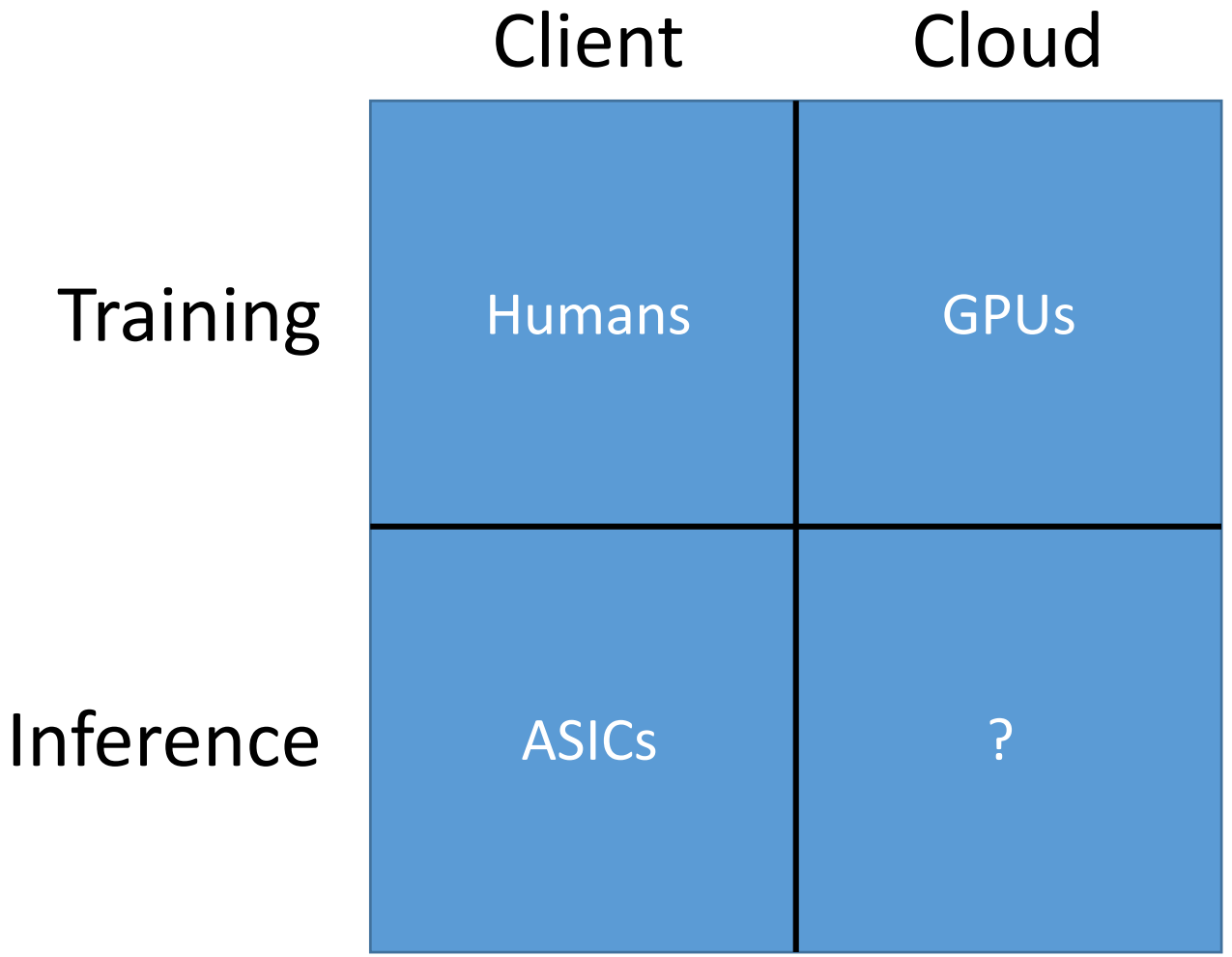
(Re)Configurable Clouds and the Dawn of a New Era

Doug Burger @ Microsoft Research NExT

FPL Keynote

August 30, 2016


Client
Cloud
Training
Inference
Humans
GPUs
ASICs
?

5.8+ billion
worldwide queries each month
bing

250+ million
active users
OneDrive

400+ million
active accounts
Outlook.com

2.4+ million
emails per day
Microsoft Exchange Hosted Services

8.6+ trillion
objects in Microsoft Azure storage
Microsoft Azure

48+ million
users in 41 markets
XBOX LIVE

50+ million
active users
Office Web Apps

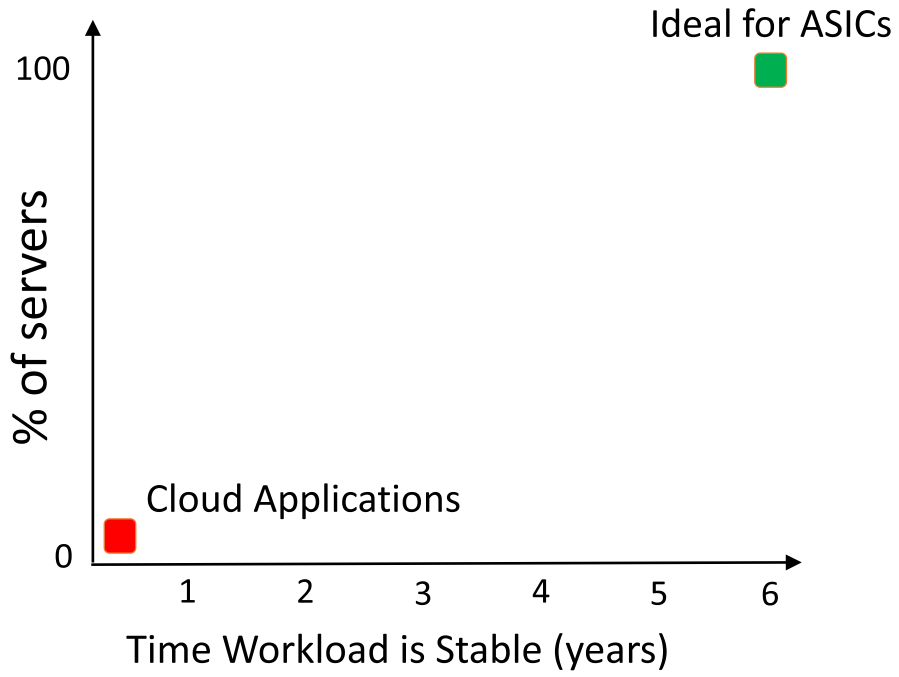
1 in 4
enterprise customers
Office 365

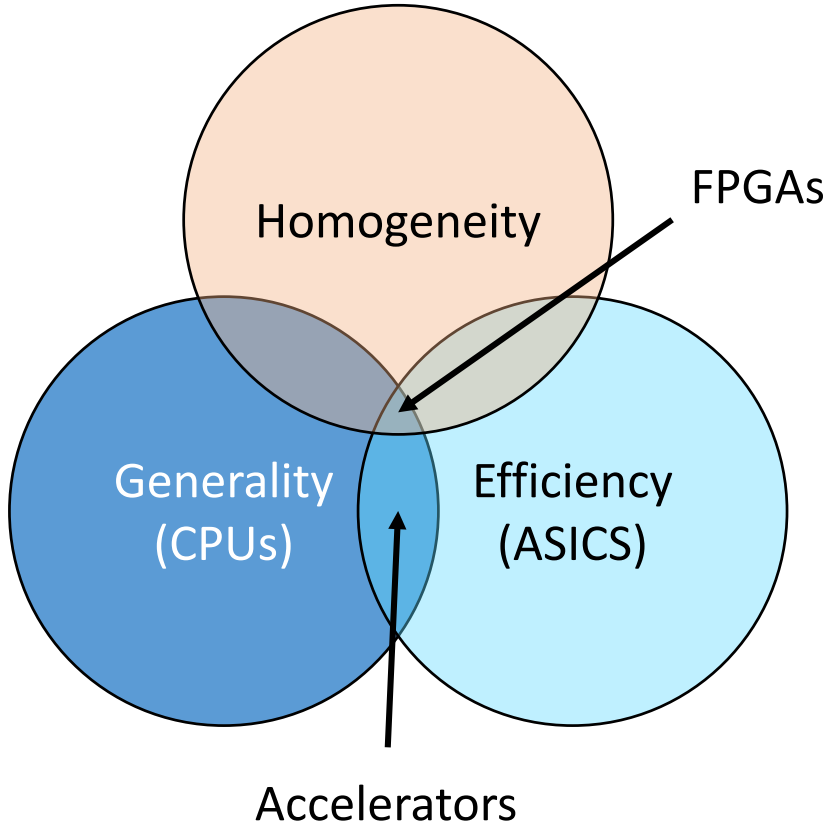
50+ billion
minutes of connections handled each month
skype

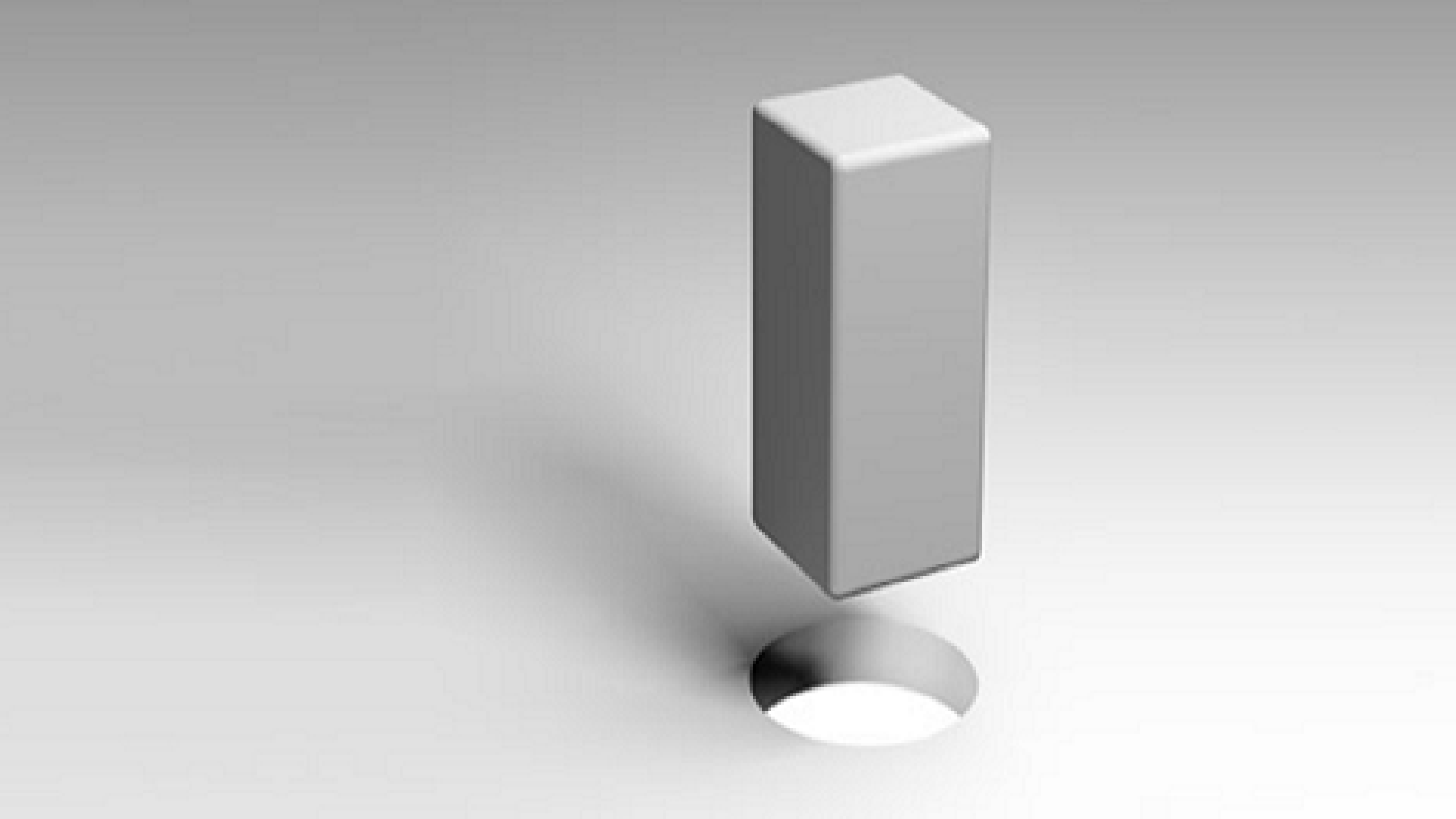
200+ Cloud Services: Diversity

1+ billion customers · 20+ million businesses · 90+ markets worldwide

What Drives a Post-CPU “Enhanced” Cloud?

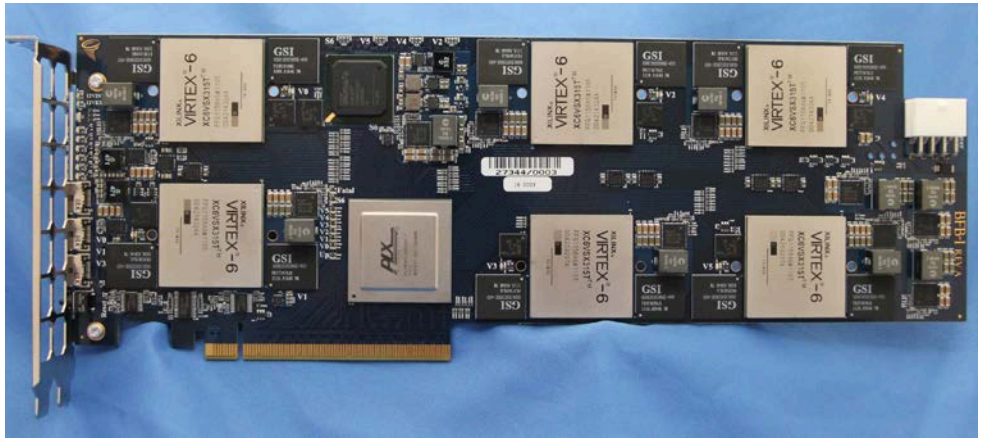


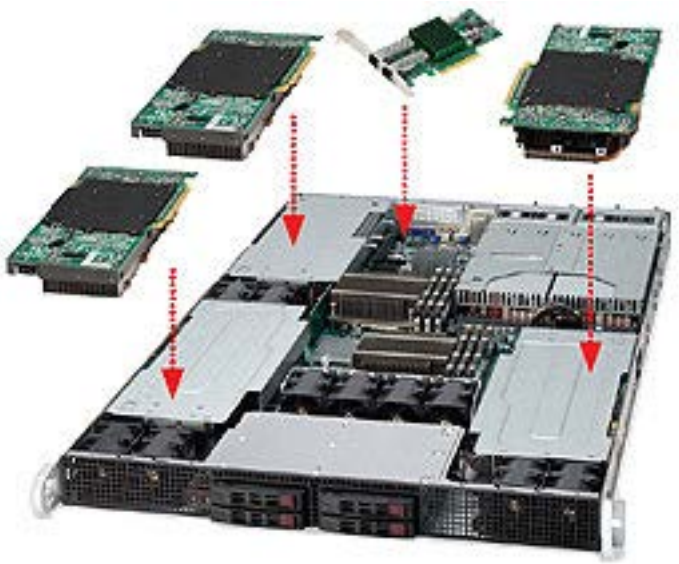




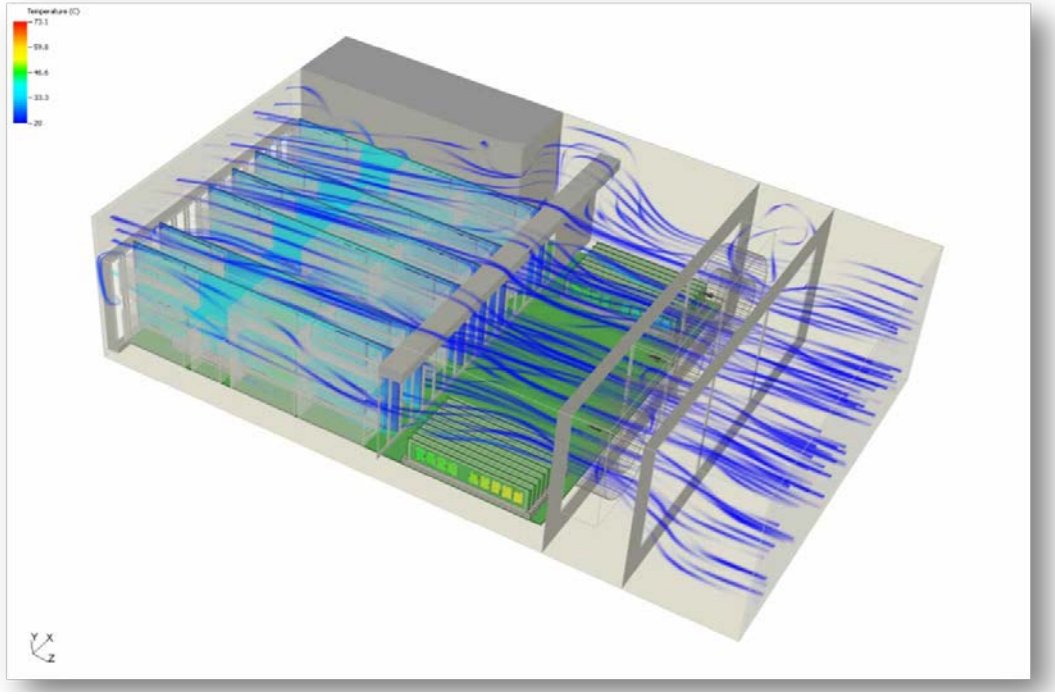
Catapult V0: BFB (2011)

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

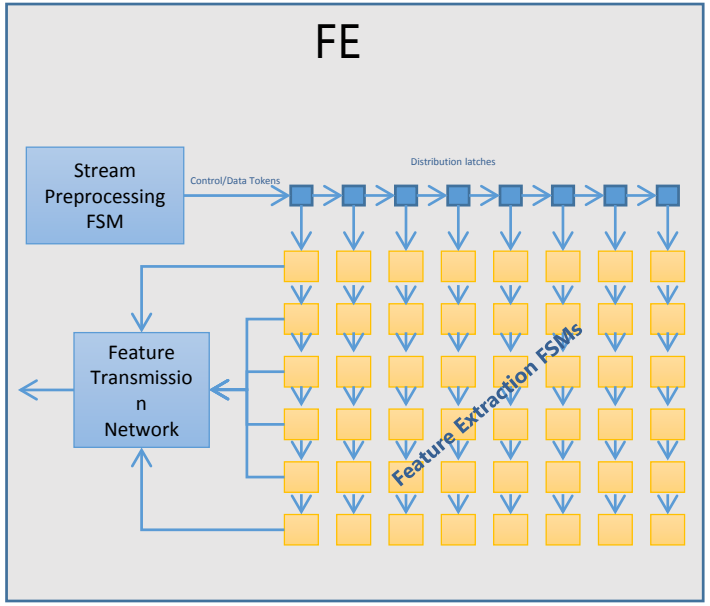


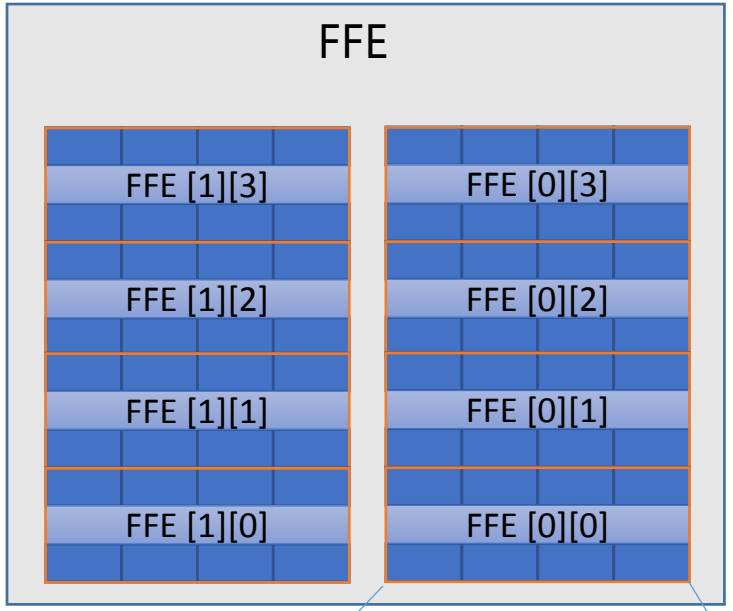


- 1U rack-mounted
- 2 x 10Ge ports
- 3 x16 PCIe slots
- 12 Intel Westmere cores (2 sockets)



Bing Ranking Implementation Details





DTS

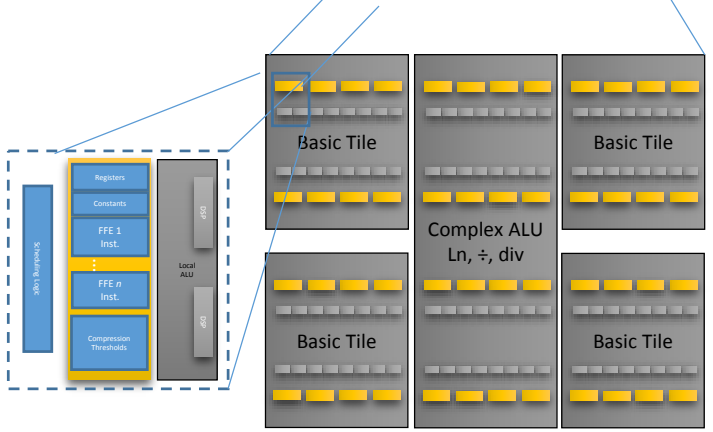
DTT [3][11]	DTT [2][11]	DTT [1][11]	DTT [0][11]
DTT [3][10]	DTT [2][10]	DTT [1][10]	DTT [0][10]
DTT [3][9]	DTT [2][9]	DTT [1][9]	DTT [0][9]
DTT [3][8]	DTT [2][8]	DTT [1][8]	DTT [0][8]
DTT [3][7]	DTT [2][7]	DTT [1][7]	DTT [0][7]
DTT [3][6]	DTT [2][6]	DTT [1][6]	DTT [0][6]
DTT [3][5]	DTT [2][5]	DTT [1][5]	DTT [0][5]
DTT [3][4]	DTT [2][4]	DTT [1][4]	DTT [0][4]
DTT [3][3]	DTT [2][3]	DTT [1][3]	DTT [0][3]
DTT [3][2]	DTT [2][2]	DTT [1][2]	DTT [0][2]
DTT [3][1]	DTT [2][1]	DTT [1][1]	DTT [0][1]
DTT [3][0]	DTT [2][0]	DTT [1][0]	DTT [0][0]

FE0

89 Non-BodyBlock Features
34 State Machines
55 % Utilization

FE1

55 BodyBlock Features
20 State Machines
45 % Utilization

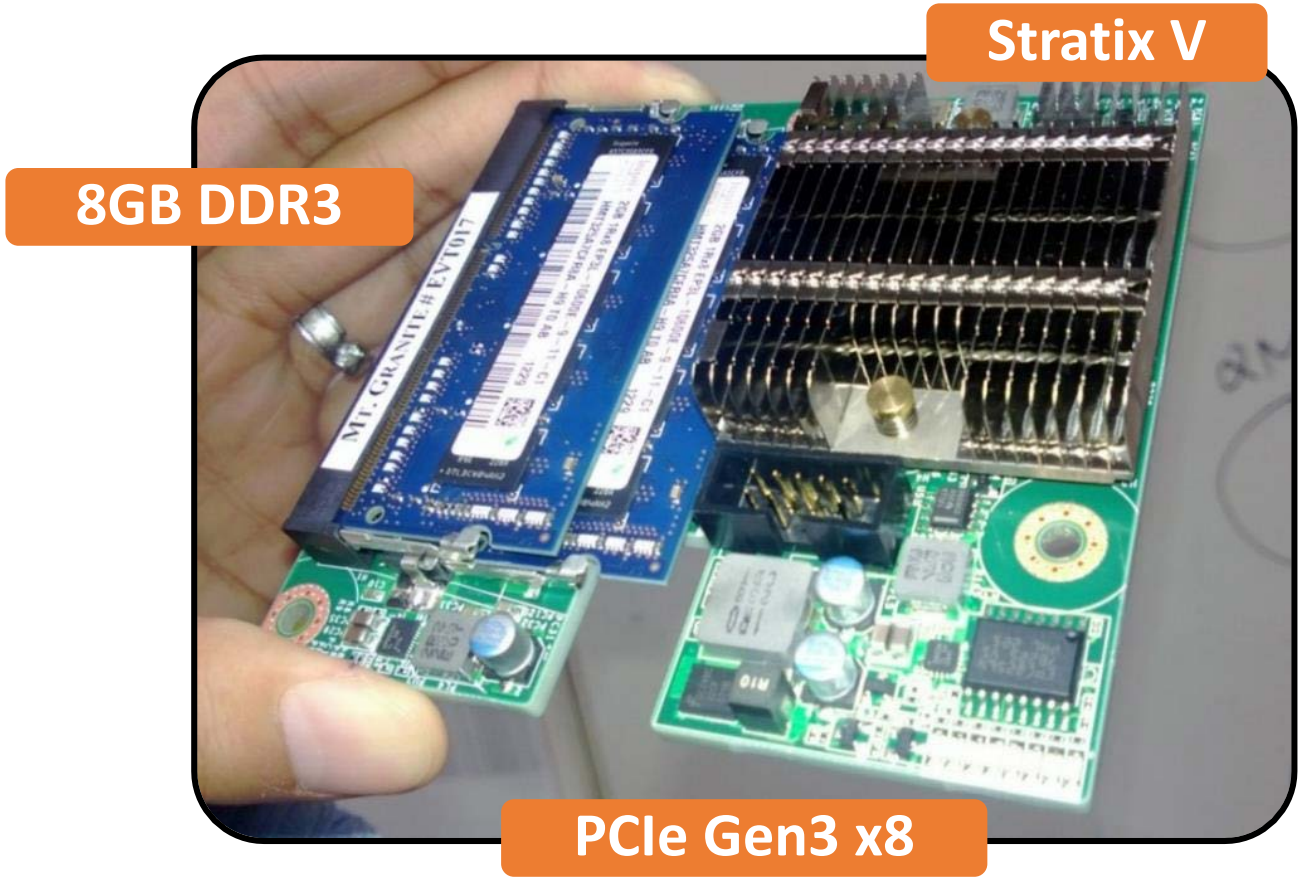


FFE: 64 cores / chip
256-512 threads
DTT: 48 DTT tiles/chip
240 tree processors
2880 trees/chip

- Fundamental flaws:
  - Additional single point of failure
  - Additional SKU to maintain
  - Too much load on the 1Gb network
  - Inelastic FPGA scaling or stranded capacity

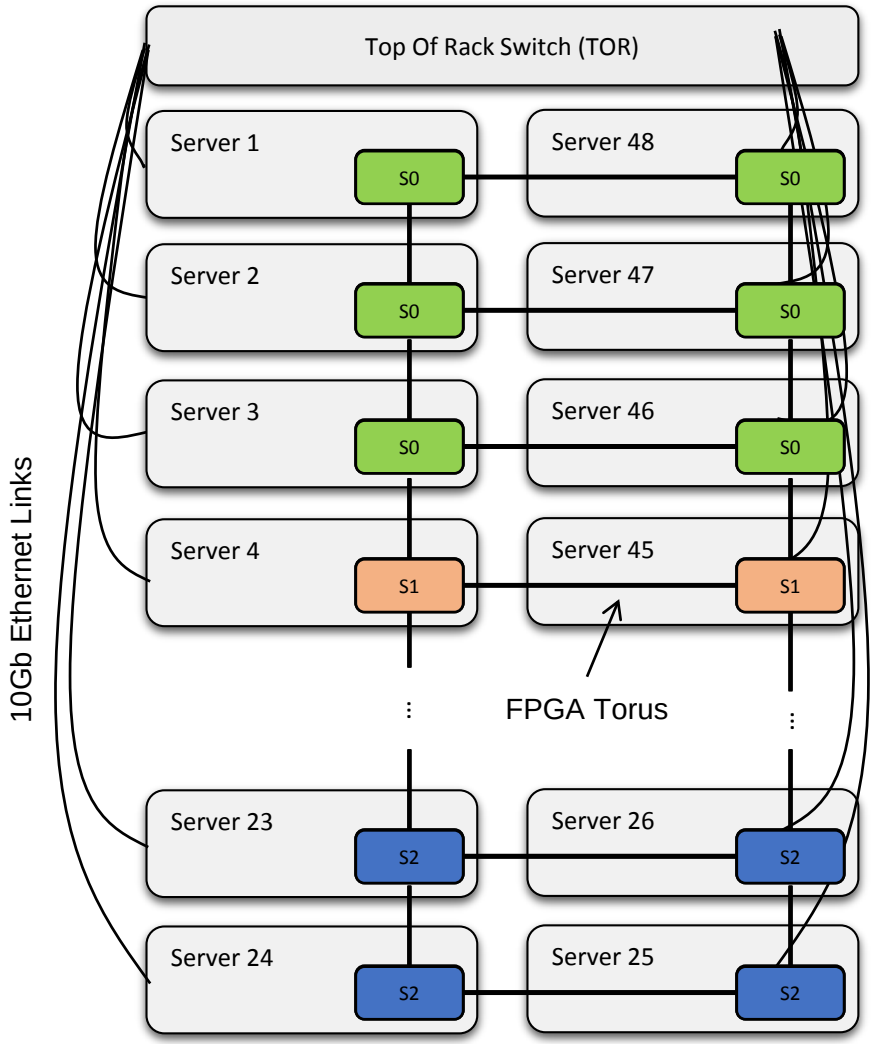
Catapult V1 Card (2012-2013)

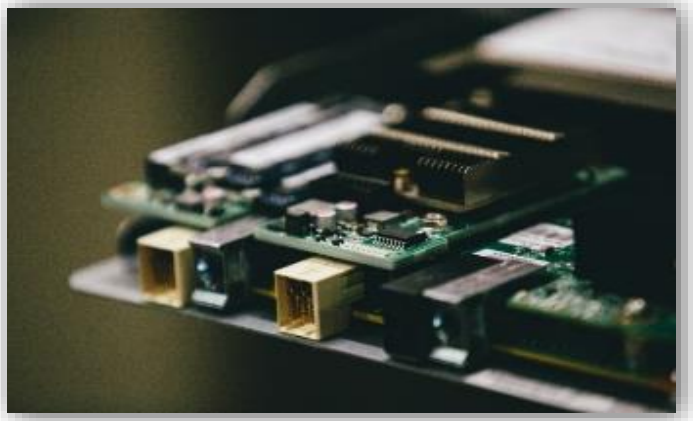
- Altera Stratix V D5
- 172.6K ALMs, 2014 M20Ks
  - 457KLEs
  - 1 KLE == ~12K gates
  - M20K is a 2.5KB SRAM
- PCIe Gen 2 x8, 8GB DDR3
- 20 Gb network among FPGAs

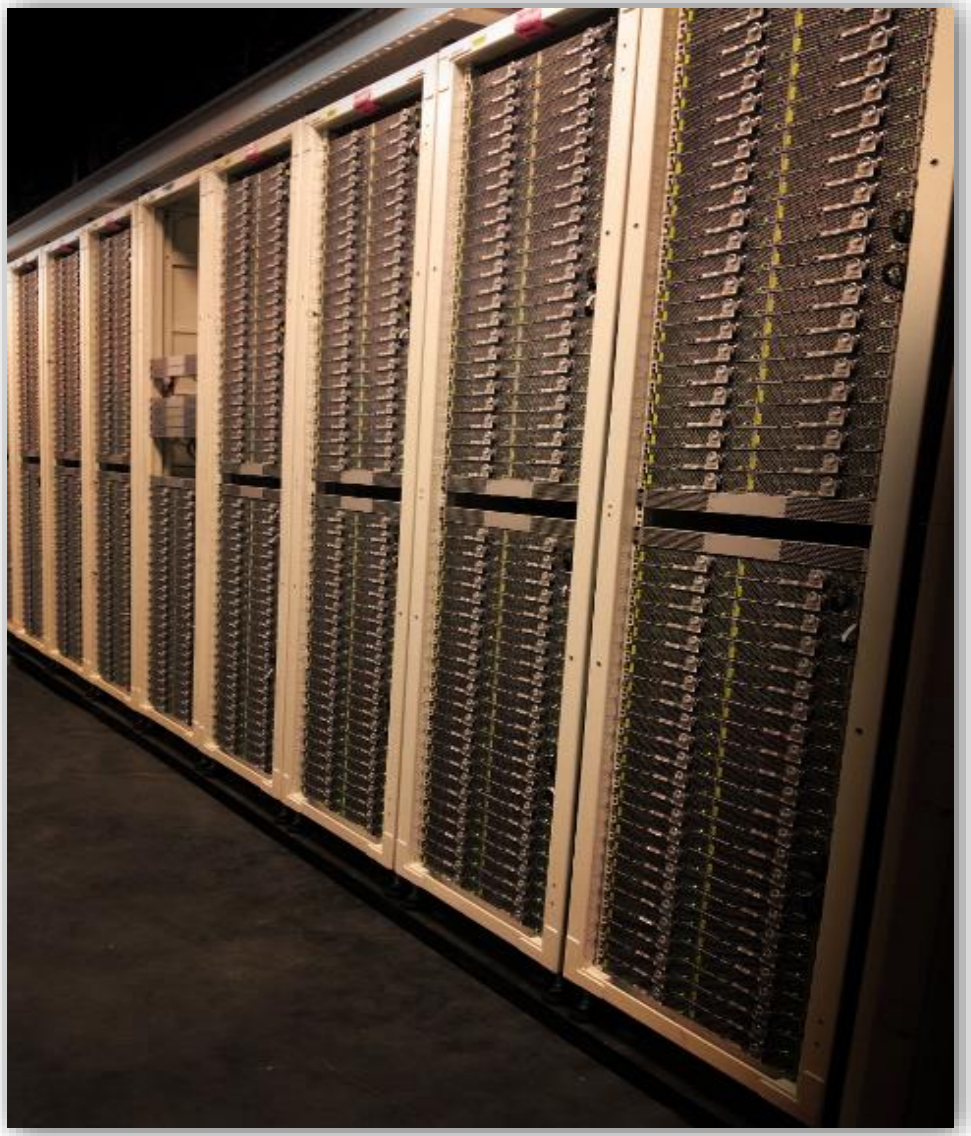


Mapped Fabric into a Pod

- 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



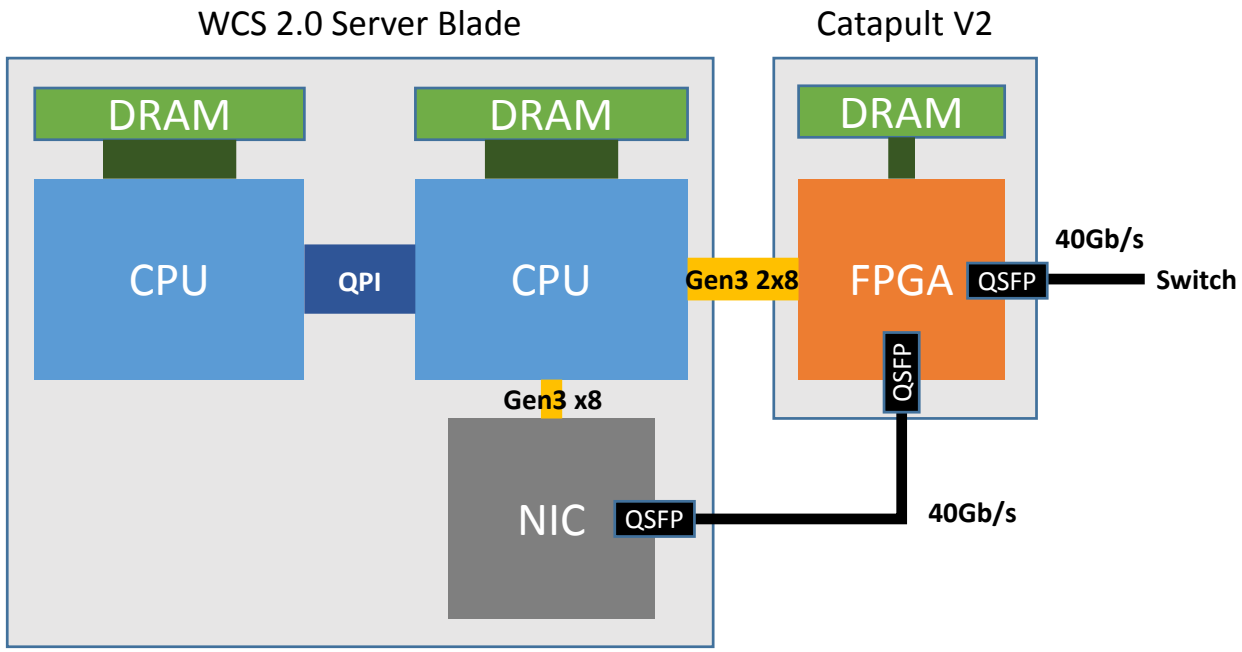




1,632 server pilot deployed in production datacenter

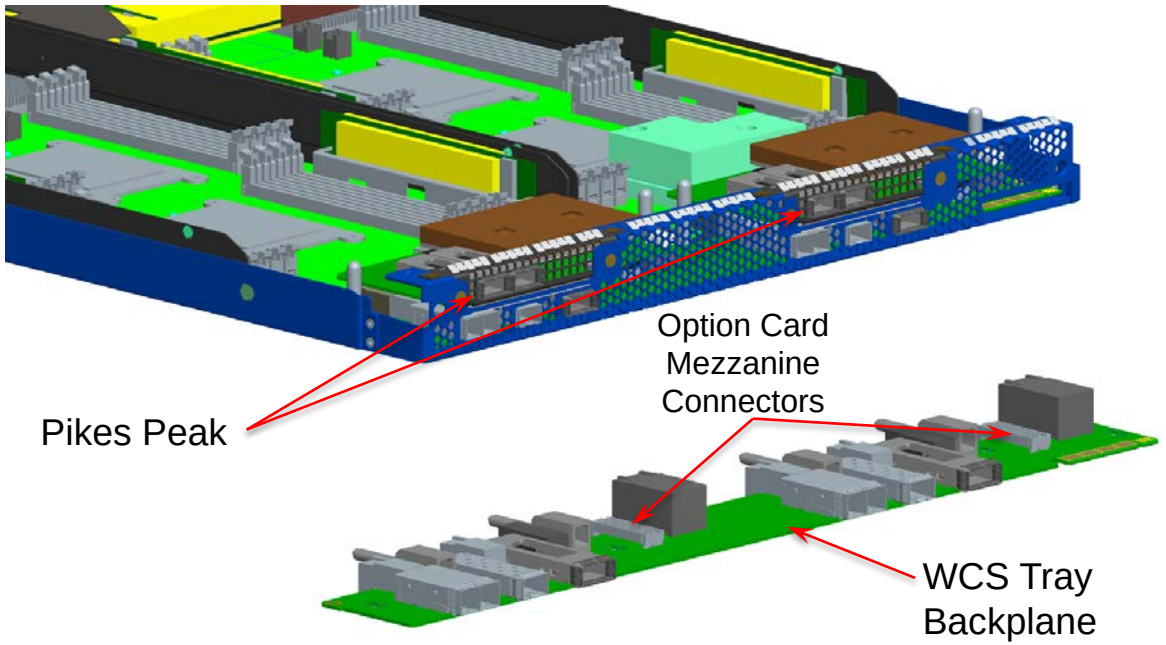
- Fundamental flaws:
  - Microsoft was converging on a single SKU
  - No one else wanted the secondary network
    - Complex, difficult to handle failures
    - Difficult to service boxes
  - No killer infrastructure accelerator
    - Application presence is too small

Catapult V2 Architecture



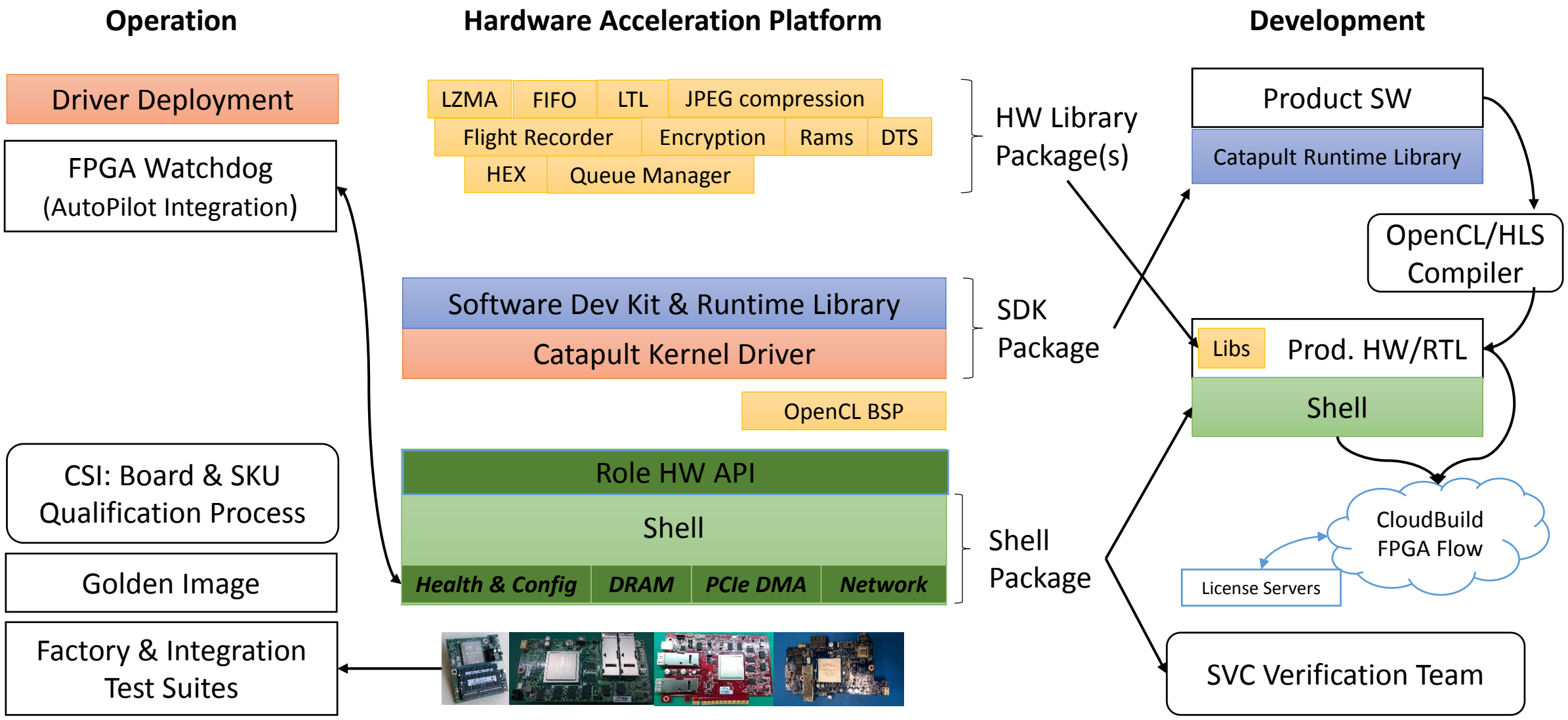
Catapult v2 Mezzanine card

WCS Gen4.1 Blade with Mellanox NIC and Catapult FPGA



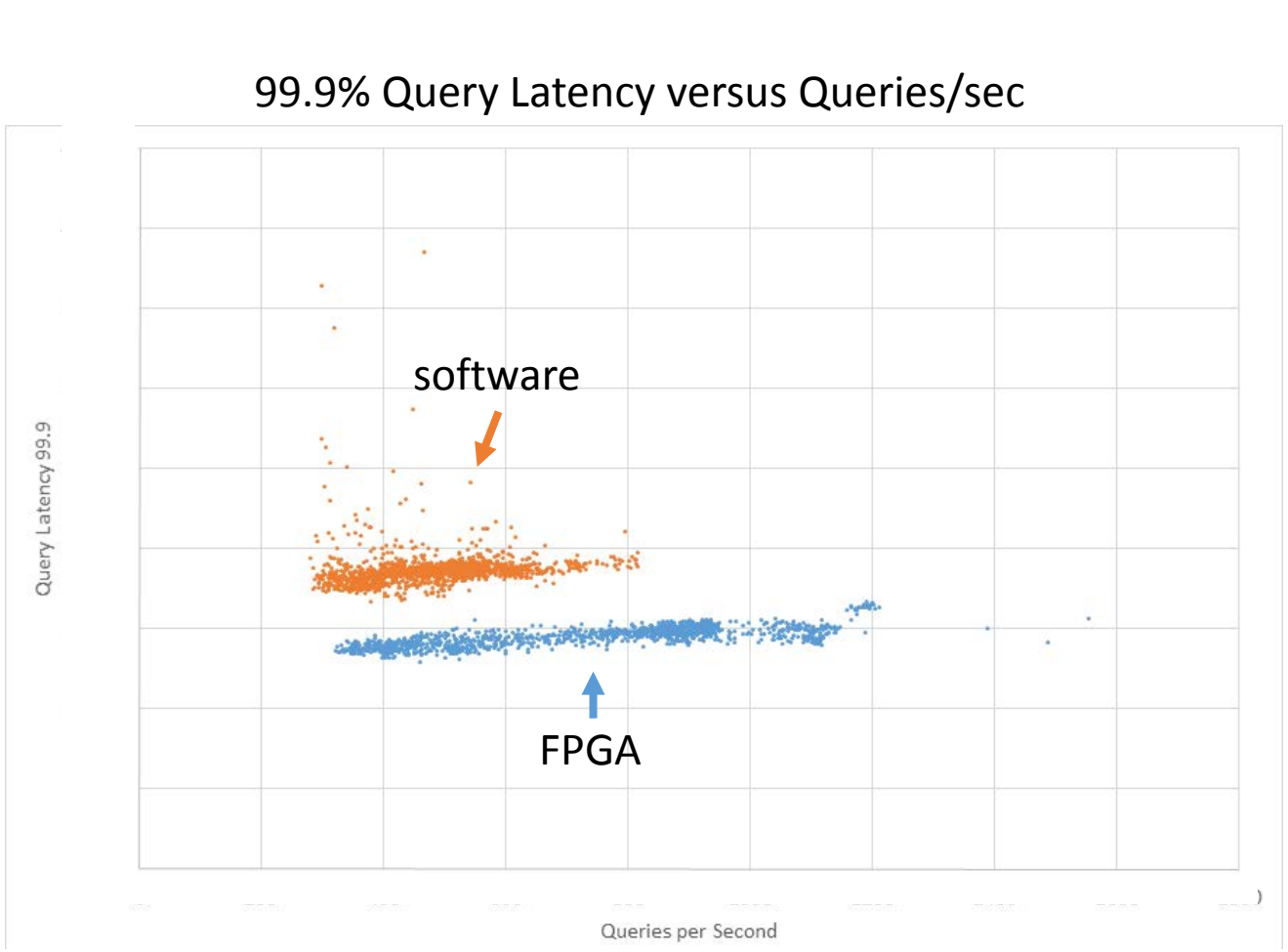
- The architecture justifies the economics
  1. Can act as a local compute accelerator
  2. Can act as a network/storage accelerator
  3. Can act as a remote compute accelerator

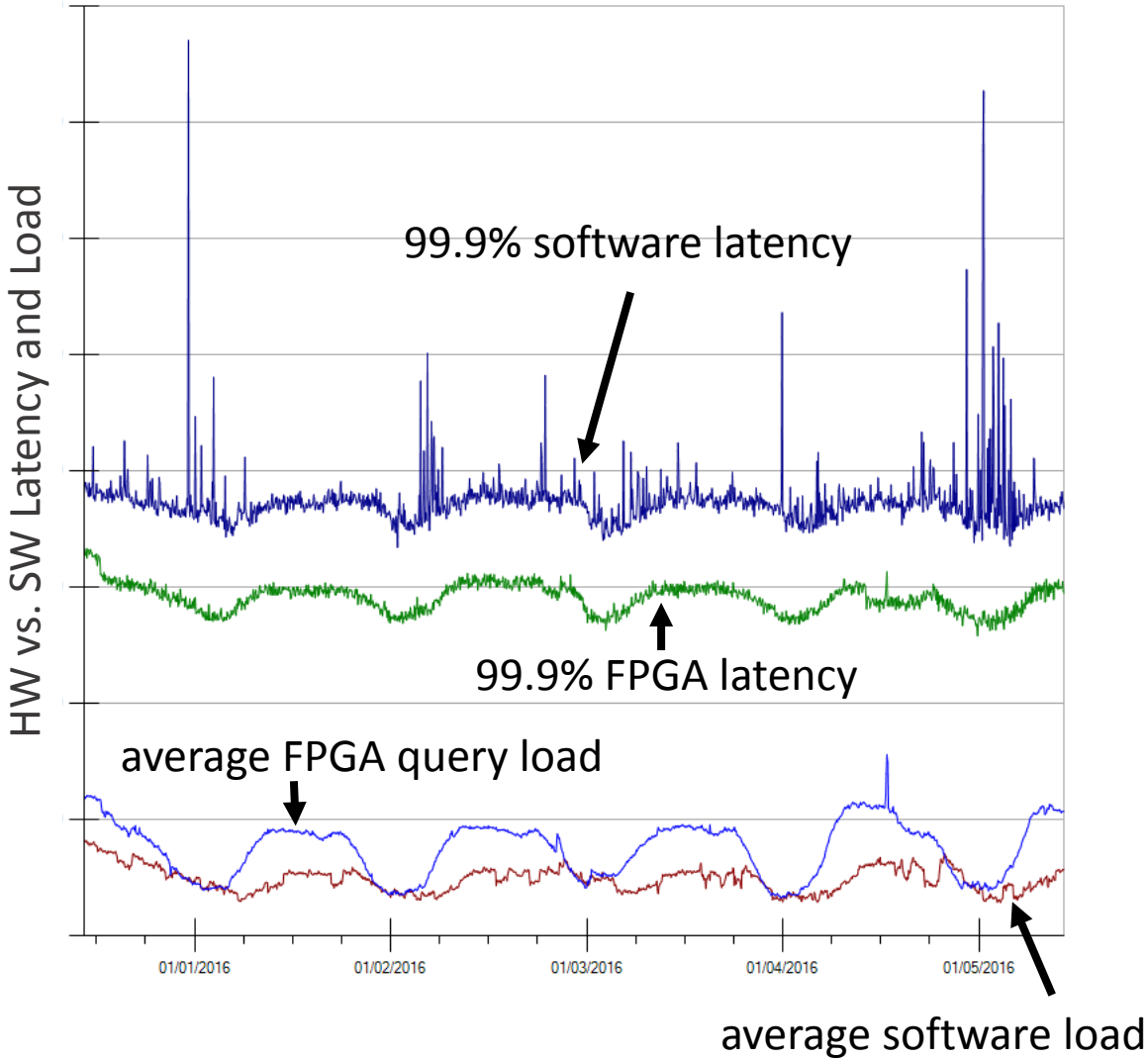
(Also need to build a complete platform)



Case 1: Use as a local accelerator

Production Results (December 2015)

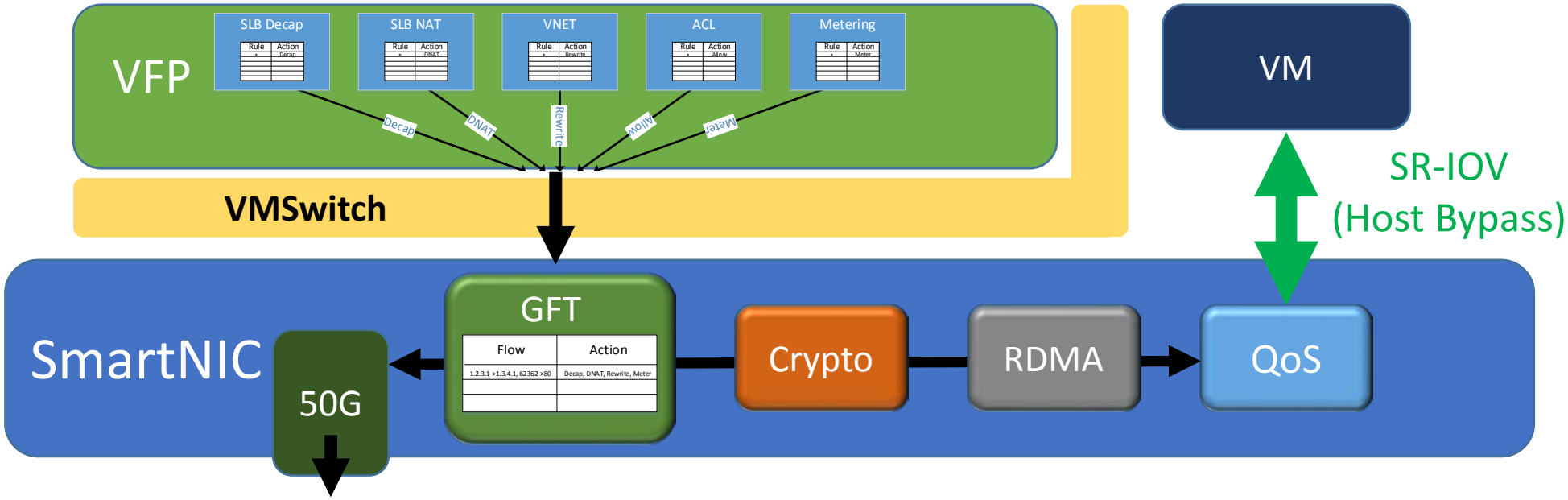




Case 2. Use as an infrastructure accelerator

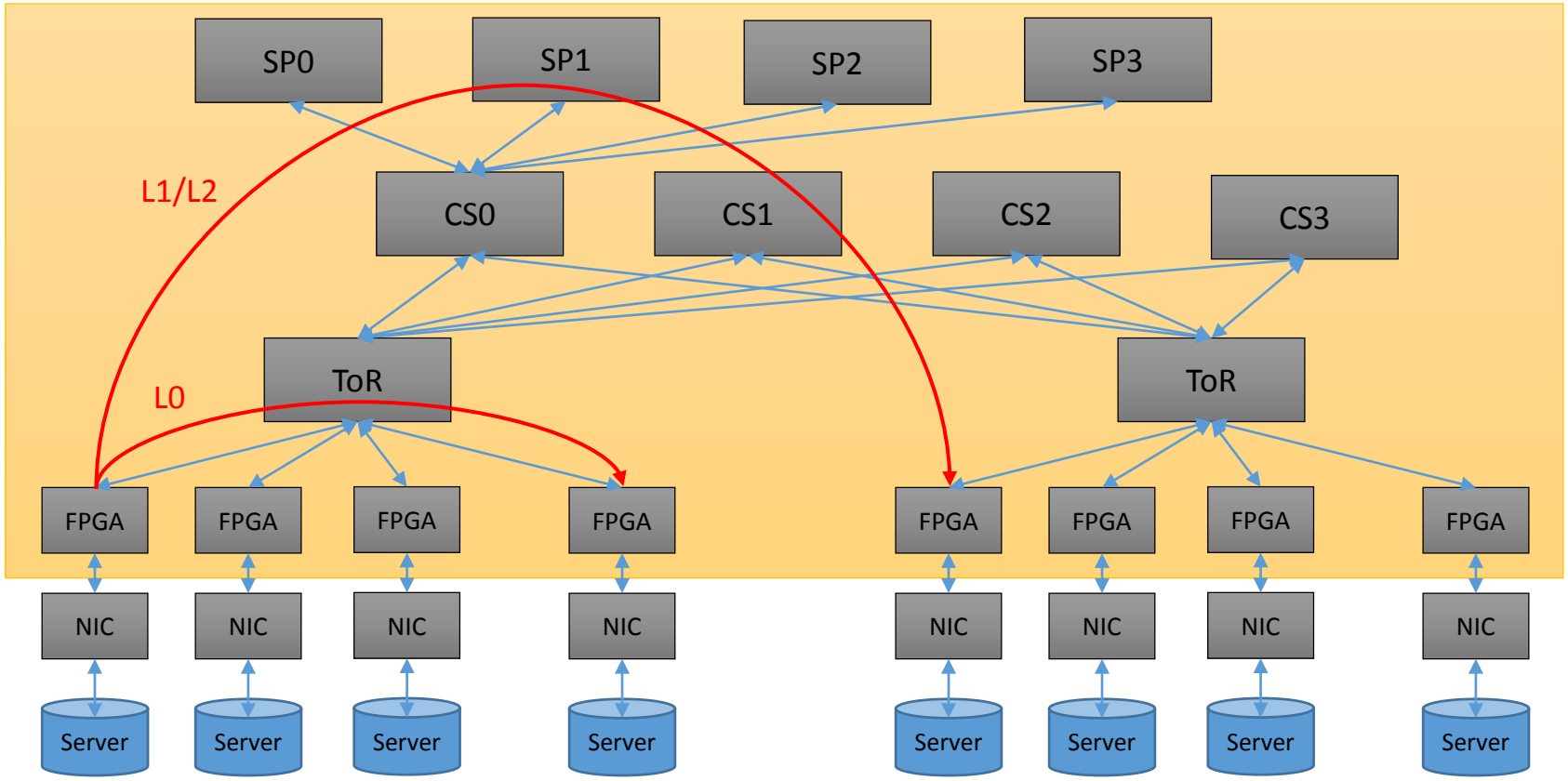
FPGA SmartNIC for Cloud Networking

- Azure runs Software Defined Networking on the hosts
  - Software Load Balancer, Virtual Networks – new features each month
- We rely on ASICs to scale and to be COGS-competitive at 40G+
  - But 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 will provide common functions like crypto, GFT, QoS, RDMA on all hosts
  - 40Gb/s network, 20Gb/s crypto takes a significant fraction of a 24-core machine
  - Example: crypto and vswitch inline on the FPGA: 0% CPU cost



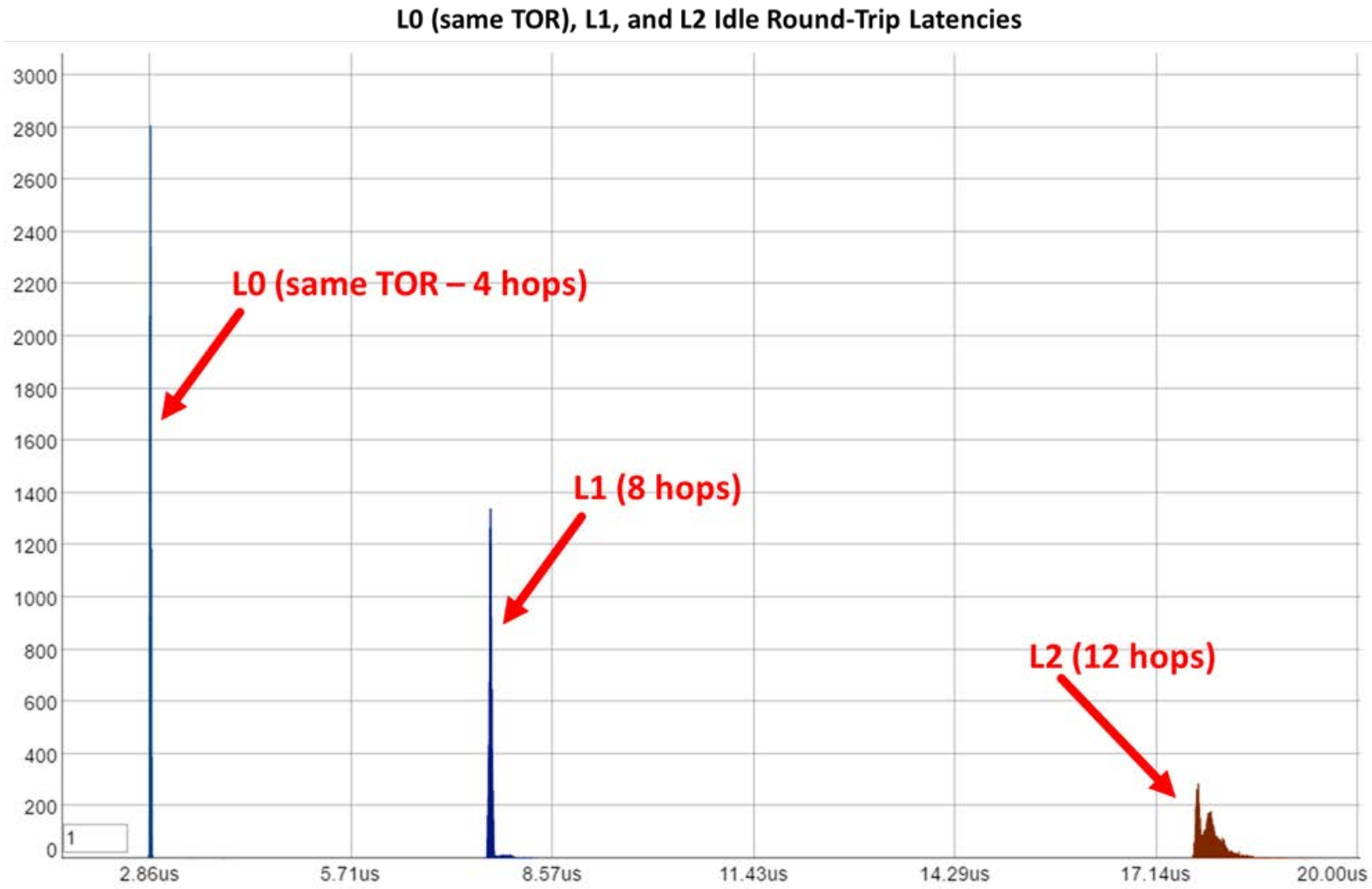
Case 3: Use as a remote accelerator

Inter-FPGA communication

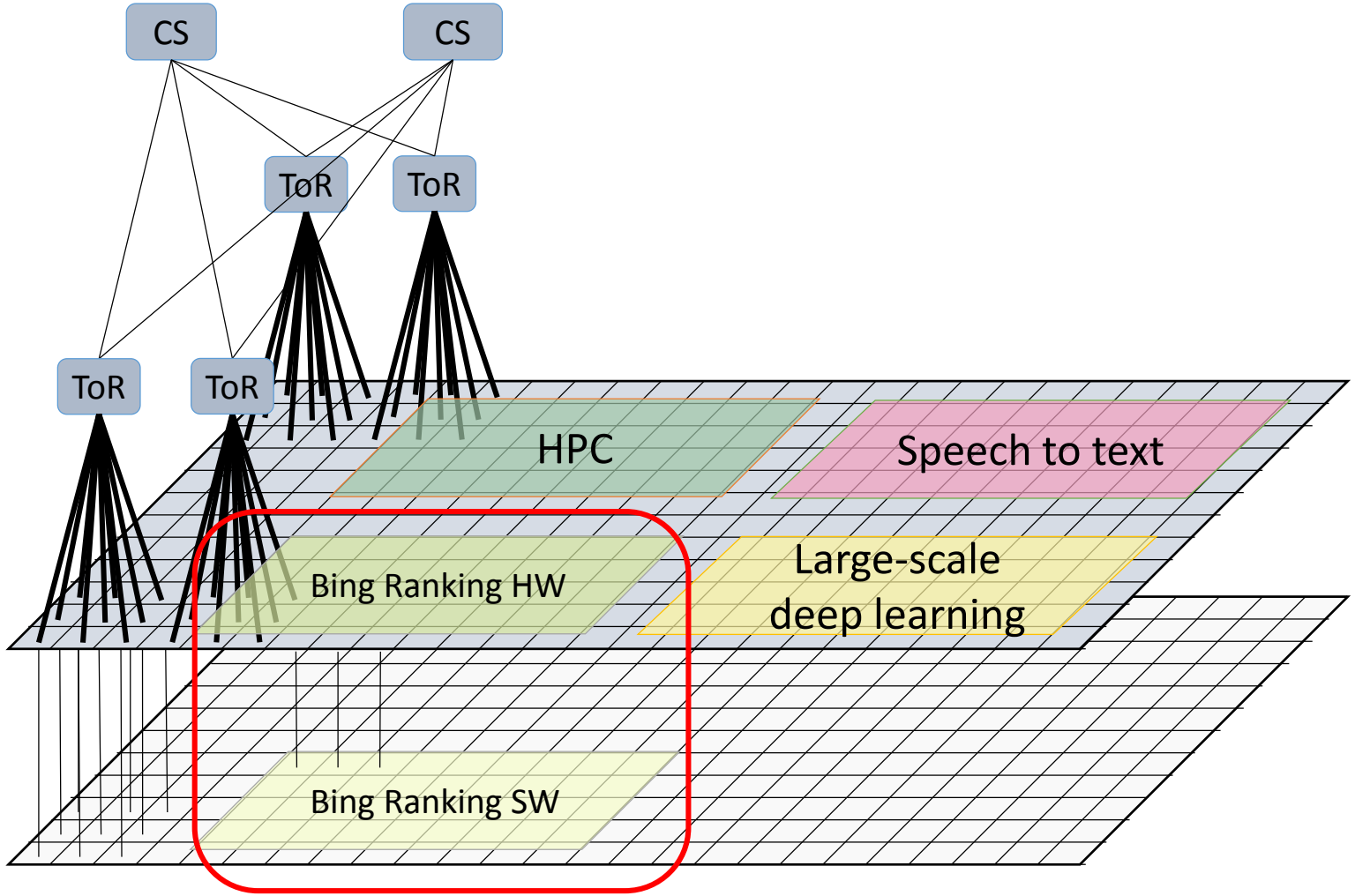


- FPGAs can encapsulate their own UDP packets
- Low-latency inter-FPGA communication (LTL)
- Can provide strong network primitives
- But this topology opens up other opportunities

FPGA-to-FPGA LTL Round-Trip Latencies



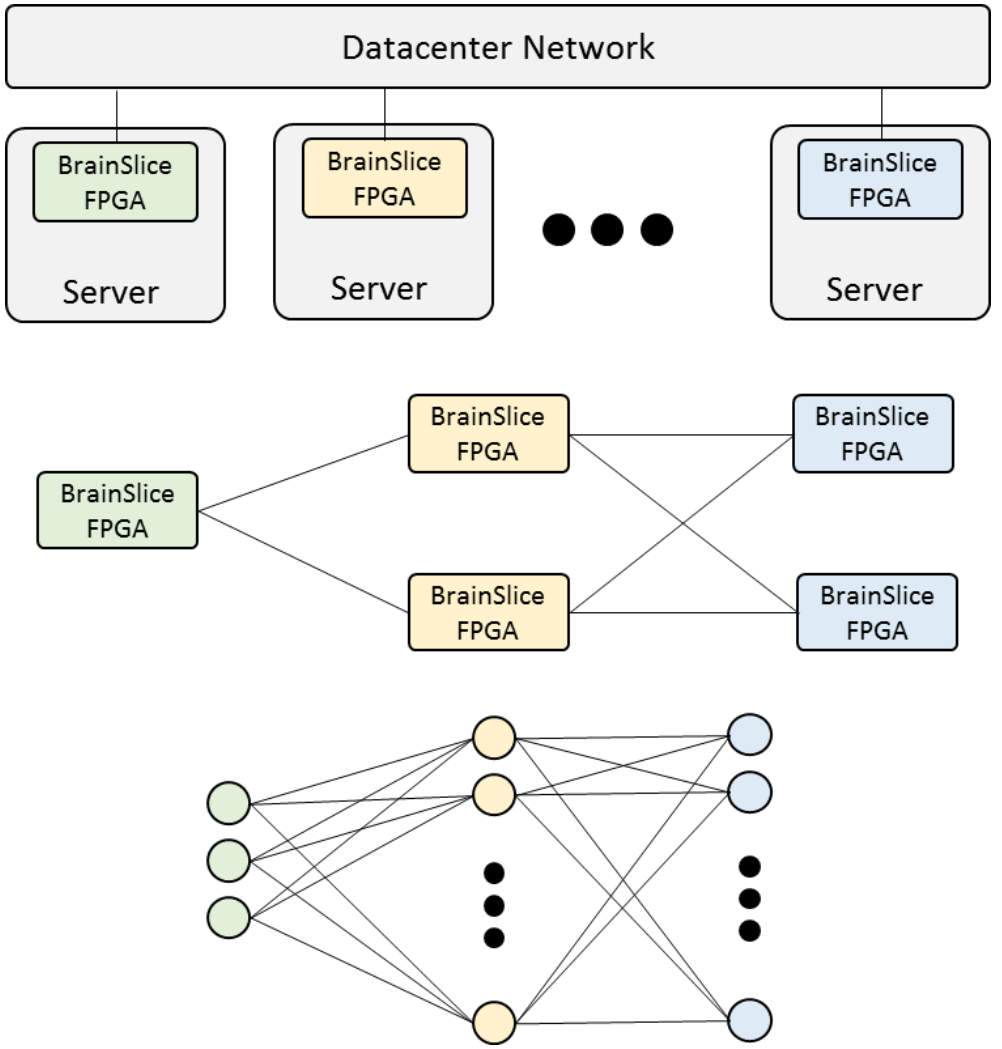
Hardware Acceleration as a Service



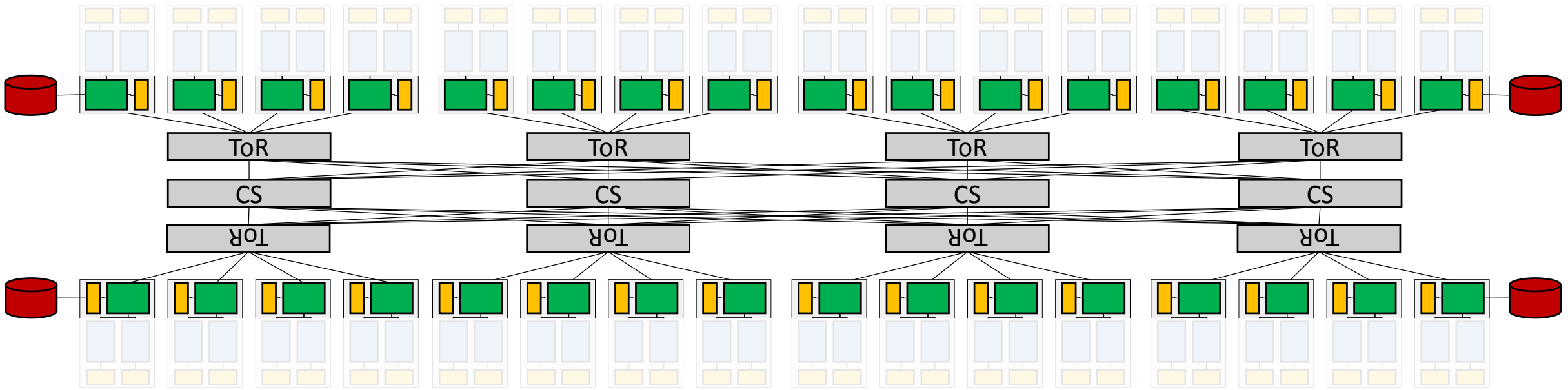
- Thanks to Stuart Byma
- Services may co-design with their local FPGAs or allocate a HaaS service remotely.
- Currently Bing ranking co-locates SW and HW fabric, but decoupling is trivial.

BrainWave: Scaling FPGAs To Ultra-Large Models

- Thanks to Eric Chung and team
- Distribute NN models across as many FPGAs as needed (up to thousands)
  - Recent Imagenet competition: 152-layer model
- Use HaaS and LTL to manage multi-FPGA execution
  - Very close to live production
- Only vectors travel over network
  - Low FPGA-FPGA latency at ~1.8us per L0 hop



V2 Architecture Makes Configurable Clouds Possible



- Massive amounts of programmable logic will change datacenter architecture broadly
- Is an independent computer running outside of the CPU domain
- Will affect network architecture (protocols, switches), storage architecture, security models

Will Catapult v2 be Deployed at Scale?

Configurable Clouds will Change the World

- Ability to reprogram a datacenter's hardware protocols
  - Networking, storage, security
- Can turn homogenous machines into specialized SKUs dynamically
- Unprecedented performance and low latency at hyperscale
  - Exa-ops of performance with a 10 microsecond diameter

- What would you do with the world's most powerful fabric?