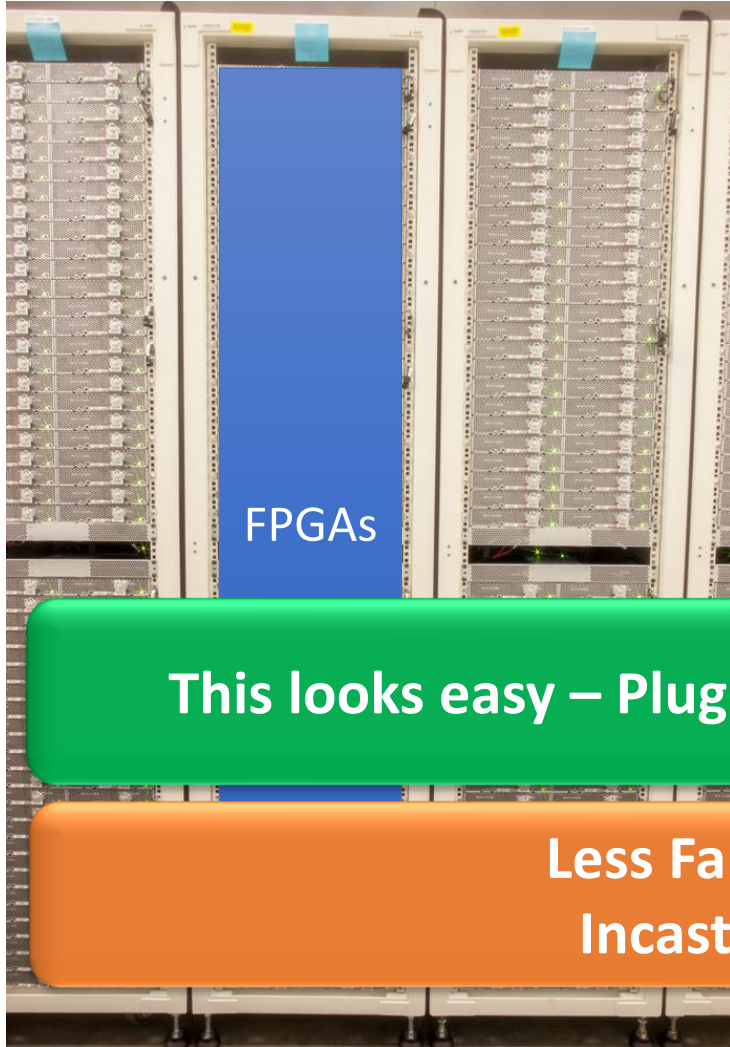
Don't Break Anything

Work in existing servers
No Network Modifications
Do not increase hardware failure rate



Version 1: Designed for Bing

Configuration?



This looks easy – Plug into the network and go!

Less Fault Tolerant
Incast Problems

Centralized



Fault tolerant
Homogeneous

Multiple FPGAs?

Distributed

~2013 Microsoft Open Compute Server

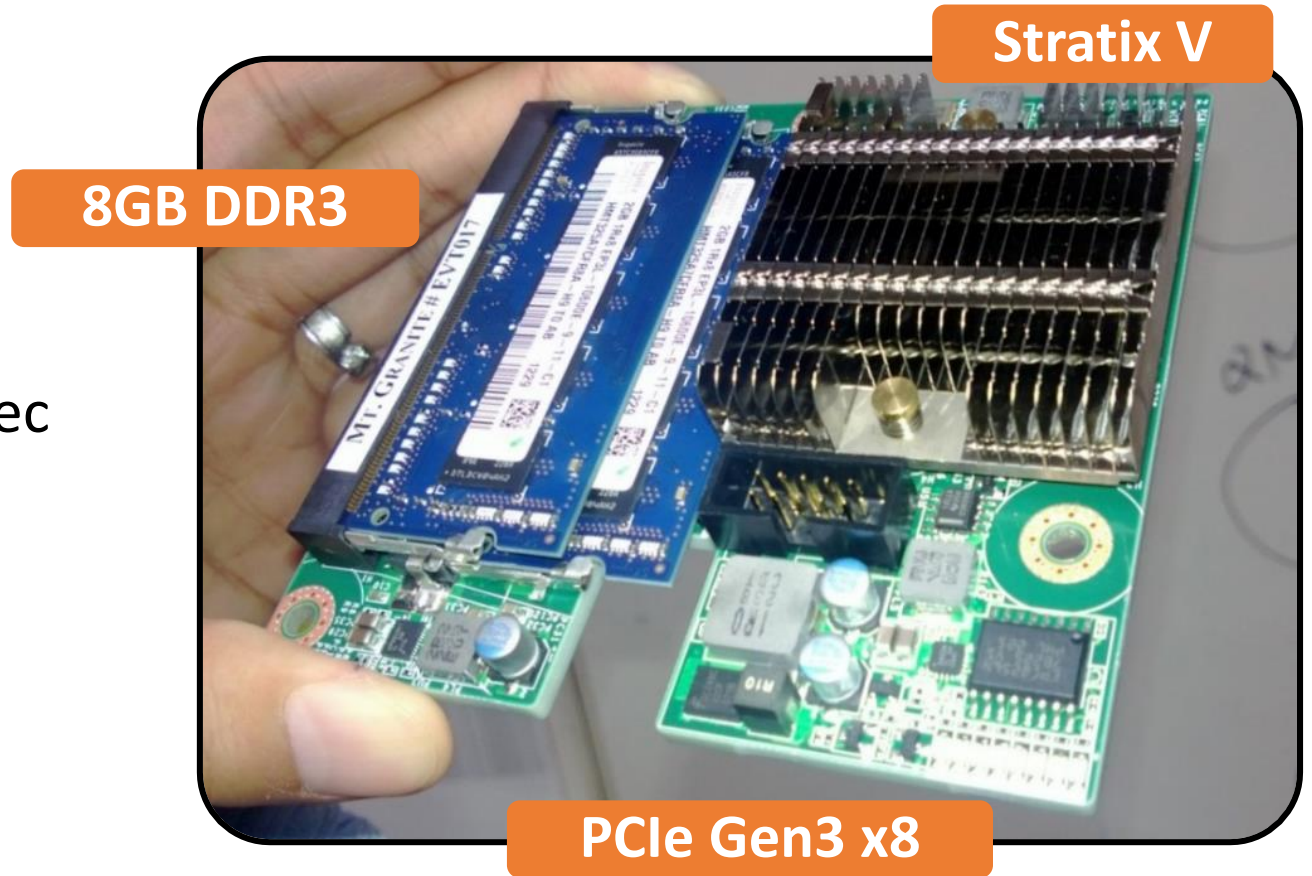


Two 8-core Xeon 2.1 GHz CPUs
64 GB DRAM
4 HDDs, 2 SSDs
10 Gb Ethernet

Air flow

Catapult V1 Accelerator Card

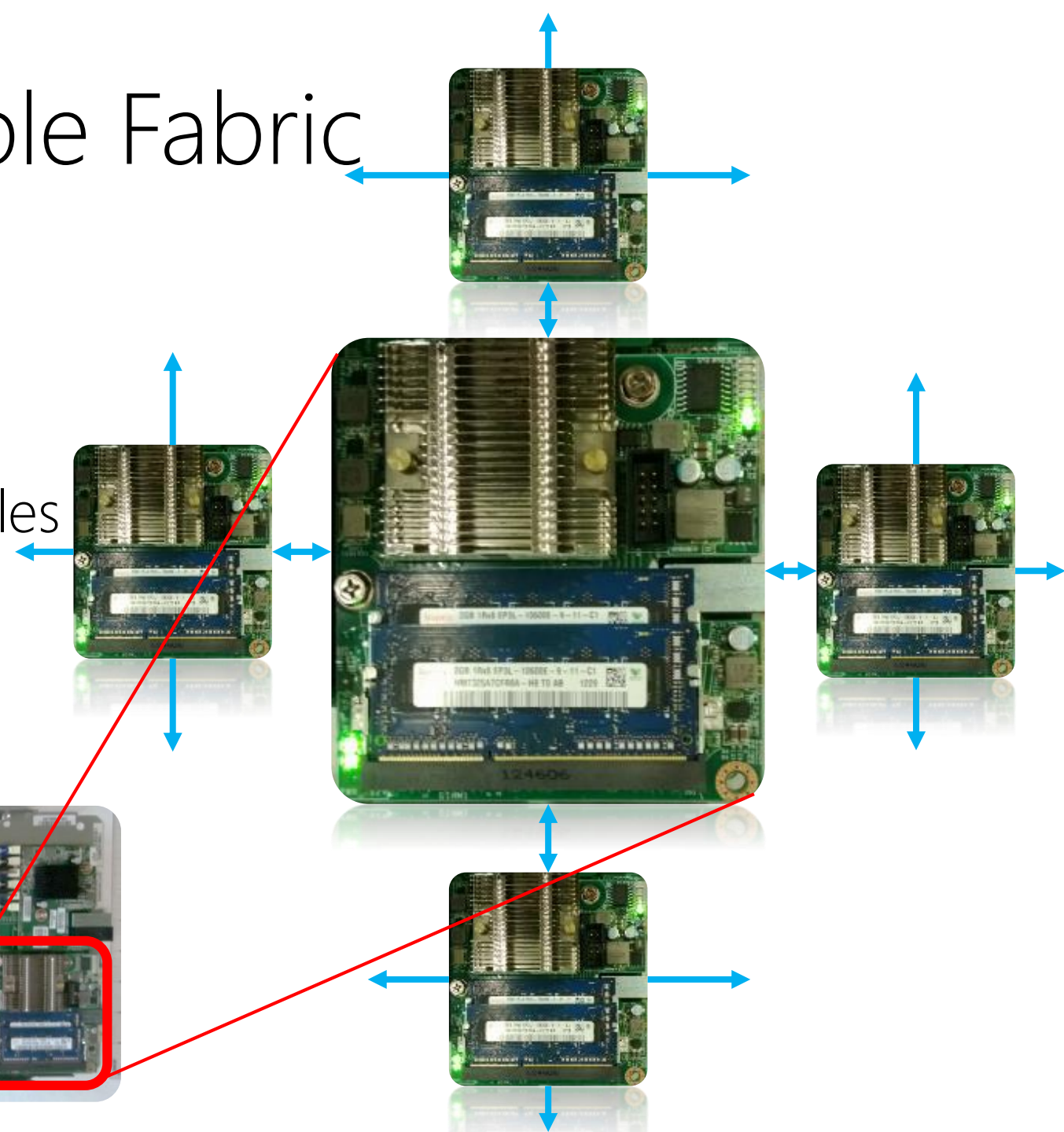
- Altera Stratix V D5 (2011 part)
 - Logic: ~5.5M ASIC gates+registers
 - 2014 20Kb memories
 - $2 \times 40b @ 200MHz = 4TB/sec$
- PCIe Gen 2 x8, 8GB DDR3
- *What if single FPGA too small?*




Scalable Reconfigurable Fabric

- 1 FPGA board per Server
- 48 Servers per ½ Rack
- 6x8 Torus Network among FPGAs
 - 12.5Gb over SAS SFF-8088 cables

Data Center Server (1U, ½ width)





Version 2:
Designed for all Microsoft

Economies of Scale

In data center, sameness == goodness

- Not possible everywhere, e.g., GPU SKUs

Have enough difference (divergence) from new components

- E.g., Intel processors change every 1-2 years

Divergence is very costly

- Stock spares for each type of machine
 - Many parts already end-of-lifed
- Keeping track of what needs to be purchases
- Reduced volumes == increased pricing

V1 designed for Bing, could others use V1?

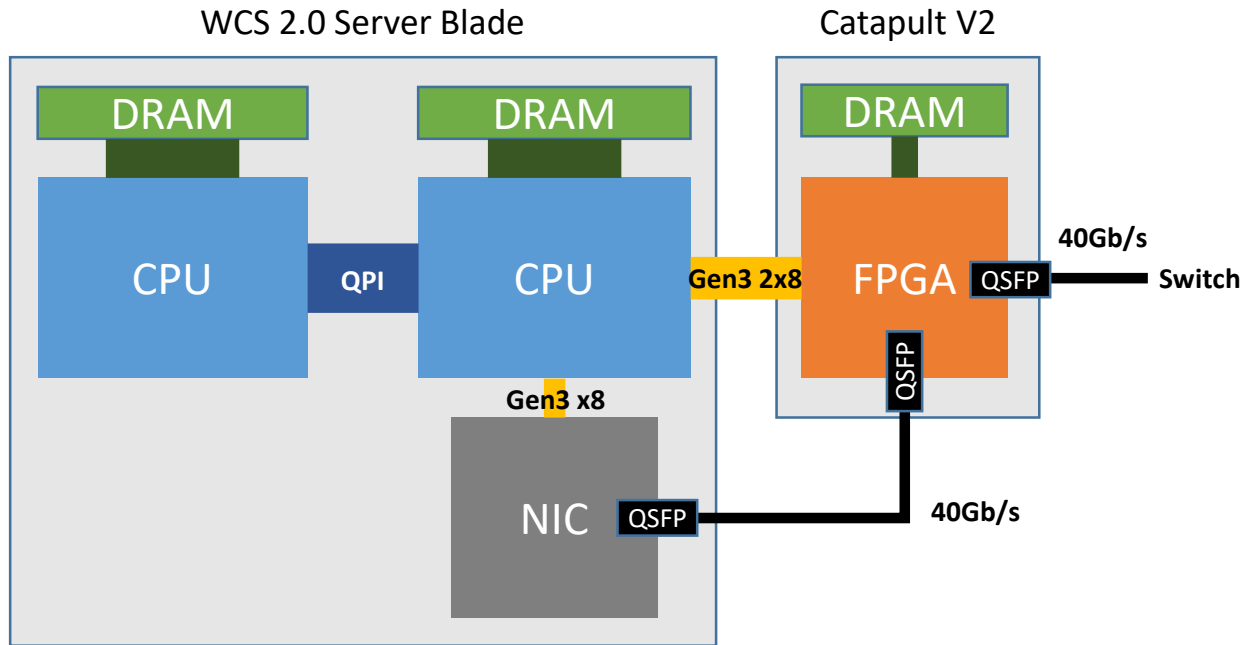
Can we design a system that is useful for all data center users?

FPGAs to Azure

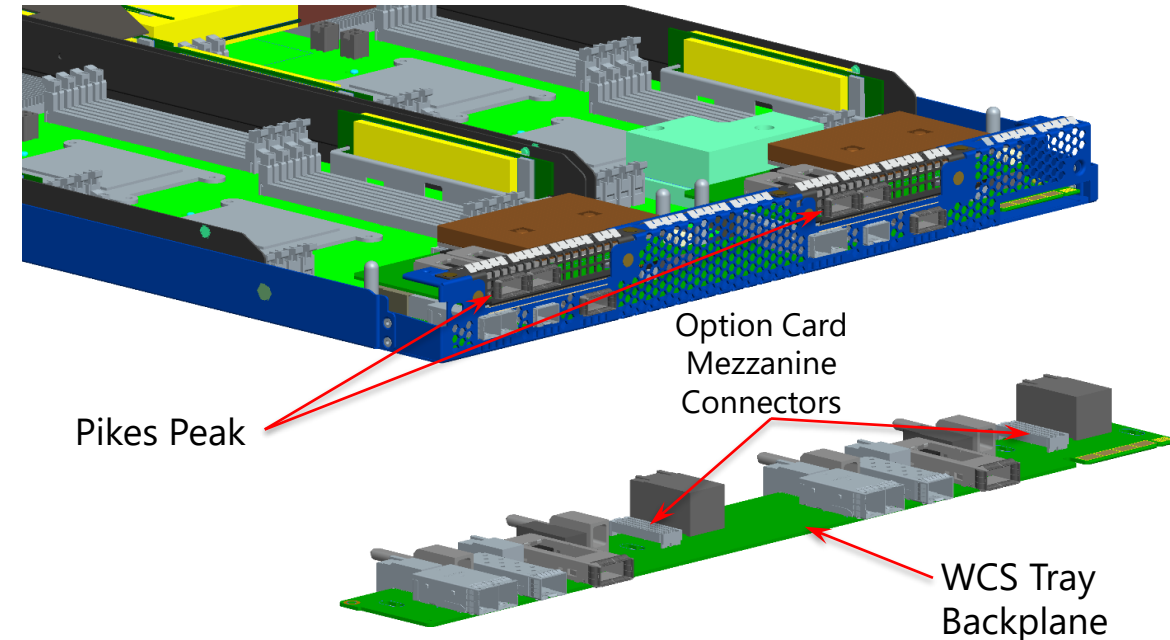
- At the time, Microsoft moving to “single SKU”
 - Azure needed to adopt FPGAs as well
- Azure networking saw a need for
 - network crypto
 - Software defined networking acceleration
- But, could you do it with our V1 architecture?
 - Was designed without network modifications because Azure not on board
- Developed new architecture to accommodate all known uses

Converged Bing/Azure Architecture

Catapult v2 Mezzanine card

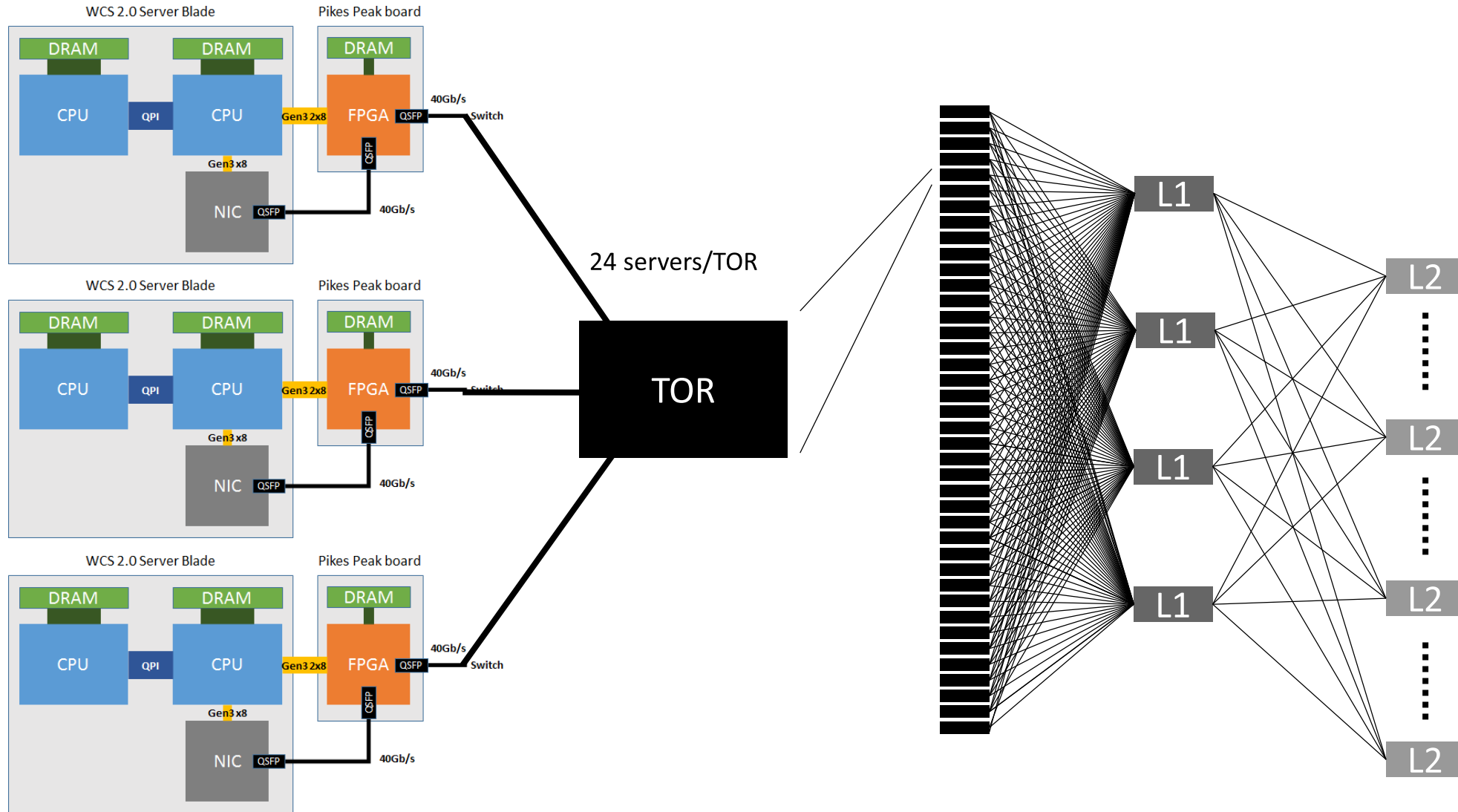


WCS Gen4.1 Blade with Mellanox NIC and Catapult FPGA



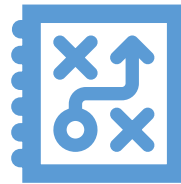
- Completely flexible architecture
 1. Can act as a local compute accelerator
 2. Can act as a network/storage accelerator
 3. Can act as a remote compute accelerator

Network Connectivity (IP)

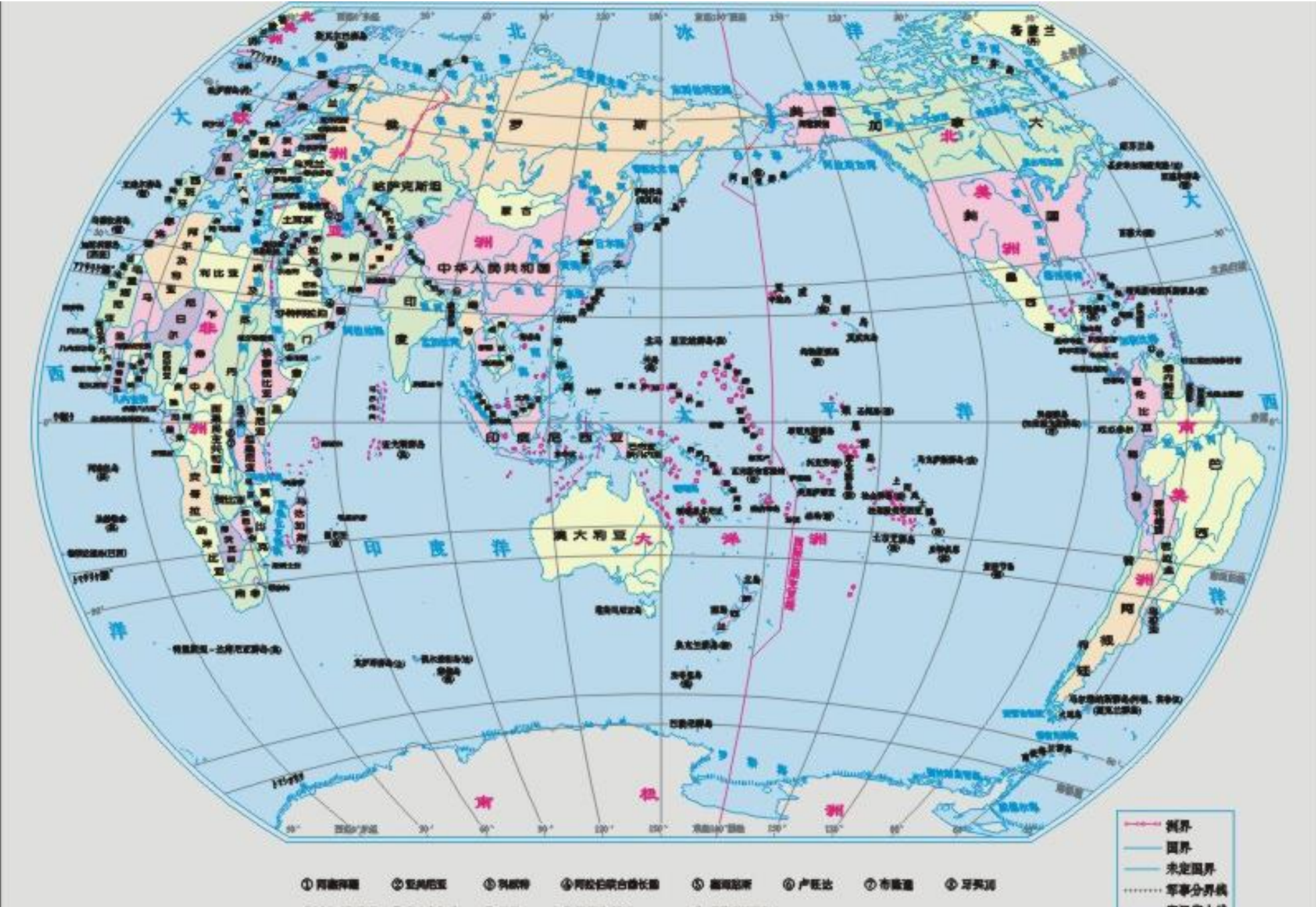


How Should We View Our Data Center Servers?

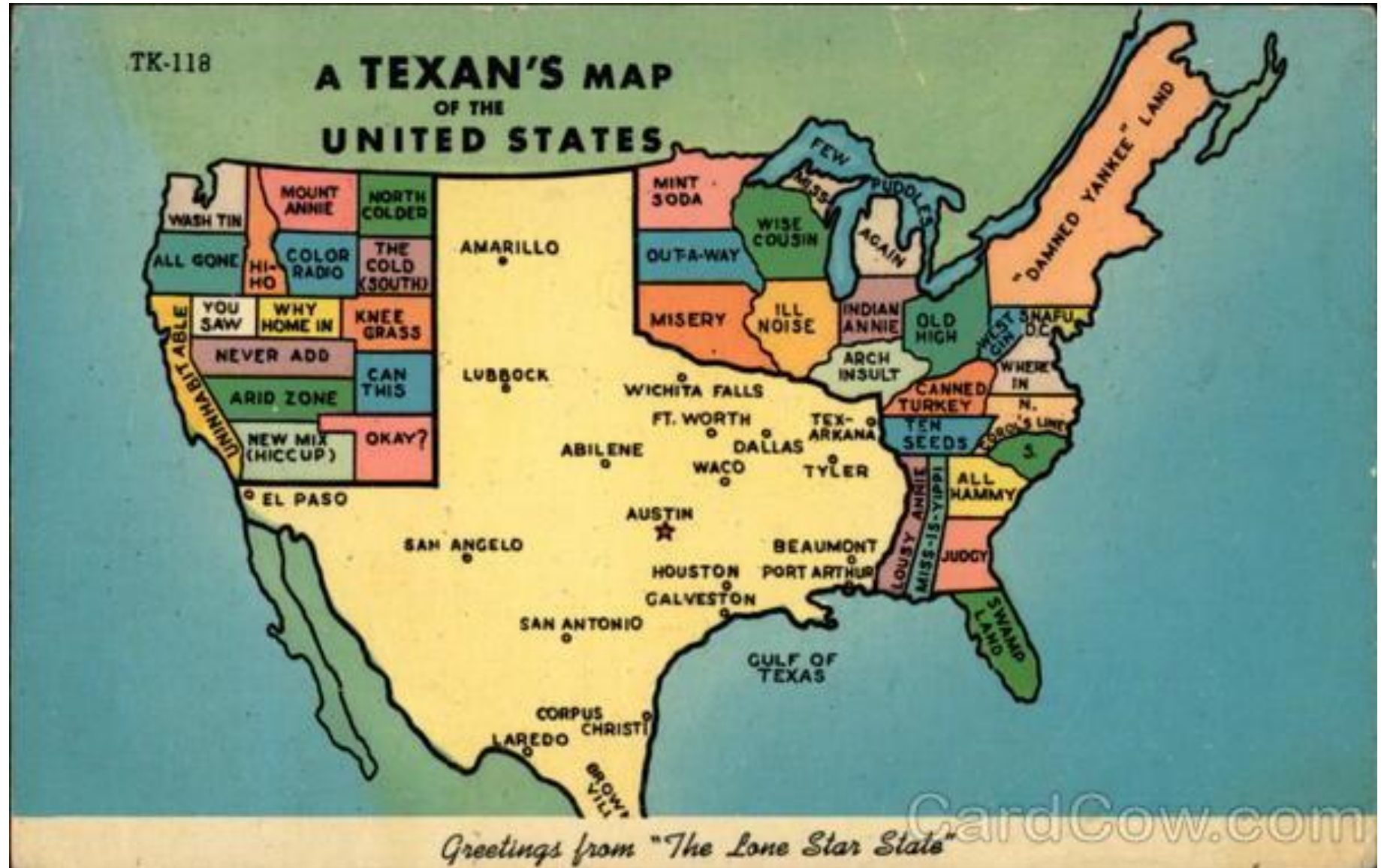
Depends on your point of view!



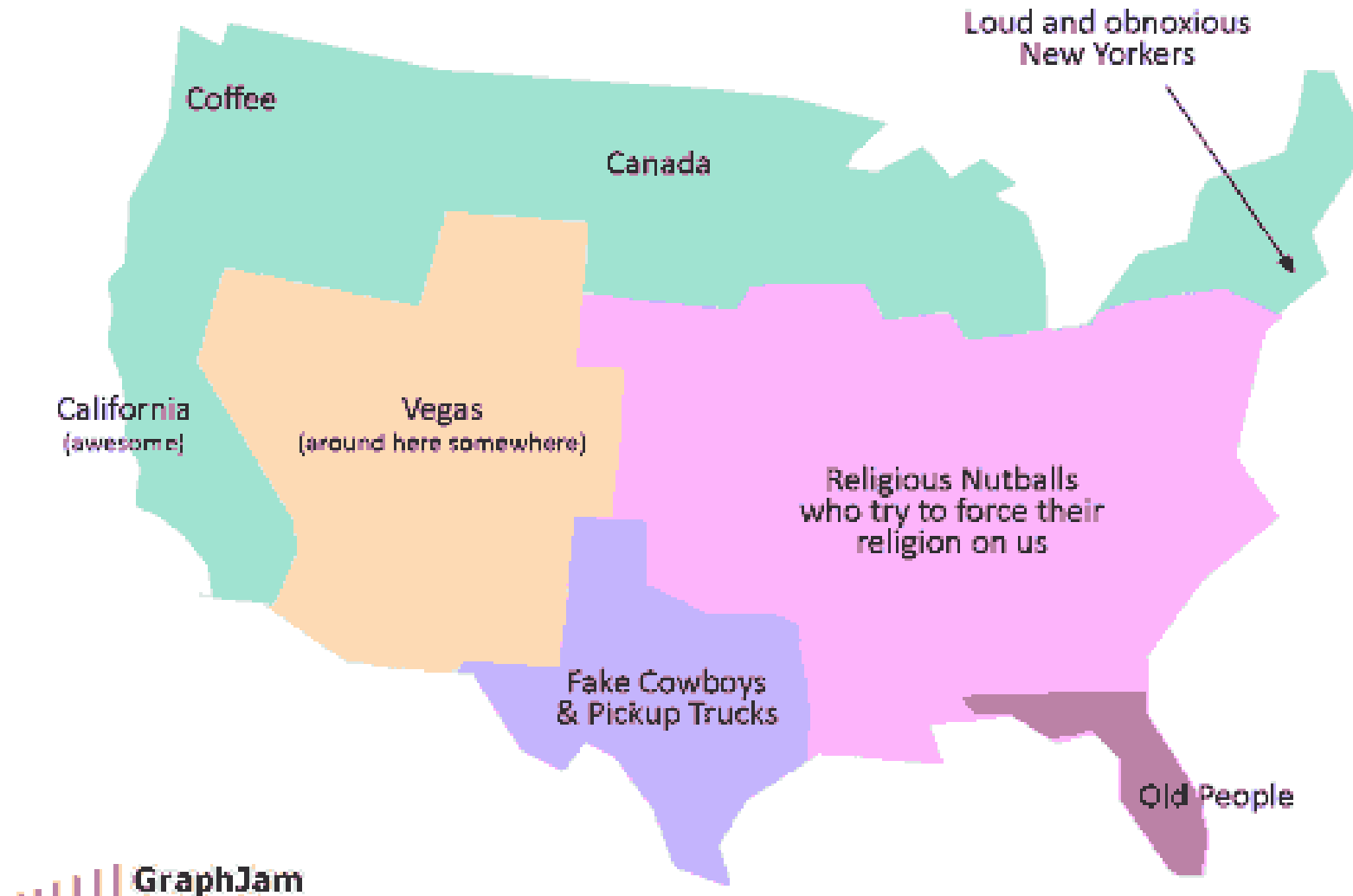




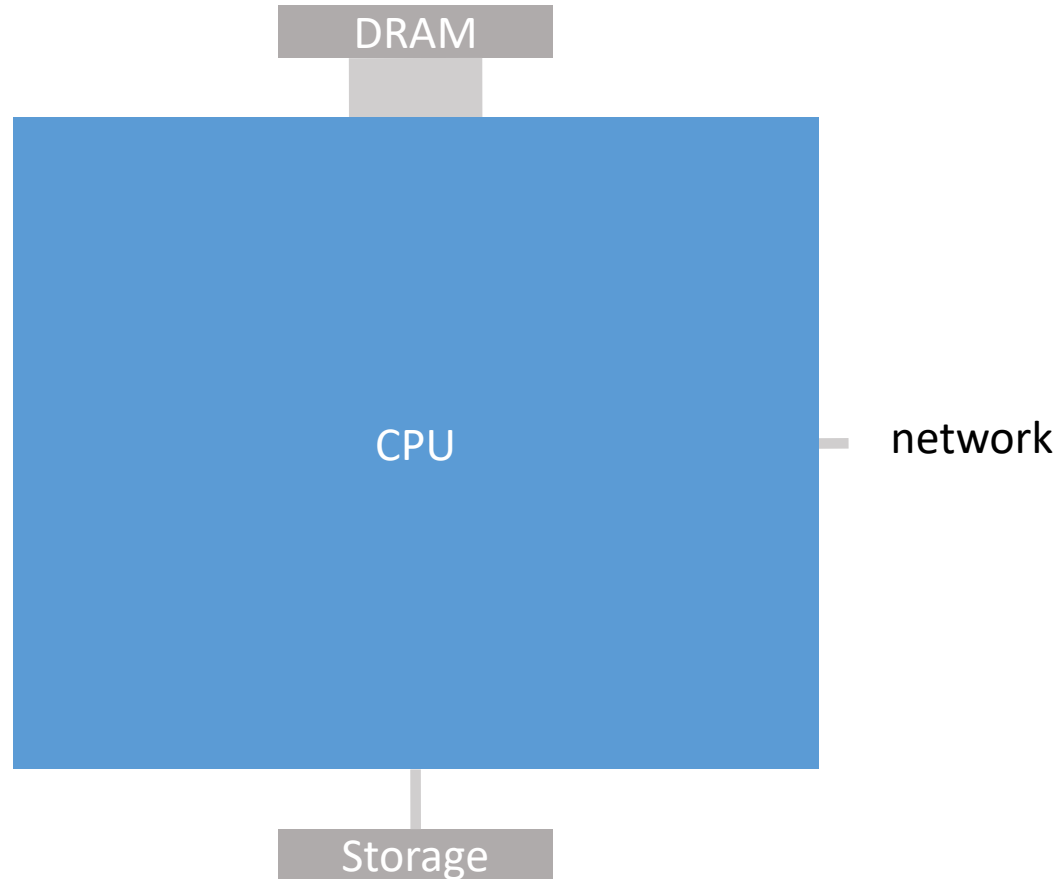
A Texan's View of the US



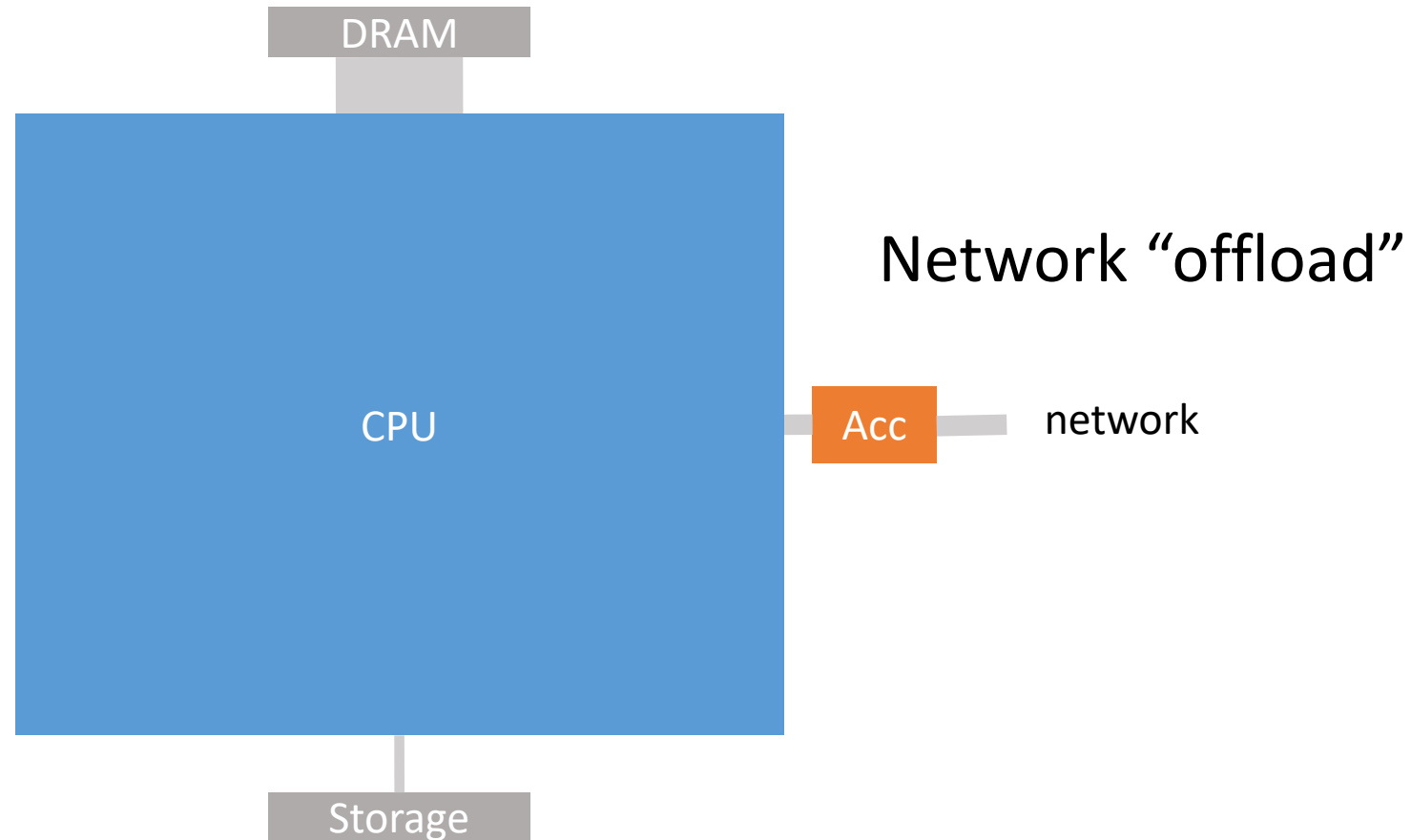
How Californians See America



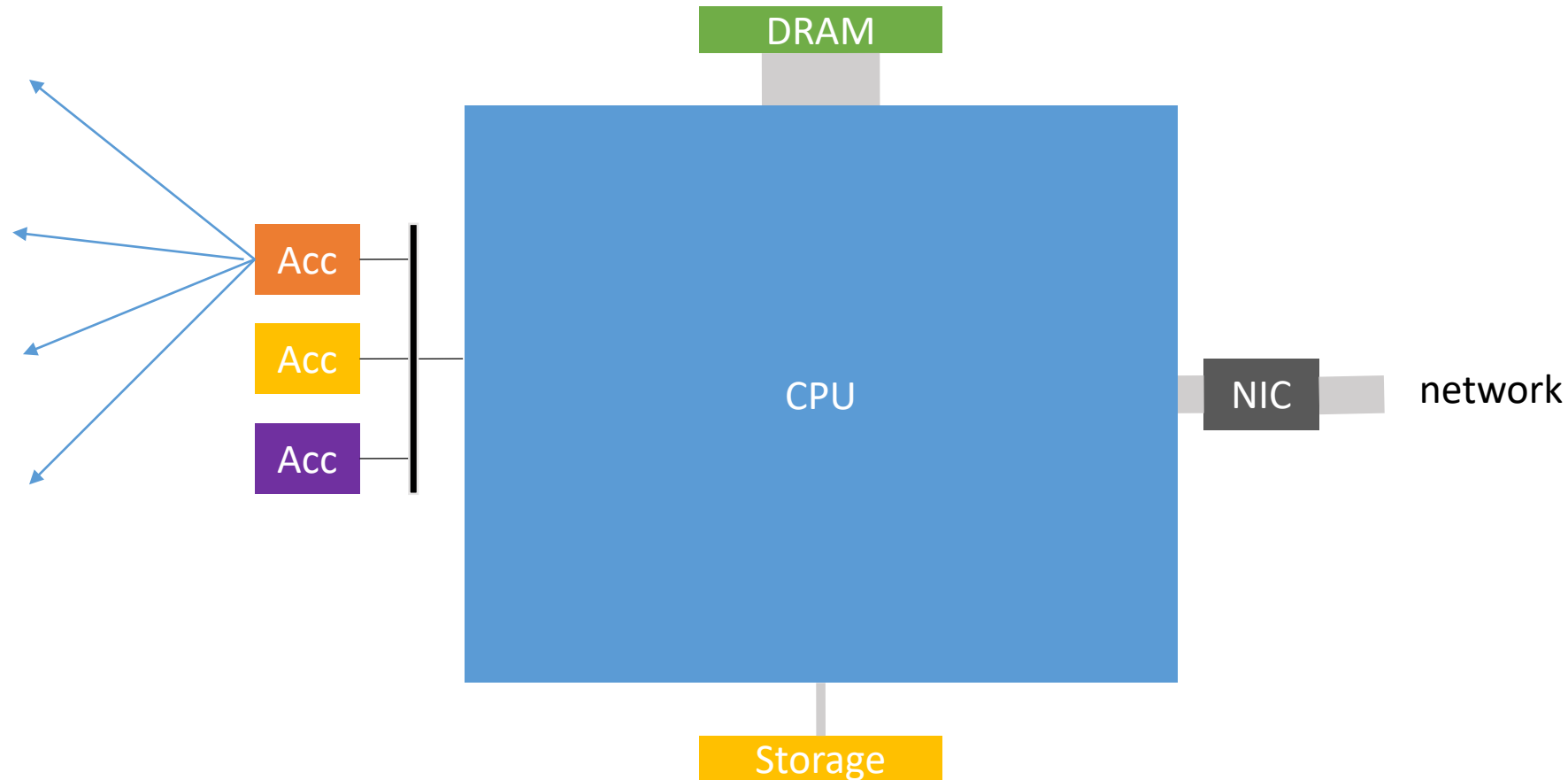
Classic View of Computer



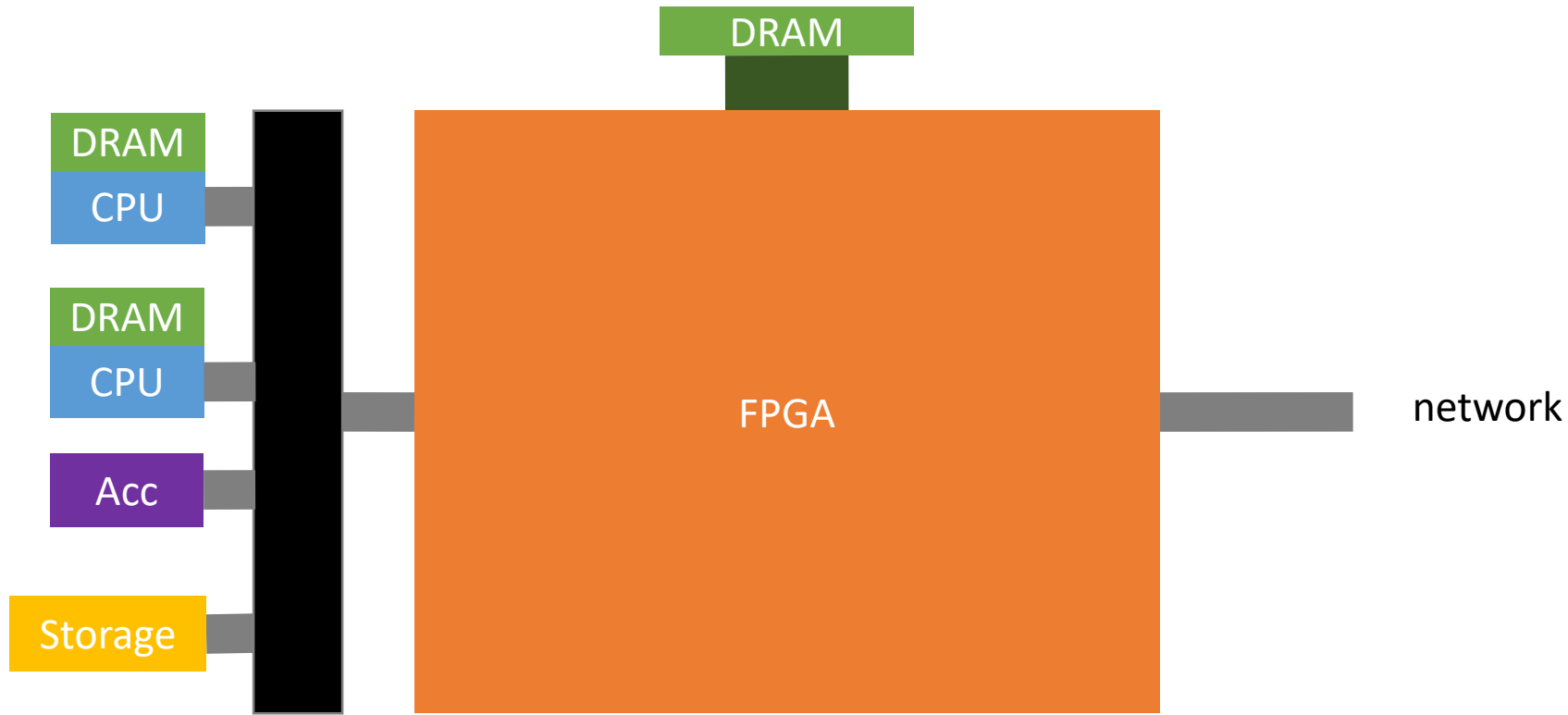
Networking View of Computer



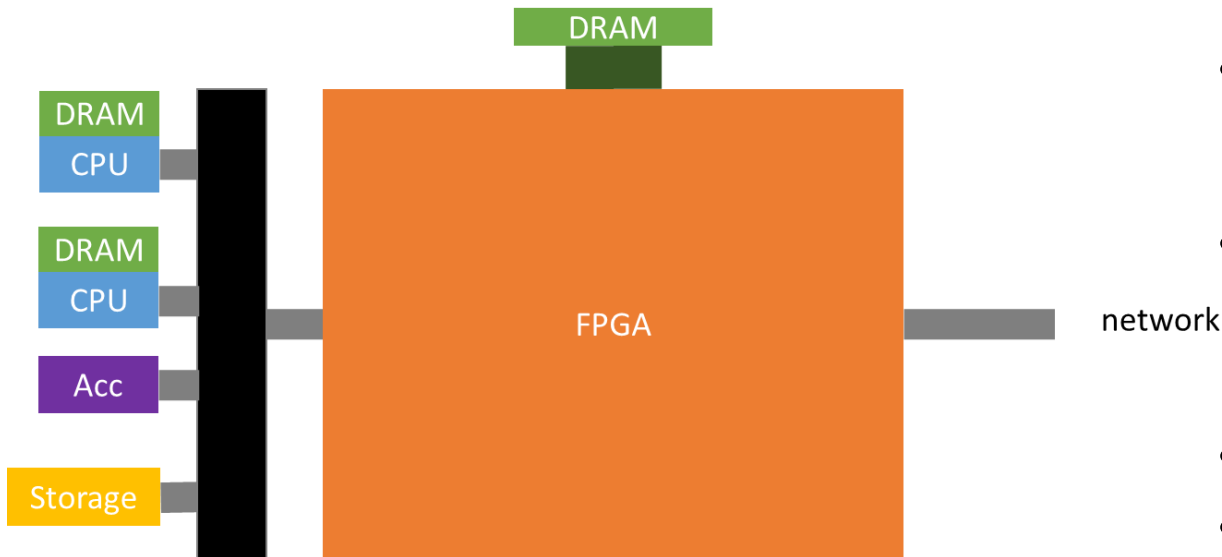
“Offload” Accelerator view of Server



Our View of a Data Center Computer



Benefits



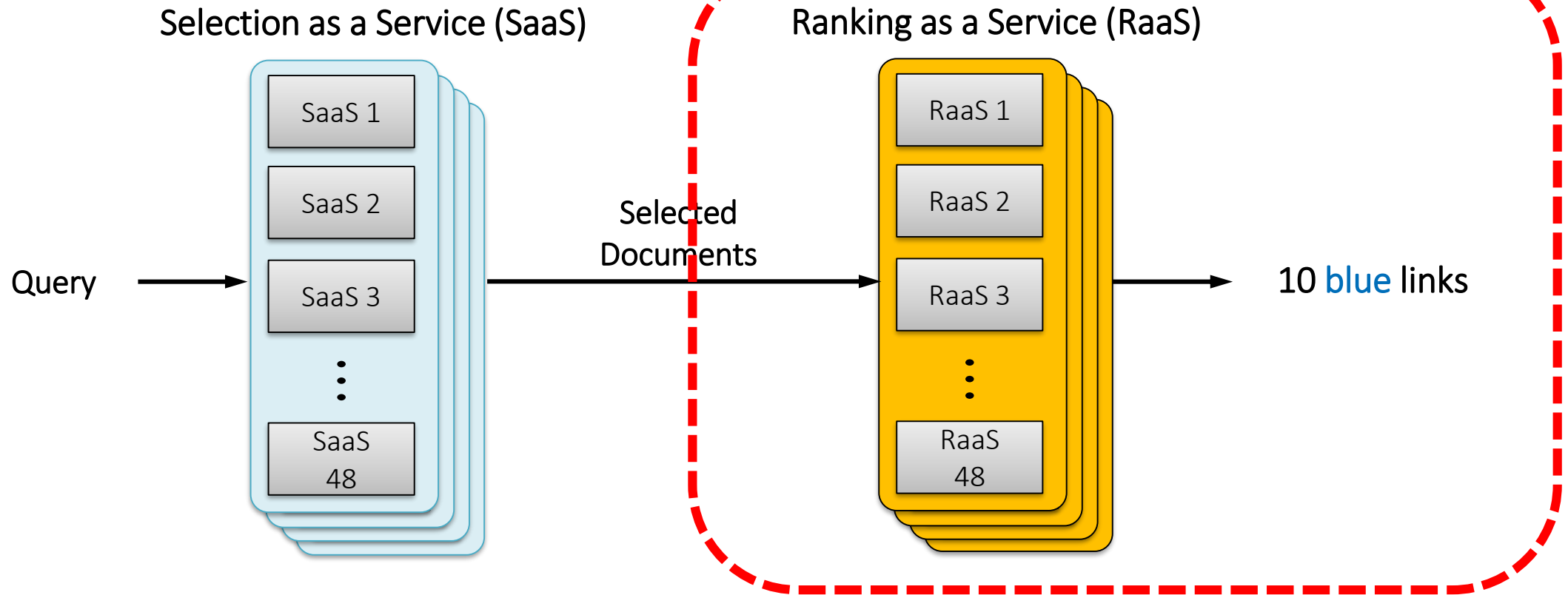
- Software receives packets slowly
 - Interrupt or polling
 - Parse packet, start right work
- FPGA processes every packet anyways
 - Packet arrival is an event that FPGA deals with
 - Identify FPGA work, pass CPU work to CPU
- Map common case work to FPGA
 - Processor never sees packet
 - Can read/modify system memory to keep app state consistent
- ***CPU is complexity offload engine for FPGA!***
- Many possibilities
 - Distributed machine learning
 - Software defined networking
 - Memcached get

Case 1

Use as a local accelerator

Bing Ranking as a Service

Bing Document Ranking Flow



Selection-as-a-Service (SaaS)

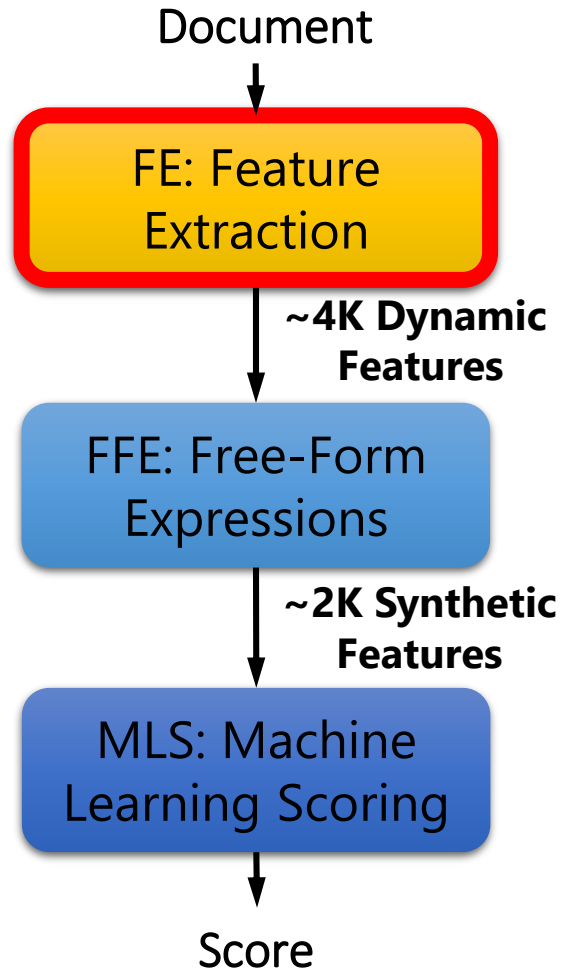
- Find all docs that contain query terms,
- Filter and select candidate documents for ranking

Ranking-as-a-Service (RaaS)

- Compute scores for how relevant each selected document is for the search query
- Sort the scores and return the results

FE: Feature Extraction

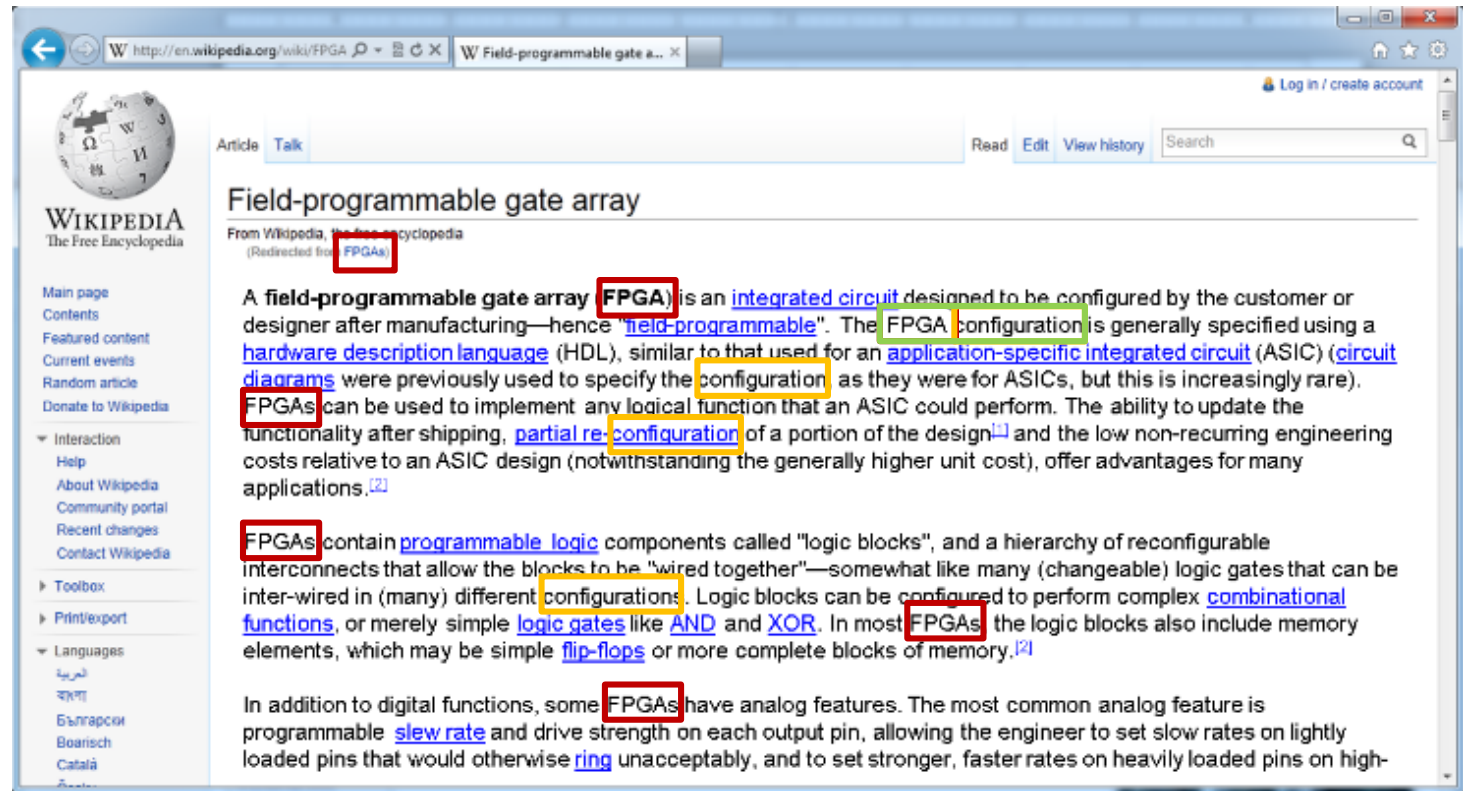
Query: "FPGA Configuration"



NumberOfOccurrences_0 = 7

NumberOfOccurrences_1 = 4

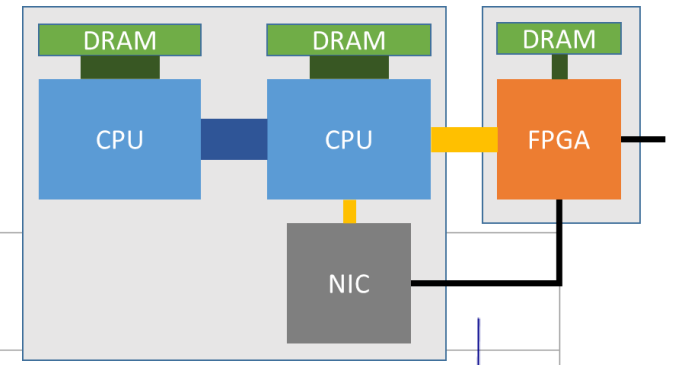
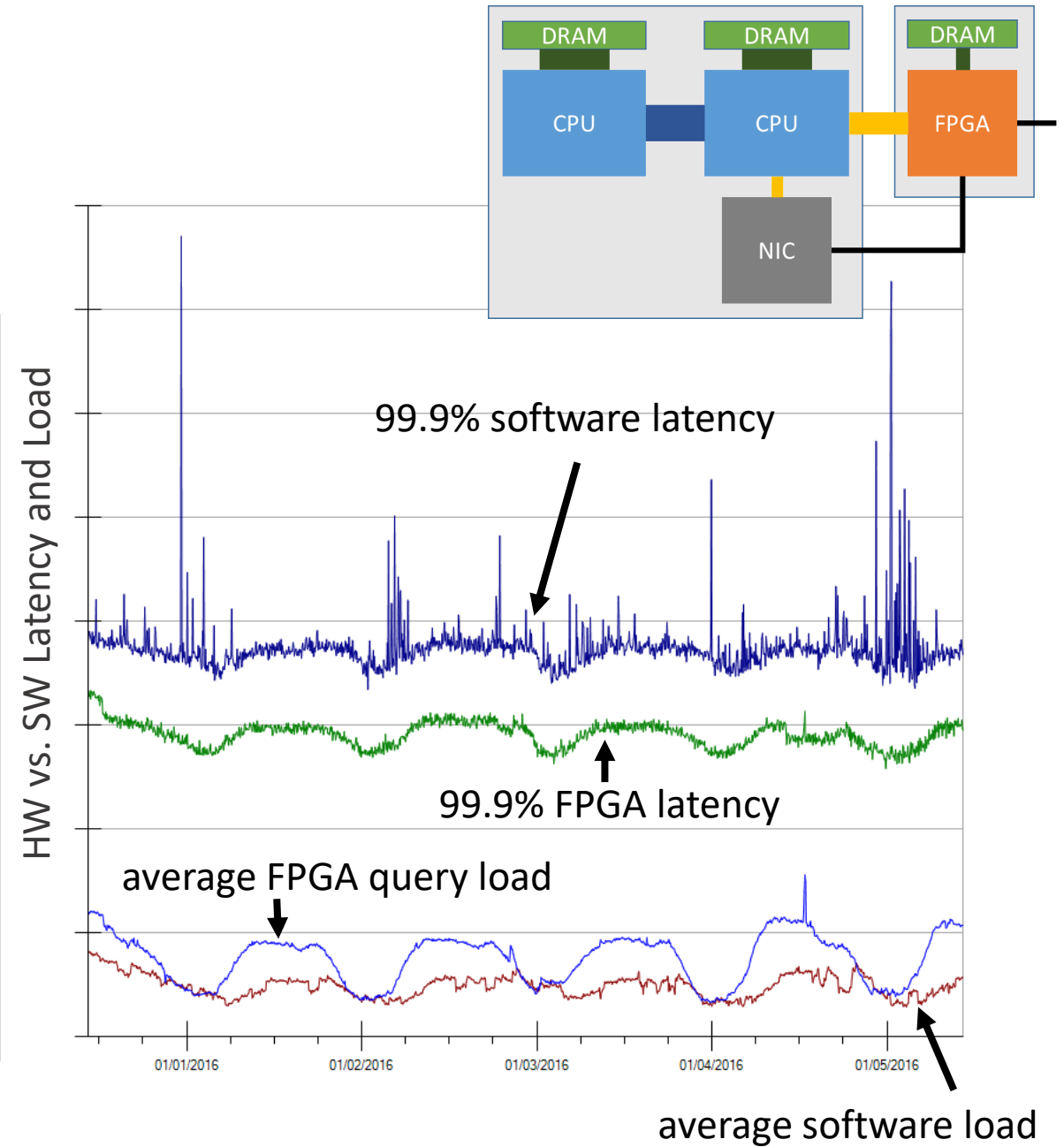
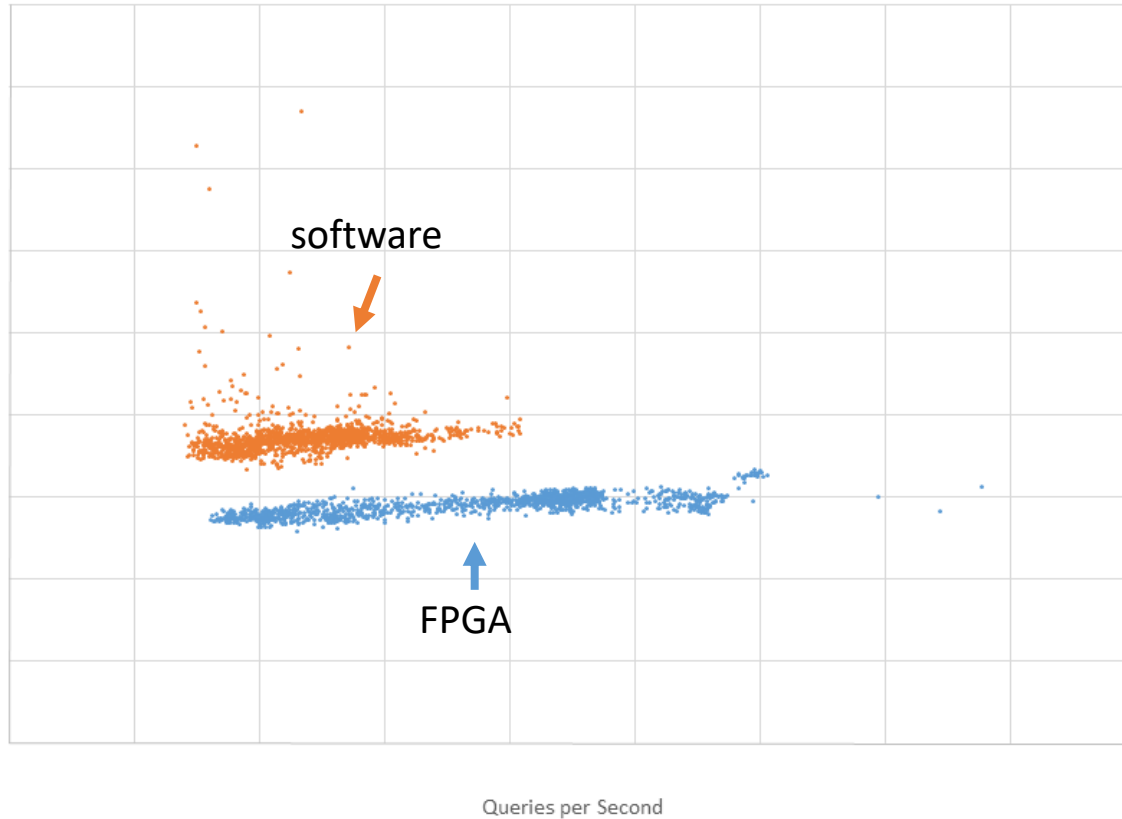
NumberOfTuples_0_1 = 1



Bing Production Results

99.9% Query Latency versus Queries/sec

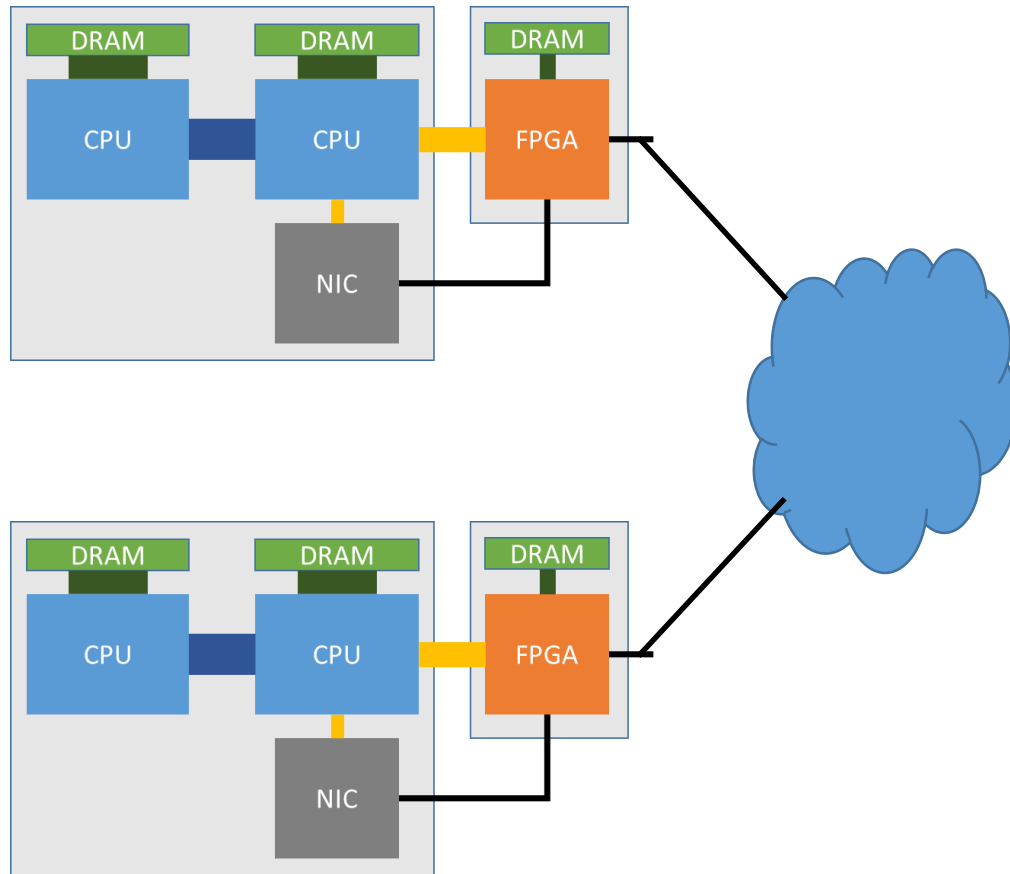
Query Latency 99.9



Case 2

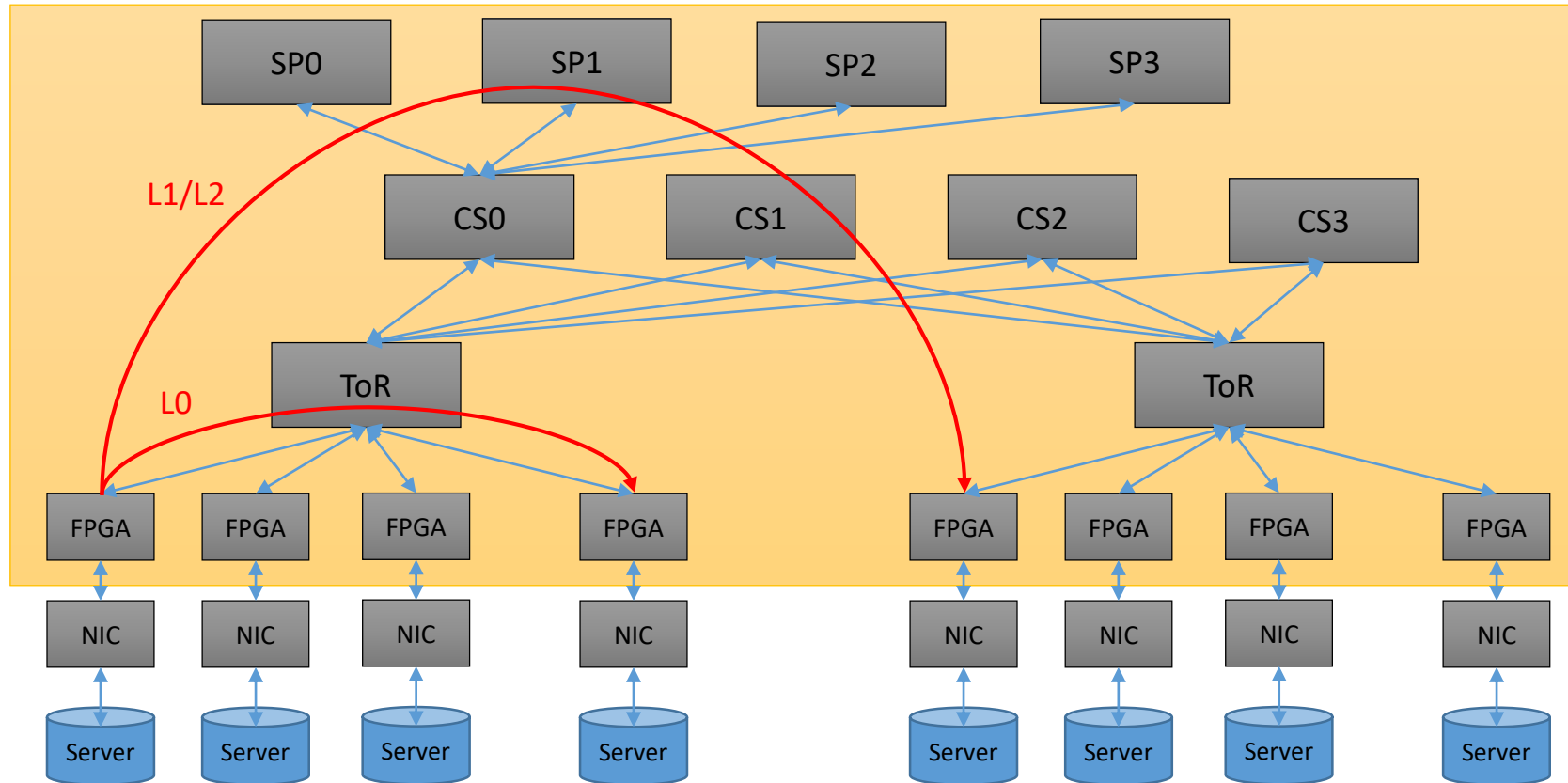
Use as a remote accelerator

Feature Extraction FPGA faster than needed



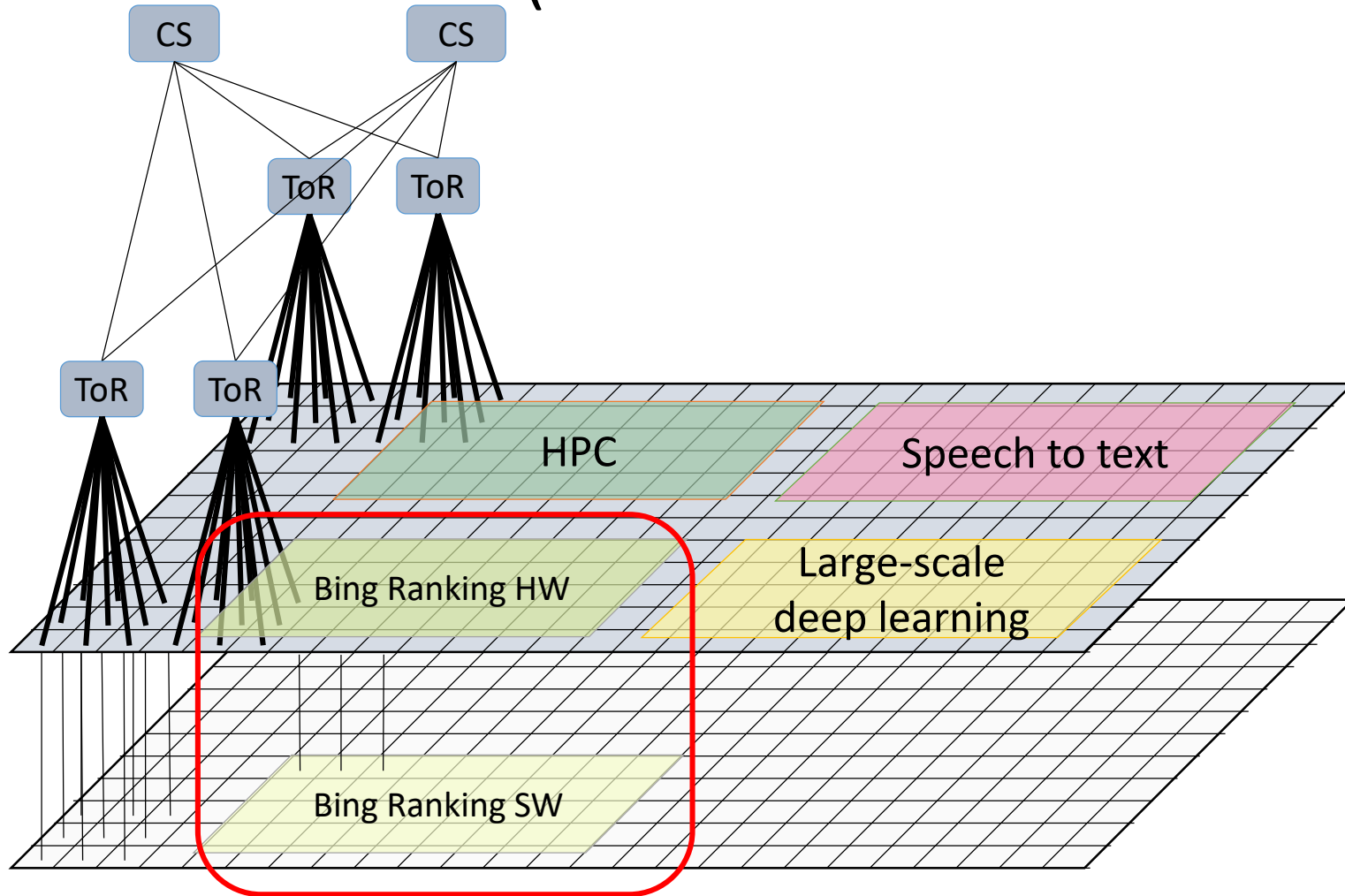
- Single feature extraction FPGA much faster than single server
- Wasted capacity and/or wasted FPGA resources
- Two choices
 - Somehow reduce performance and save FPGA resources
 - Allow multiple servers to use single FPGA?
- Use network to transfer requests and return responses

Inter-FPGA communication



- FPGAs can encapsulate their own UDP packets
- Low-latency inter-FPGA communication (LTL)
- Can provide strong network primitives
- But this topology opens up other opportunities

Hardware Acceleration as a Service Across Data Center (or even across Internet)

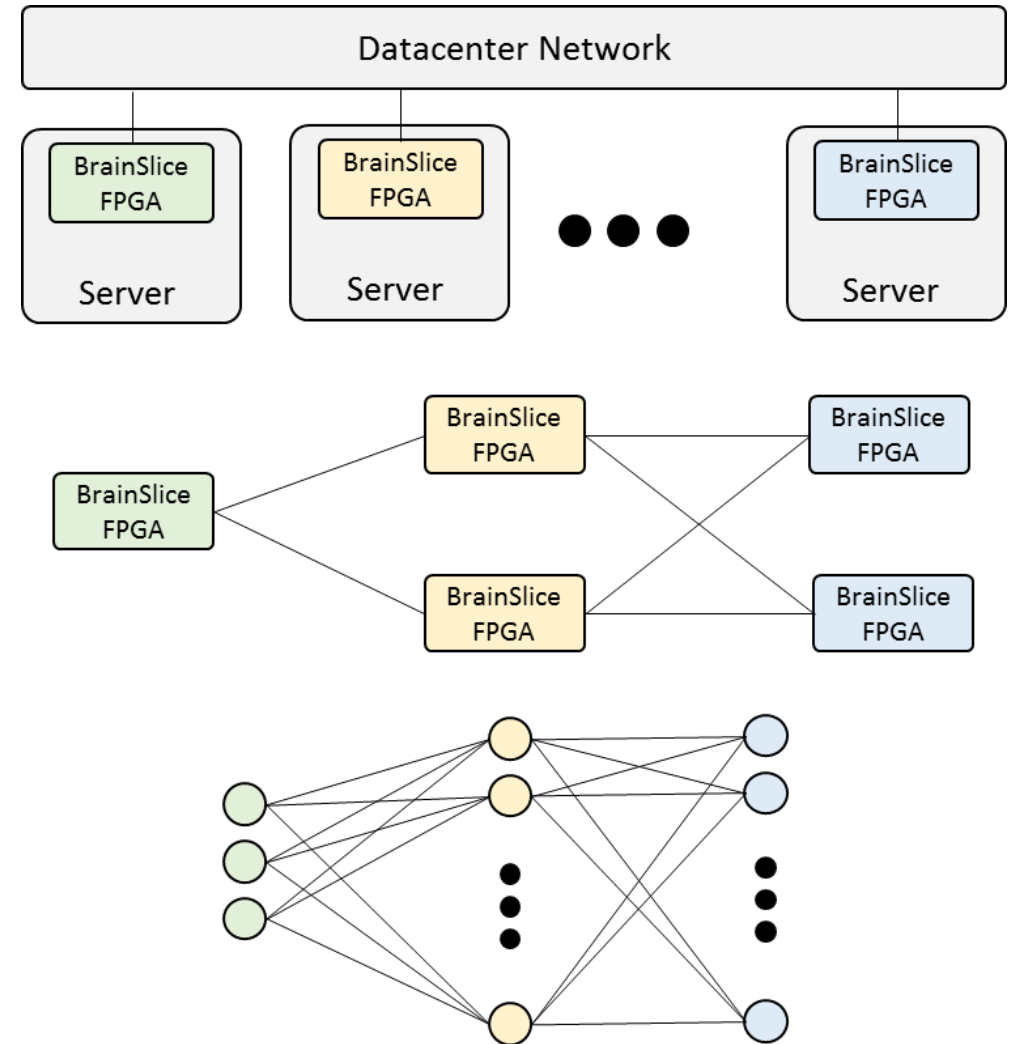


- $O(100K)$ servers
- Services may co-design with their local FPGAs or allocate a HaaS service remotely.
- FPGA is independent of server

Current Bing Acceleration: DNNs

BrainWave: Scaling FPGAs To Ultra-Large DNN Models

- Use FPGAs to implement Deep Neural Network evaluation (inference)
- Map model weights to internal FPGA memories
 - Huge amounts of bandwidth
- Since FPGA memories are limited, distribute models across as many FPGAs as needed
- Use HaaS to manage multi-FPGA execution, LTL to communicate
- Designed for batch size of 1
 - GPUs, Google TPU designed for larger batch sizes, increases queuing delay or decreases efficiency

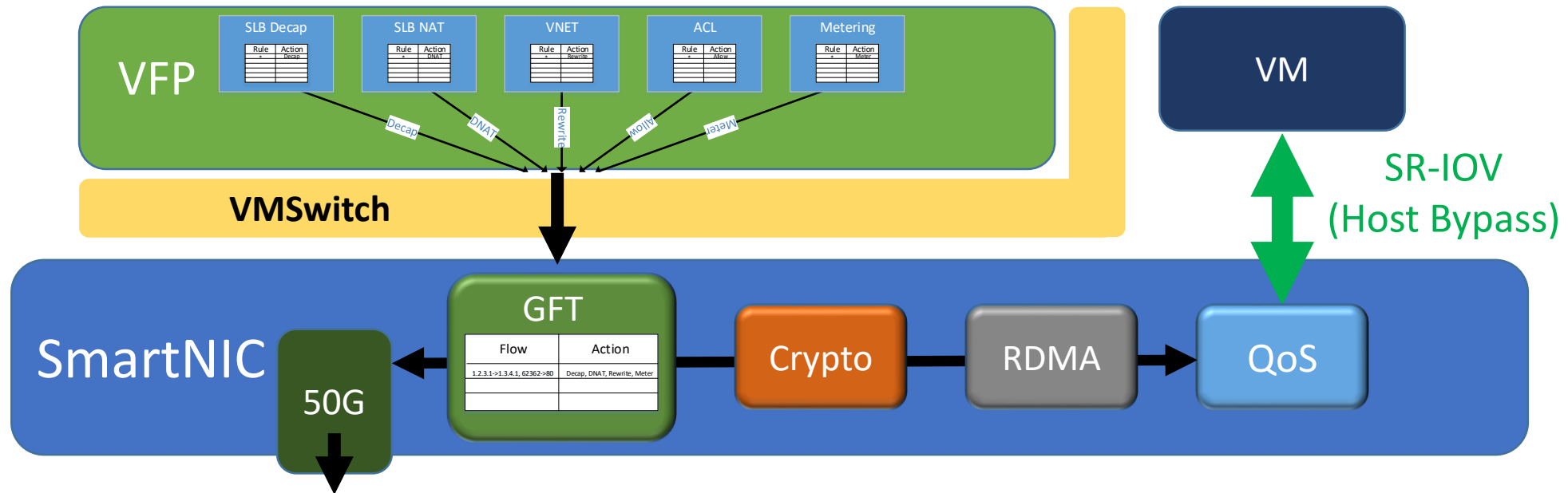


Case 3

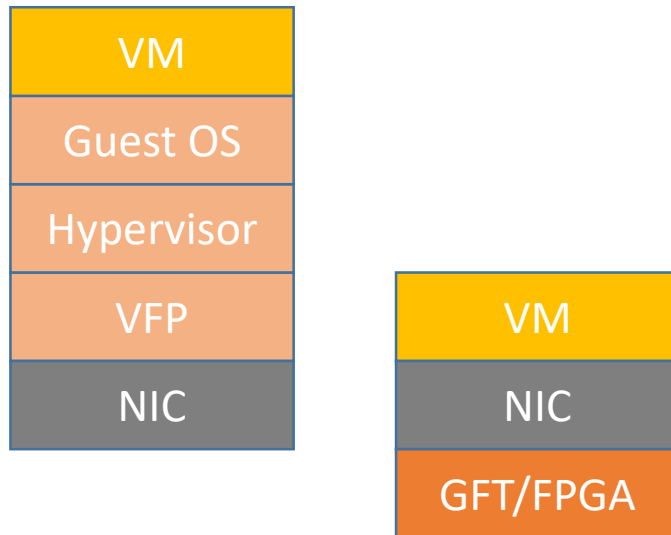
Use as an infrastructure accelerator

FPGA SmartNIC for Cloud Networking

- Azure runs Software Defined Networking on the hosts
 - Software Load Balancer, Virtual Networks – new features each month
- Before, we relied on ASIC to scale and to be COGS-competitive at 40G+
 - 12 to 18 month ASIC cycle + time to roll out new HW is too slow to keep up with SDN
- SmartNIC gives us the agility of SDN with the speed and COGS of HW
 - Base SmartNIC provide common functions like crypto, GFT, QoS, RDMA on all hosts



Azure Accelerated Networking



- SR-IOV turned on
 - VM accesses NIC hardware directly, sends messages with no OS/hypervisor call
- FPGA determines flow of each packet, rewrites header to make data center compatible
- Reduces latency to roughly bare metal
- Azure now has the ***fastest*** public cloud network
 - 25Gb/s at 25us latency
- Fast crypto developed

Catapult Academic Program



- Jointly funded by Intel and Microsoft
 - Some system administration/tools servers funded by NSF under FAbRIC
- In UTexas supercomputer facility under FAbRIC project
- Provide
 - PCIe device driver, shell (initially compiled/encrypted, discussing source access under NDA with lawyers)
 - “Hello, world!” programs
 - Individual V1 boards sent to you
 - Remote access to 6*48 servers each with V1 board
 - Accessible with one page proposal catapult@microsoft.com
- See <https://aka.ms/catapult-academic> for details



Conclusion: Hardware Specialization on Homogeneous Machines

- FPGAs being deployed for all new Azure and Bing machines
 - Many other properties as well
- Ability to reprogram a datacenter's hardware
 - Converts homogenous machines into specialized SKUs dynamically
 - Same hardware supports DNNs, Bing Features, Azure Networking
- Hyperscale performance with low latency communication
 - Exa-ops of performance with a $O(10\mu s)$ diameter