Catapult: A Reconfigurable Fabric for Petaflop Computing in the Cloud

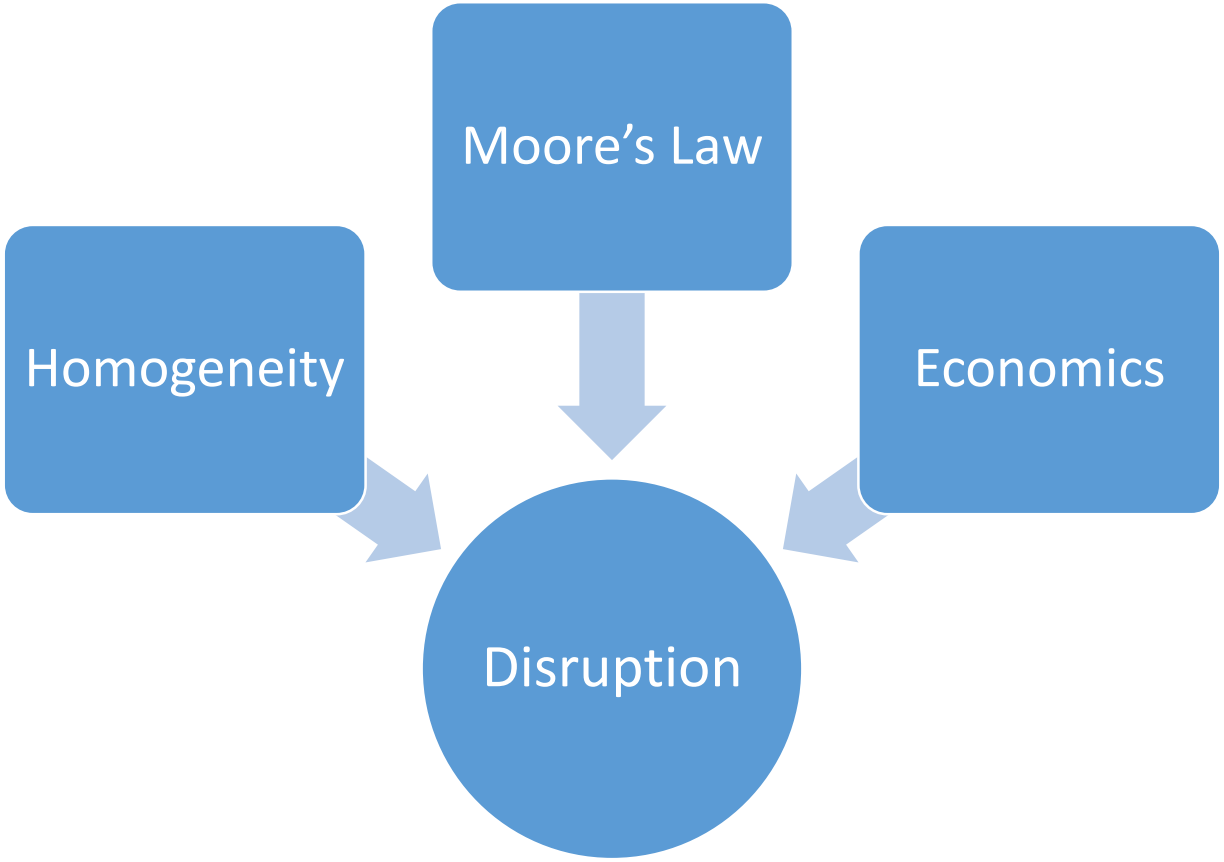
Doug Burger

Director, Hardware, Devices, & Experiences

MSR NExT

November 15, 2015

The Cloud is a Growing Disruptor for HPC



A 2-3 Horse Race

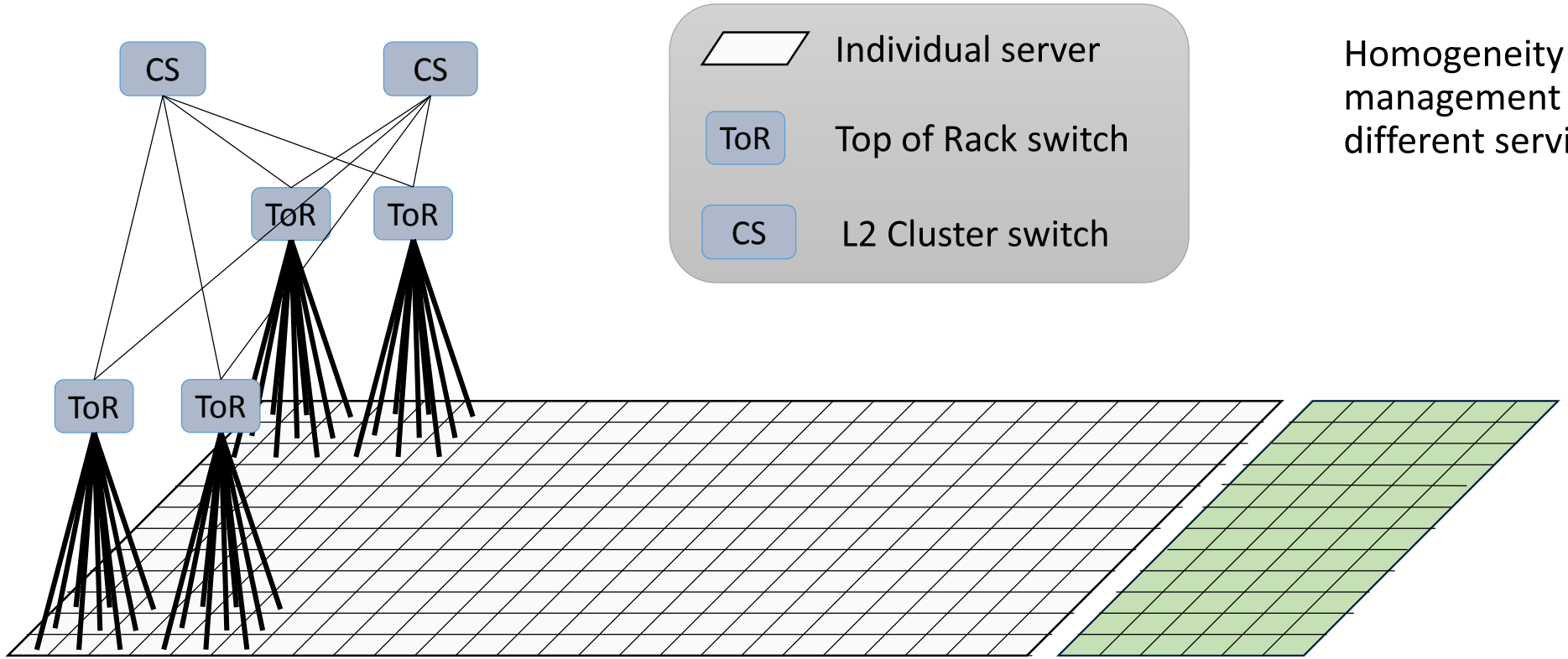
Figure 1. Magic Quadrant for Cloud Infrastructure as a Service, Worldwide

Source: Gartner (May 2015)

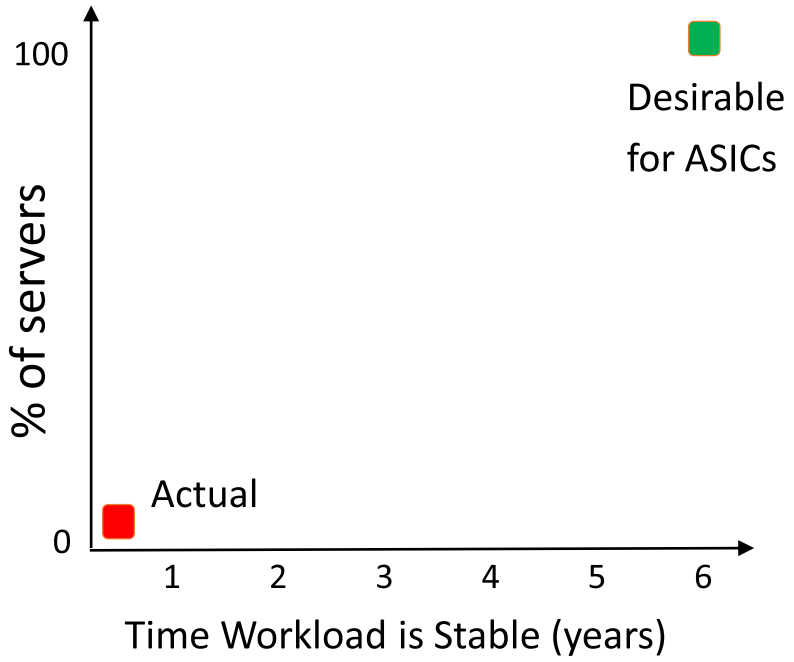
Hyperscale Cloud Fabrics

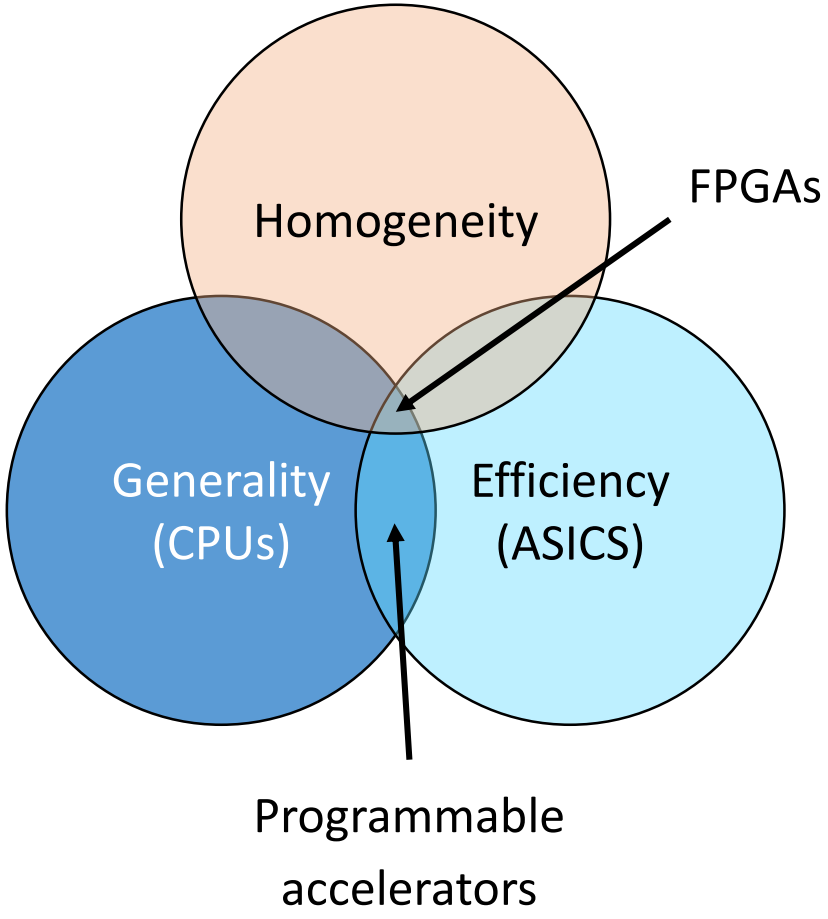
Datacenters today are formed of a software-based compute substrate; software running on CPUs in servers, connected via TCP over Ethernet through multiple layers of switches.

Homogeneity is highly desirable for management and fungibility across different services.



Accelerator Constraints of the Cloud





Catapult Project History

- December 9, 2010 – initial meeting
  - Christmas break 2010: feasible to accelerate ranking?
  - January 12, 2011 – Meeting with Bing leadership
- 2011 – v0: ported then Bing ranking stack, built BFB board
- 2012 – v1: developed distributed architecture
- 2013 – Took v1 to scale, Bing pilot
- 2014 – v2: developed new architecture, commenced work with Azure
- 2015 – Mainstreamed: production and expansion
  - Intel announced Altera acquisition, $16.7B

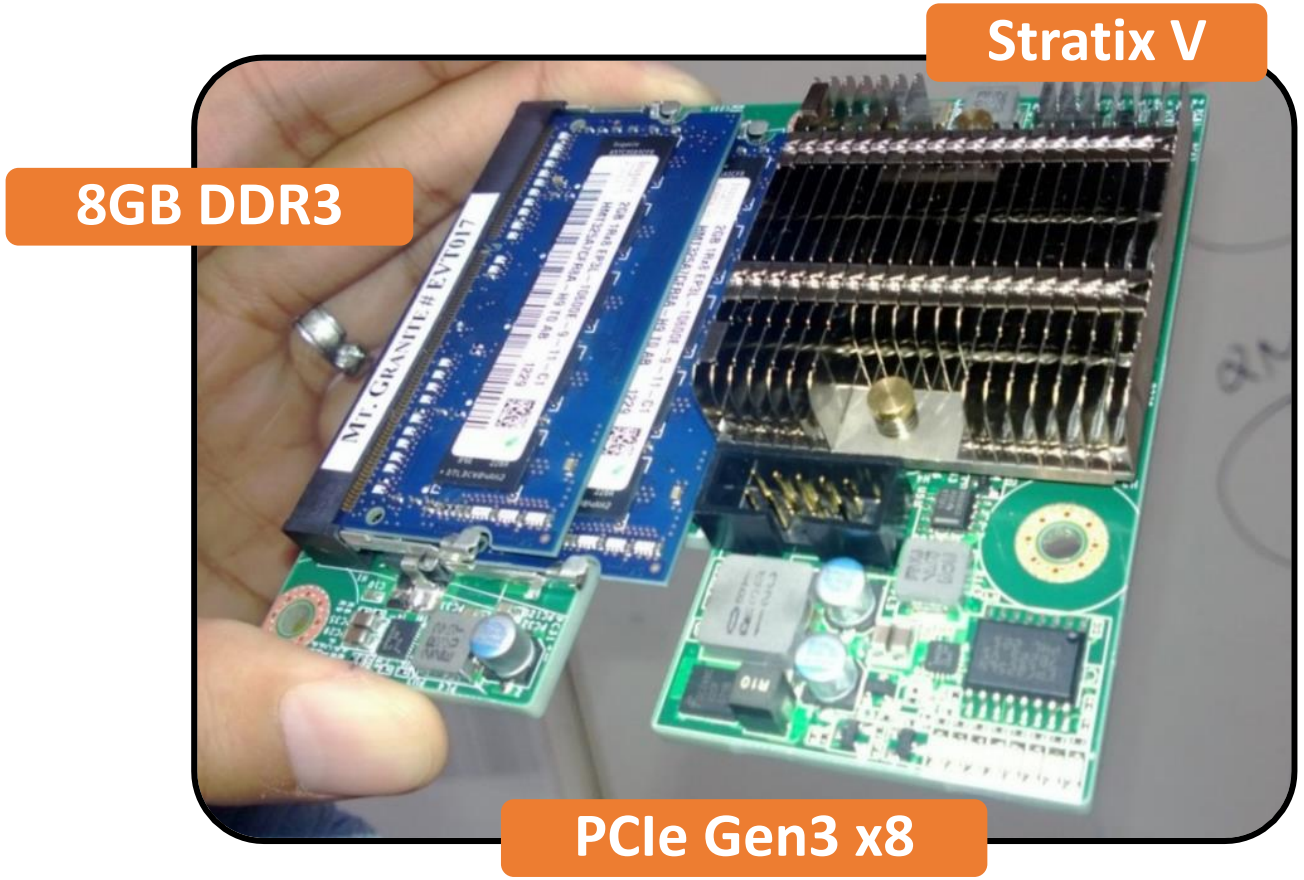
Microsoft Open Compute Server

Two 8-core Xeon 2.1 GHz CPUs
64 GB DRAM
4 HDDs, 2 SSDs
10 Gb Ethernet
No cable attachments to server

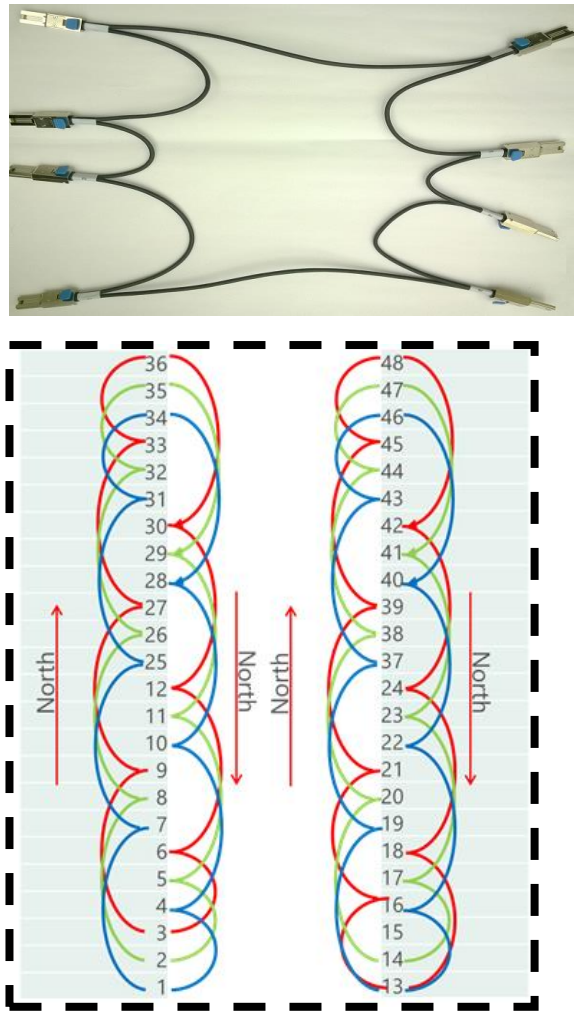
Air flow

Catapult V1 Accelerator Card

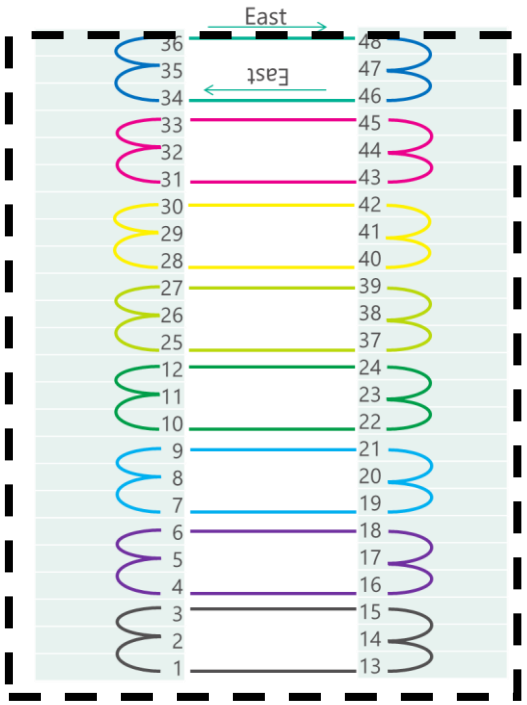
- Altera Stratix V D5
- 172.6K ALMs, 2014 M20Ks
  - 457KLEs
  - 1 KLE == ~12K gates
  - M20K is a 2.5KB SRAM
- PCIe Gen 2 x8, 8GB DDR3
- 20 Gb network among FPGAs



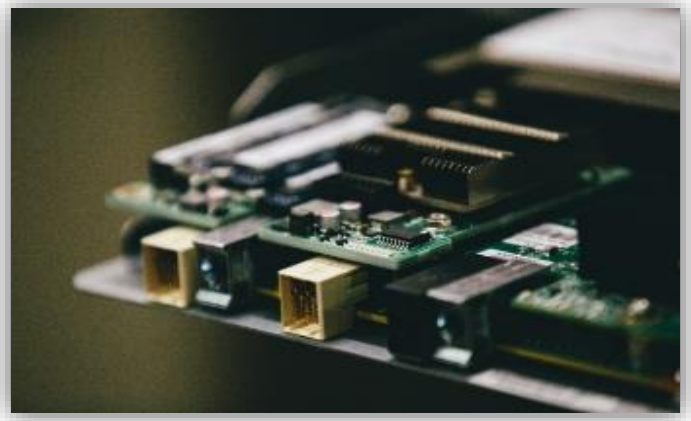
6x8 Torus in a 2x24 Server Layout



8-Shell Cables

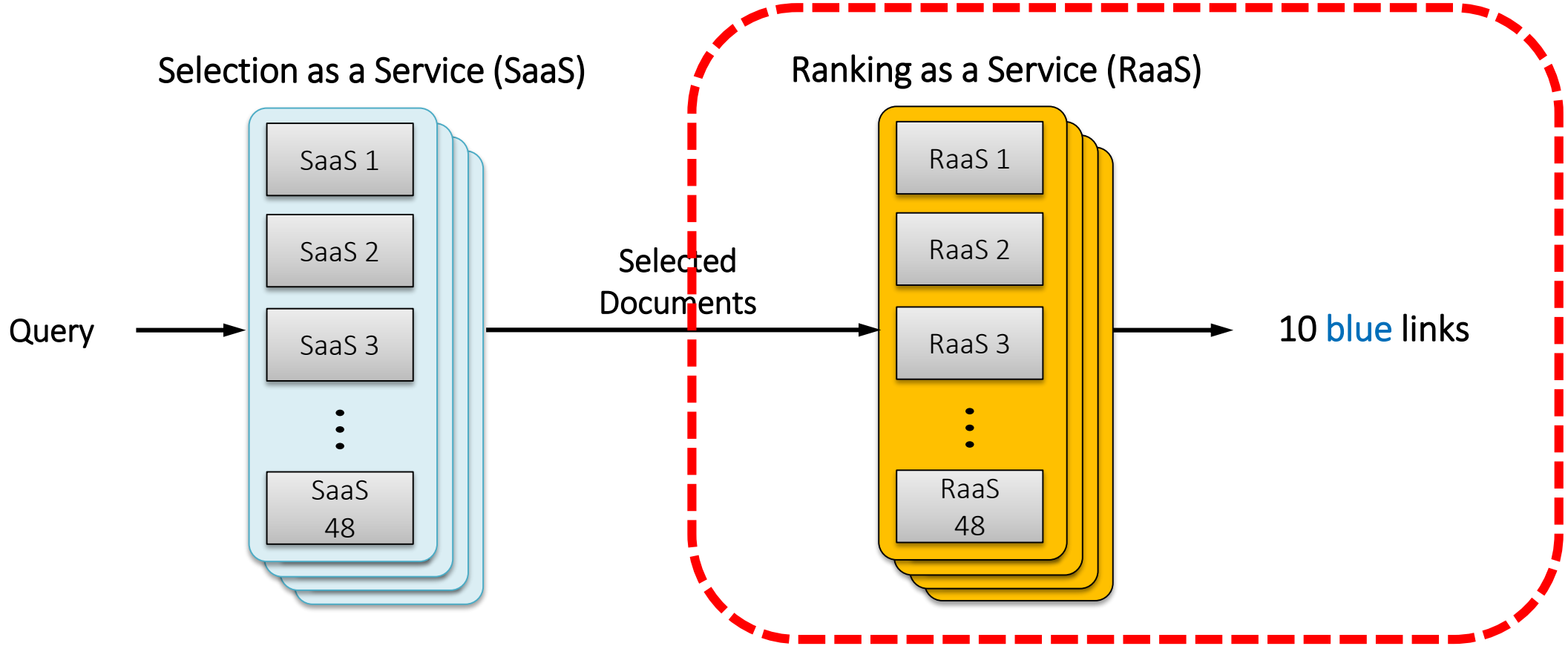


6-Shell Cables



1,632 server pilot deployed in production BN datacenter

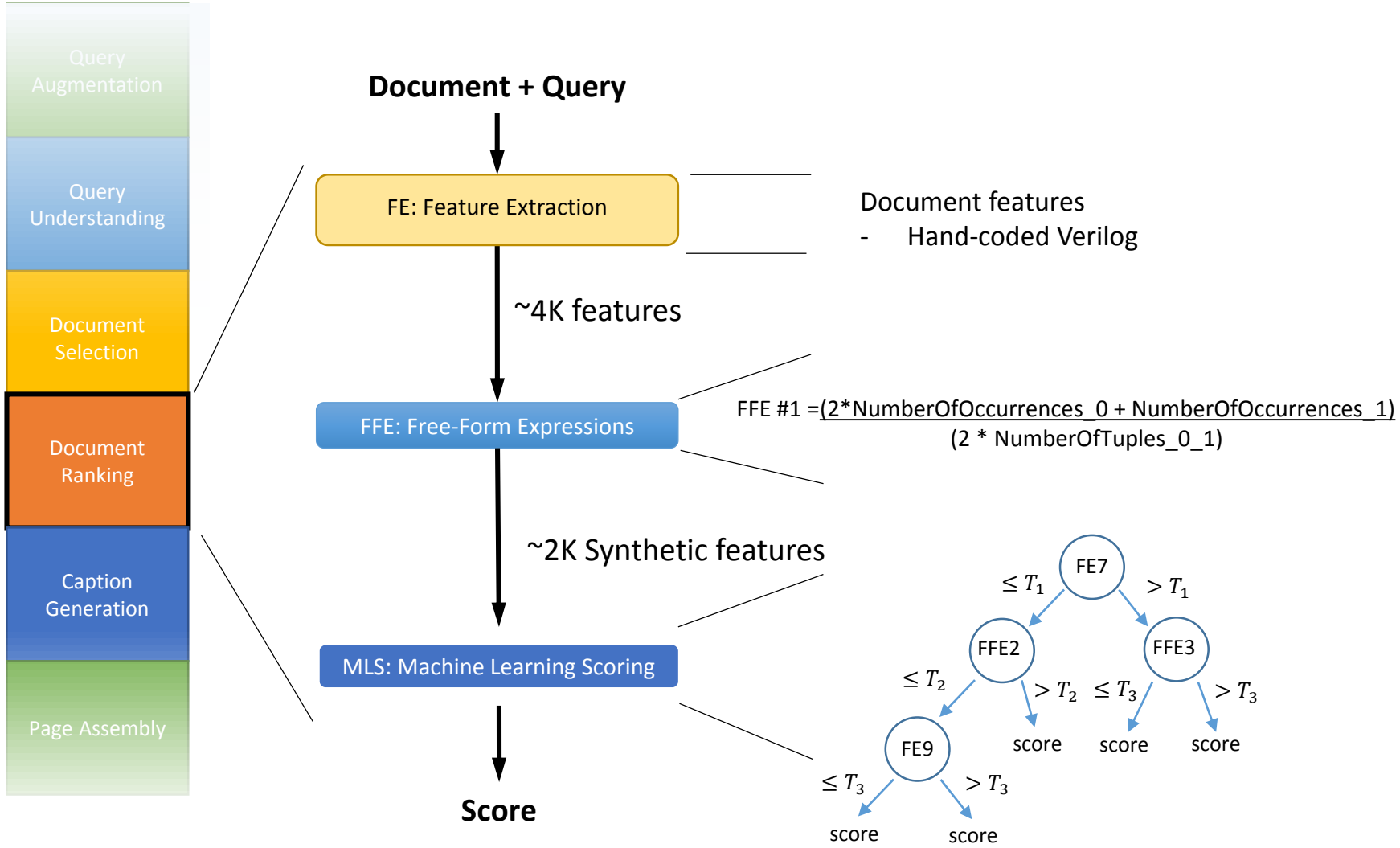
Target: Accelerate Ranking as a Service

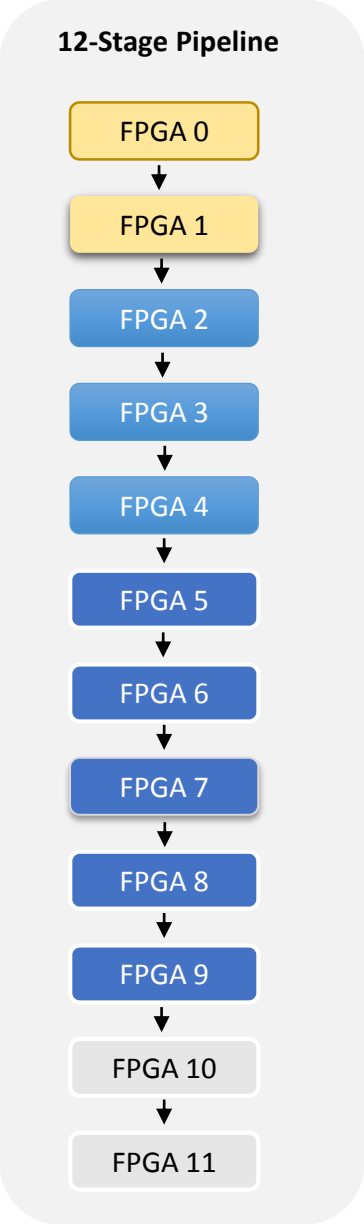


Selection-as-a-Service (SaaS)
- Find all docs that contain query terms
- Filter and select candidate documents for ranking

Ranking-as-a-Service (RaaS)
- Compute relevance scores for each selected doc
- Sort the scores and return the results

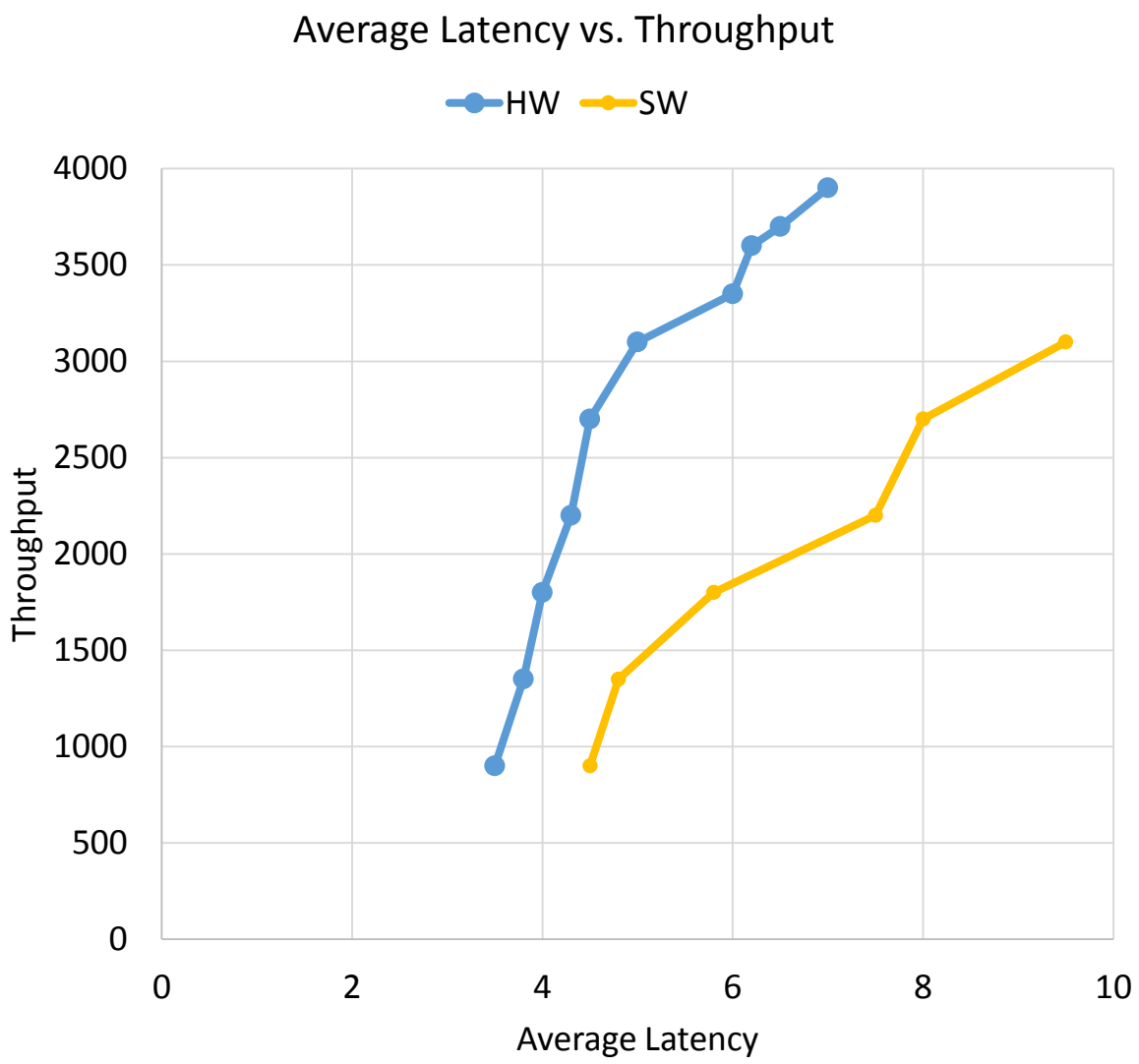
FPGA Accelerator for Bing Ranking

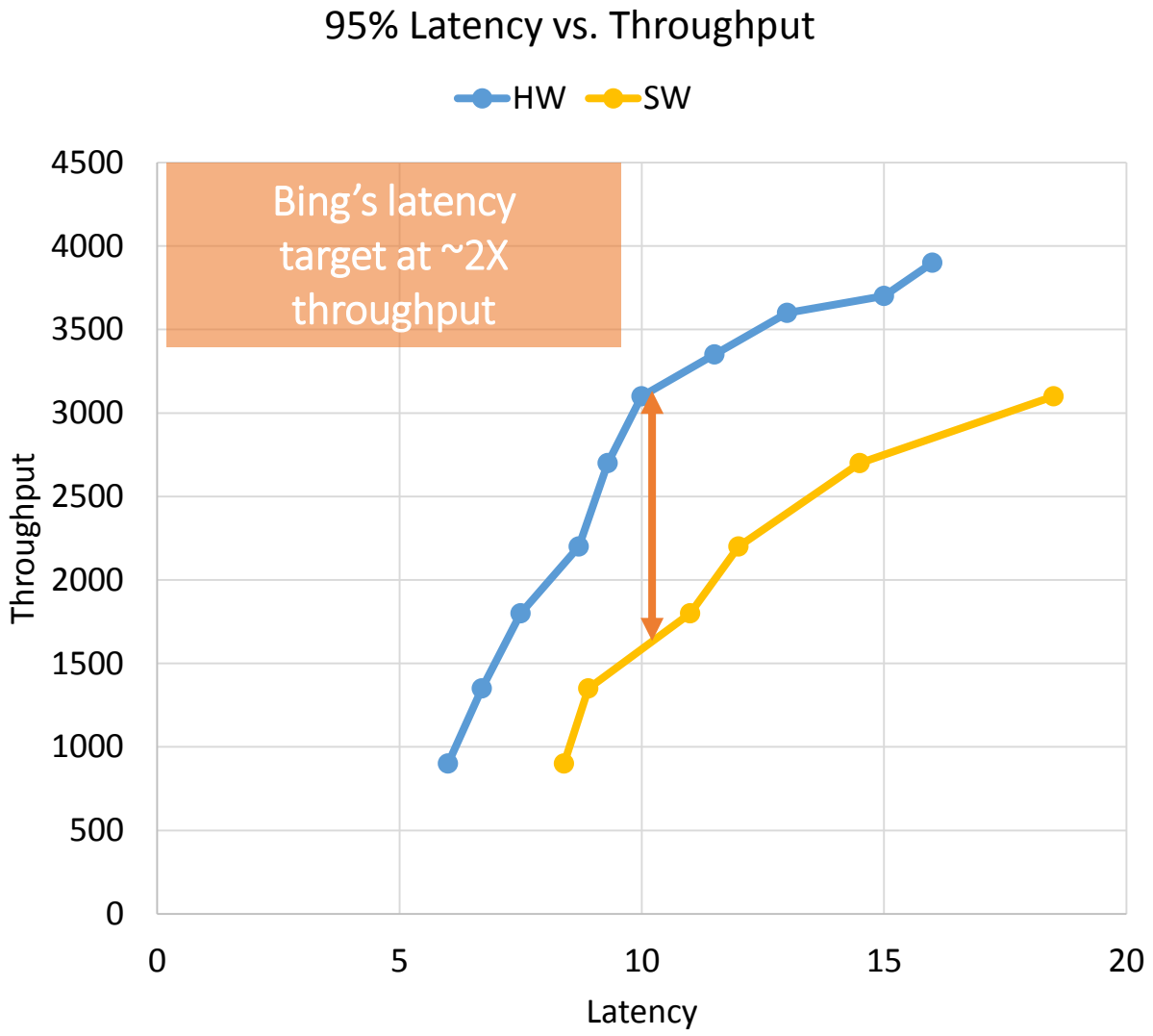




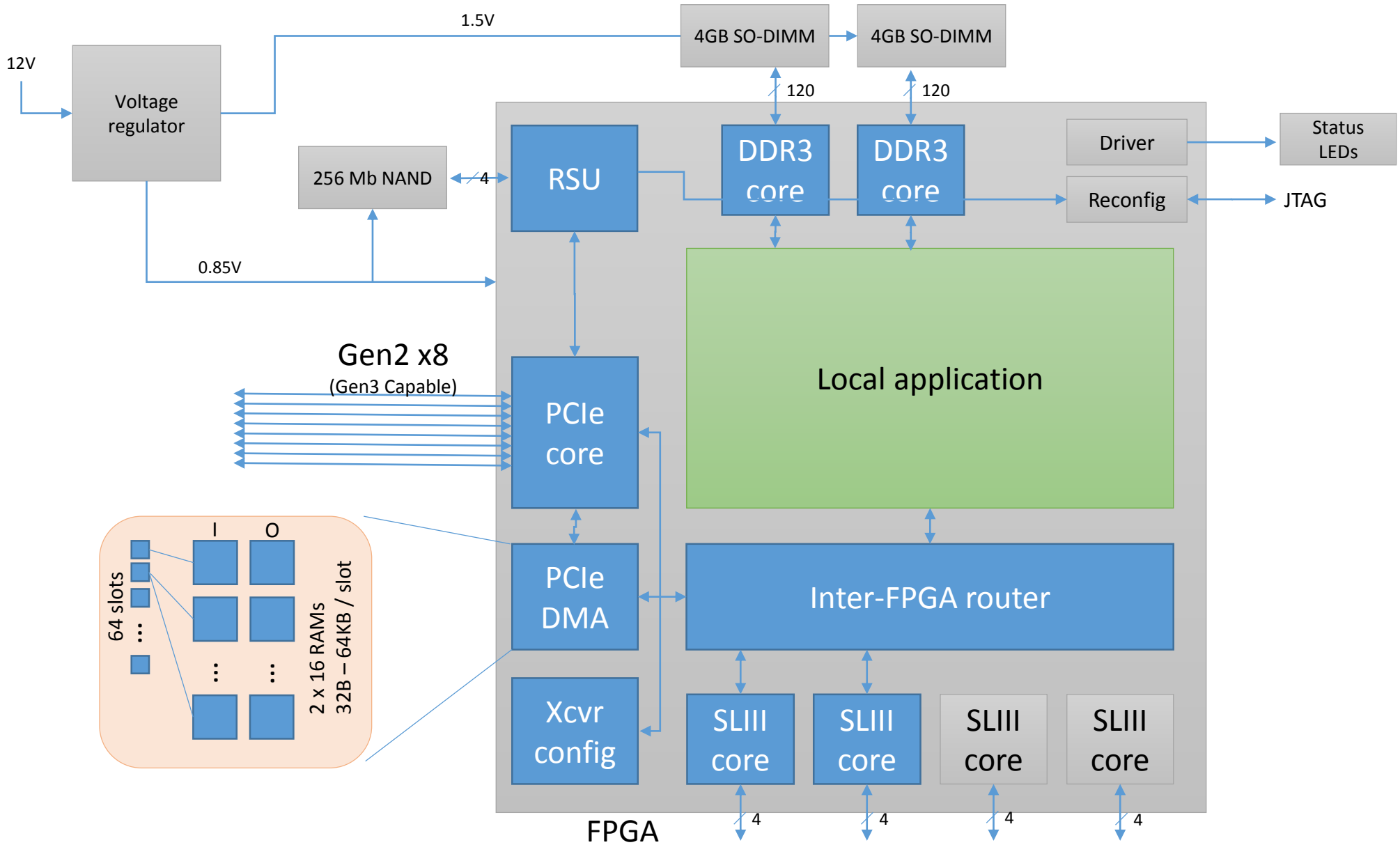
Demonstrated ~2x throughput gain and stability justifying production

Pilot Results (FPGA vs. Software)





Catapult V1 Shell Architecture

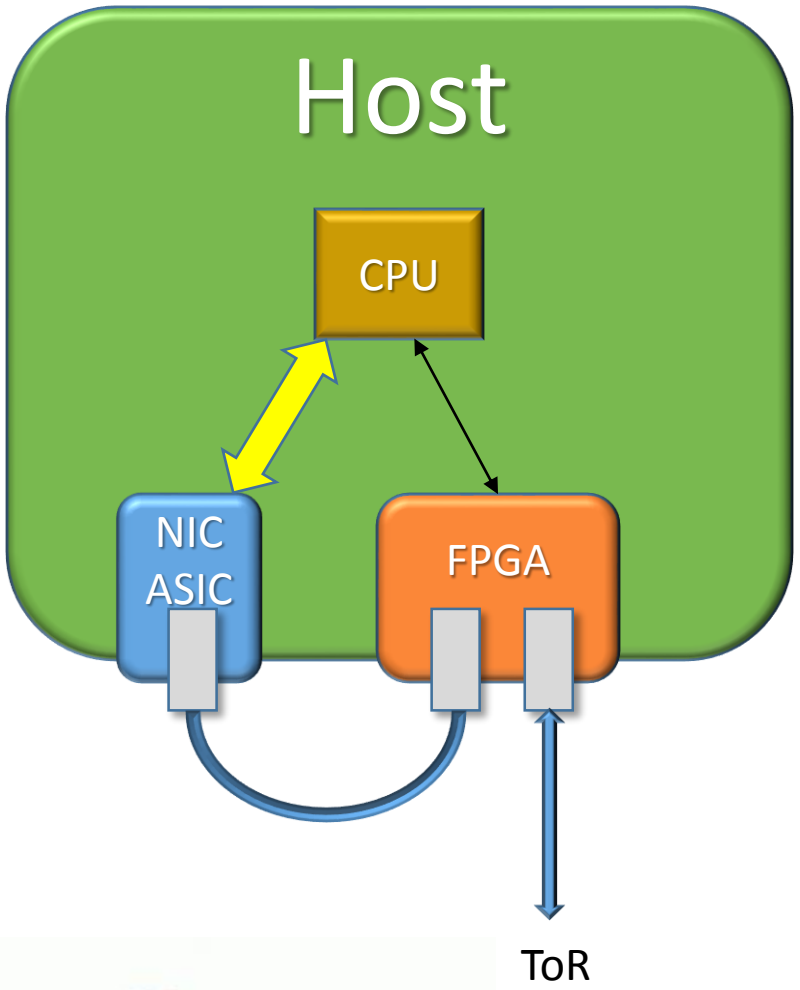


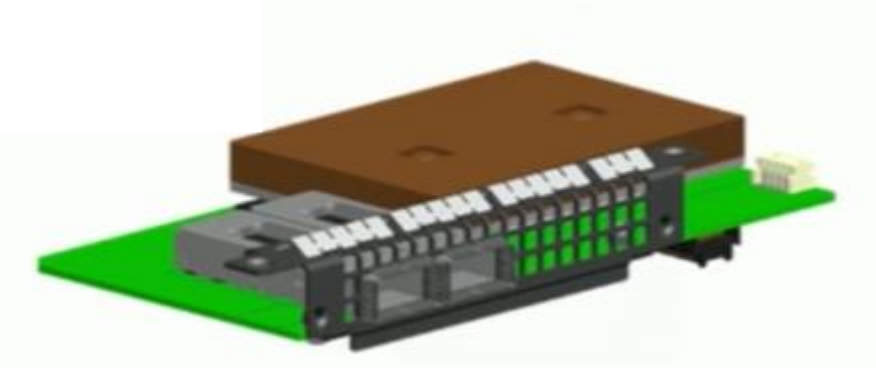
Production issues at scale

- Build system
  - License servers, availability of source, build machines
- Scale-out qualification of IP
- Clean interfaces for high-productivity development environment
- Shell/driver/application versioning and deployment
  - Backwards compatibility
- Health monitoring and failure diagnostics
  - Continuous reporting of interfaces health, soft error rate, etc.
- Debugging (esp. on livesite)
  - Flight Data Recorder to replay bug-generating condition
- System integrity testing - many servers/vendors
- Scalability of verification
- *In situ* updates to drivers, golden image, shell
- Supply chain management

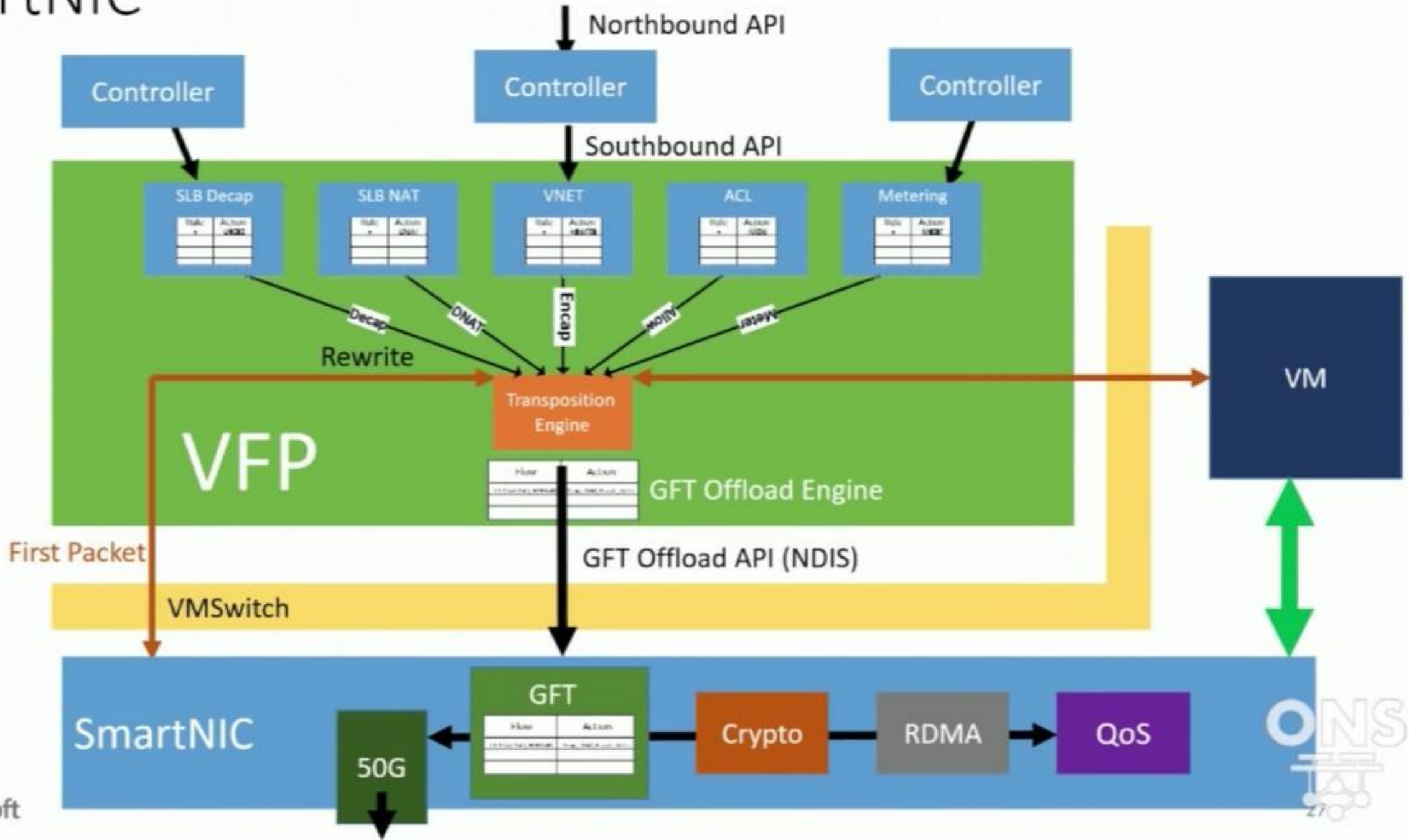
Azure SmartNIC

- Announced at ONS
- Use an FPGA for reconfigurable functions
  - FPGAs are already used in Bing (Catapult)
  - Roll out hardware as we do software
- Programmed using Generic Flow Tables (GFT)
  - Language for programming SDN to hardware
  - Uses connections and structured actions as primitives
- SmartNIC can also do Crypto, QoS, stora acceleration, and more …
  - 40Gb bidirectional AES demo





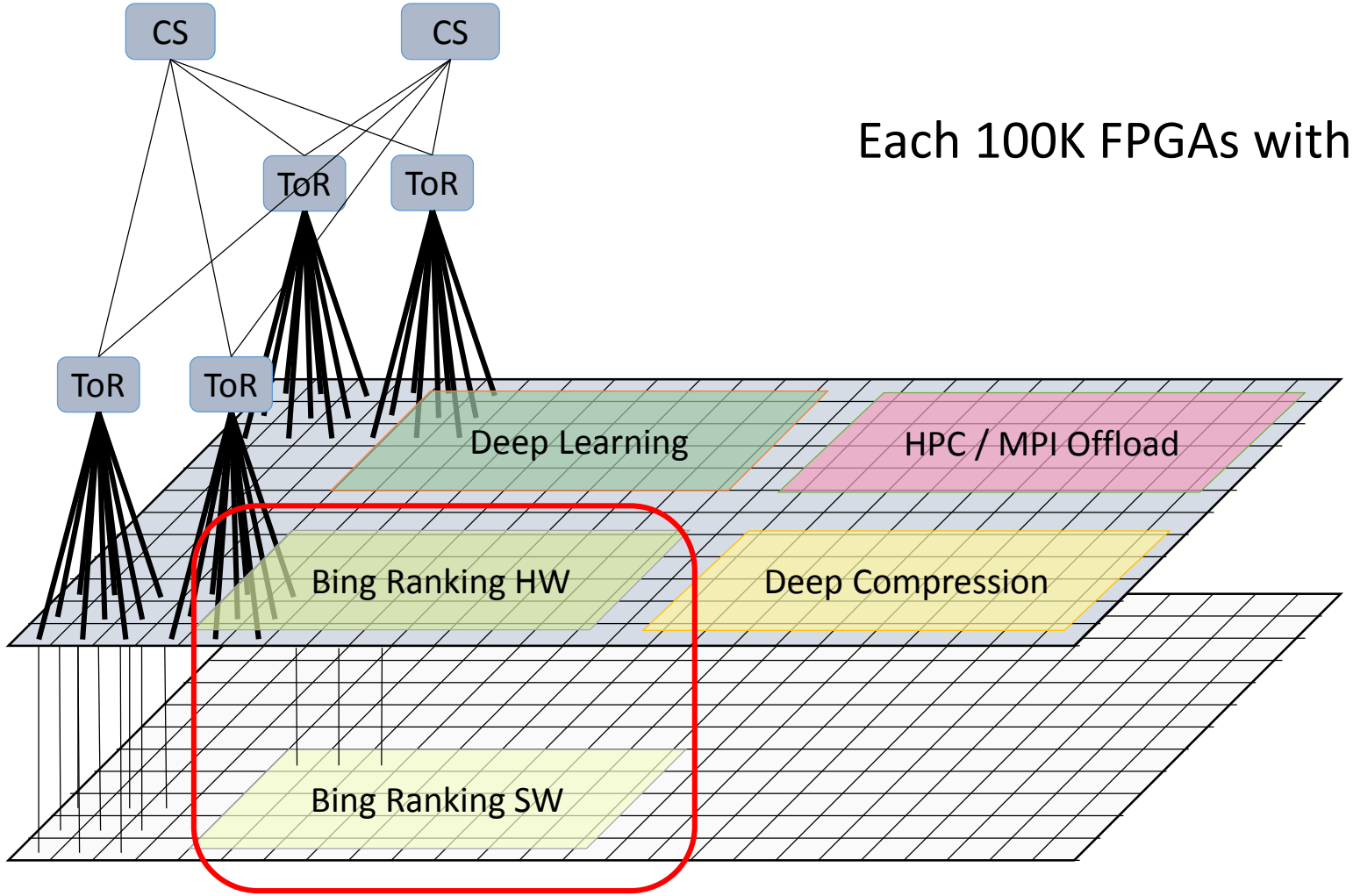
SmartNIC


Northbound API
Controller
Controller
Controller
Southbound API
SLB Decap
SLB NAT
VNET
ACL
Metering
Decap
DNAT
Encap
Allow
Meter
Rewrite
Transposition Engine
VFP
GFT Offload Engine
VM
First Packet
GFT Offload API (NDIS)
VMSwitch
SmartNIC
GFT
50G
Crypto
RDMA
QoS

FPGAs “versus” GPUs

	CPUs 	GPUs 	FPGAs 
Language	C/C++	CUDA	Verilog -> OpenCL (?)
Performance	400 Gflops	6 Tflops -> 10T	100G -> 1T -> 4T
Efficiency	5 Gflops/W	-> 20 Gflops/W	40-50 G/W -> 80-100 G/W
Scale	2M+ and growing	1s -> 10s -> 100s	10Ks -> 100Ks -> 1M+
DRAM BW	85 GB/s	2x240 GB/s	10GB/s -> 20GB/s -> 200-500GB/s

Large-Scale Reconfigurable Computing for HPC



Each 100K FPGAs with 1TFlop gives you 100 Pflops …

Programmable HW fabric

Programmable SW fabric

Conclusions

- We are at the dawn of a new era
- Programmable logic playing a central role in systems at massive scale
- “A new kind of computer”
- Will enable new applications and services to be cost effective
- Will change system architecture, both in server and at cloud scale