FPGAs in the Cloud

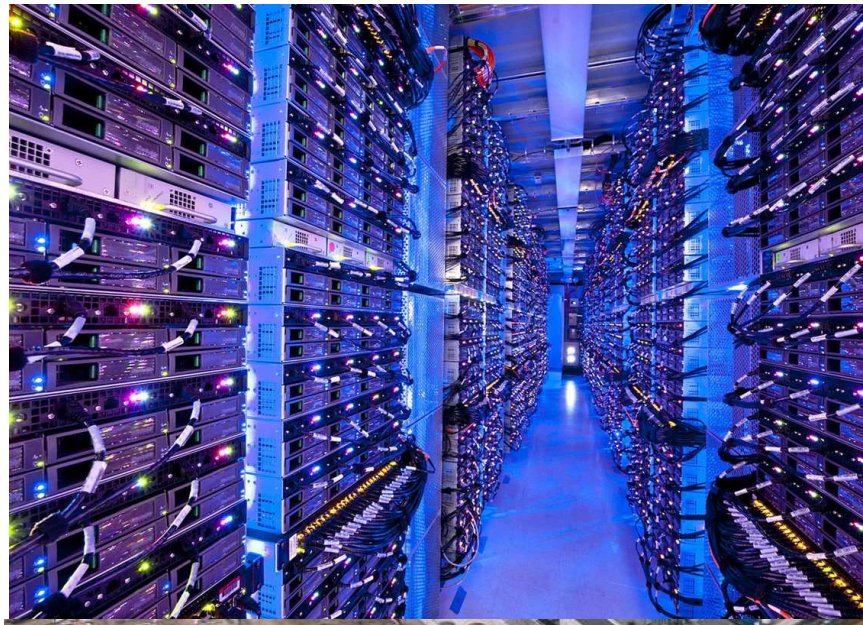
Derek Chiou

Microsoft and

The University of Texas at Austin

Today's Data Centers

- O(100K) servers/data center
- Very dense, maximize number of servers
- Tens of MegaWatts
- Strict power and cooling requirements
- Secure, hot, noisy
- Incrementally upgraded
  - 3 year server depreciation, upgraded quarterly



Microsoft Cloud Services

Microsoft Azure

bing

XBOX LIVE

COSMOS

skype

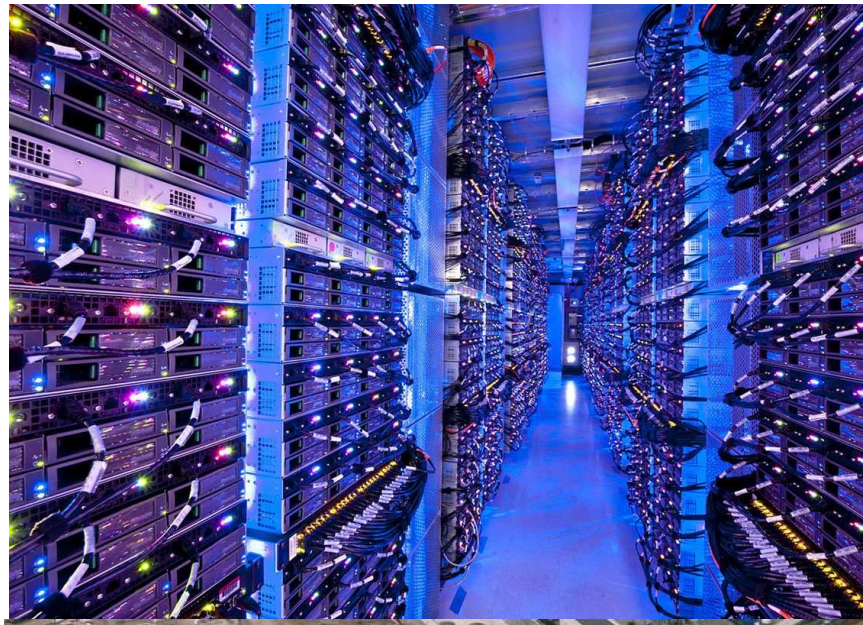
msn

OneDrive

Outlook.com

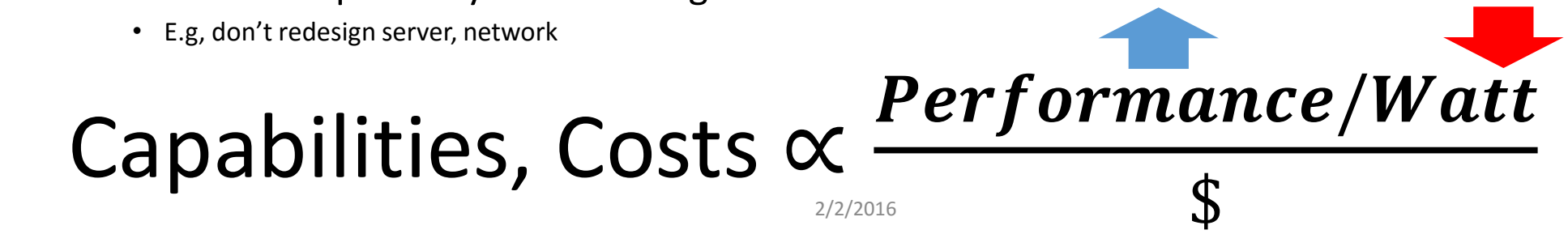
Microsoft Office 365

Microsoft SQL Server



Challenges

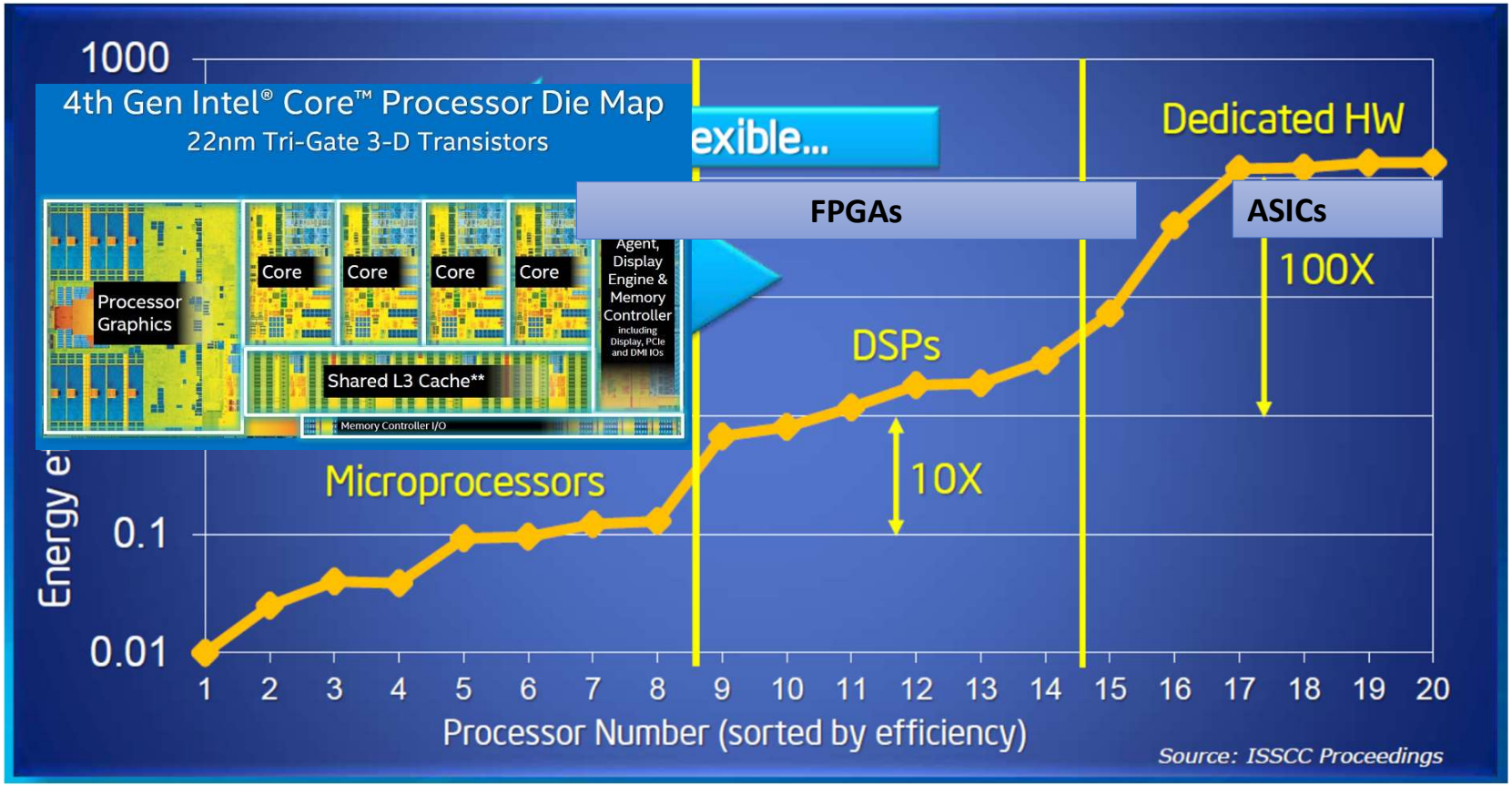
- Efficiency is the ***essence*** of data centers
  - Shared servers, management, high volume energy more efficient
- Thus, data centers are high volume, low margin
  - At data center scales, even small savings add up
  - Don't buy what you don't need, use everything you buy (fungible hardware)
- Software services change very rapidly (monthly)
- Machines switch services over time
  - E.g., new machines for latency bound, old machines for throughput bound
- Little/no HW maintenance, accessibility (cannot upgrade hardware midlife)
- Maintain compatibility with existing infrastructure
  - E.g, don't redesign server, network



Goal: Improve Data Center Efficiency

- Traditional method was to add cores
- However, Dennard scaling ended, Moore's Law almost ended
  - Energy used now ~ proportional to number of transistors switching
  - Single thread CPU efficiency leveled off
  - Cost per transistor becoming constant
- More real work per transistor switch to improve efficiency

Efficiency via Specialization



Source: Bob Broderson, Berkeley Wireless group

Application Specific Integrated Circuit vs Field Programmable Gate Array

- ASIC
  - Hardwired lc
    better than F
  - Full 10x be
  - Specific to
  - Unchangak
  - O(months) code => chips
  - O($50M) NRE requires high volumes
    - unlikely to use latest technology node
  - Potential to hold back algorithms

ASICs aren't compatible with our requirement to enable acceleration of many applications, are more costly, increase complexity to add flexibility, perform verification

ic, dual ported
e routing => ~10x worse
time

ea/performance
uit
onality/bugs

- O(hours) code => chips
- FPGAs have high volumes
  - First parts out on a technology node
- Encourages algorithm exploration

Our Design Requirements

Don't Cost Too Much

<30% Cost of Current Servers

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

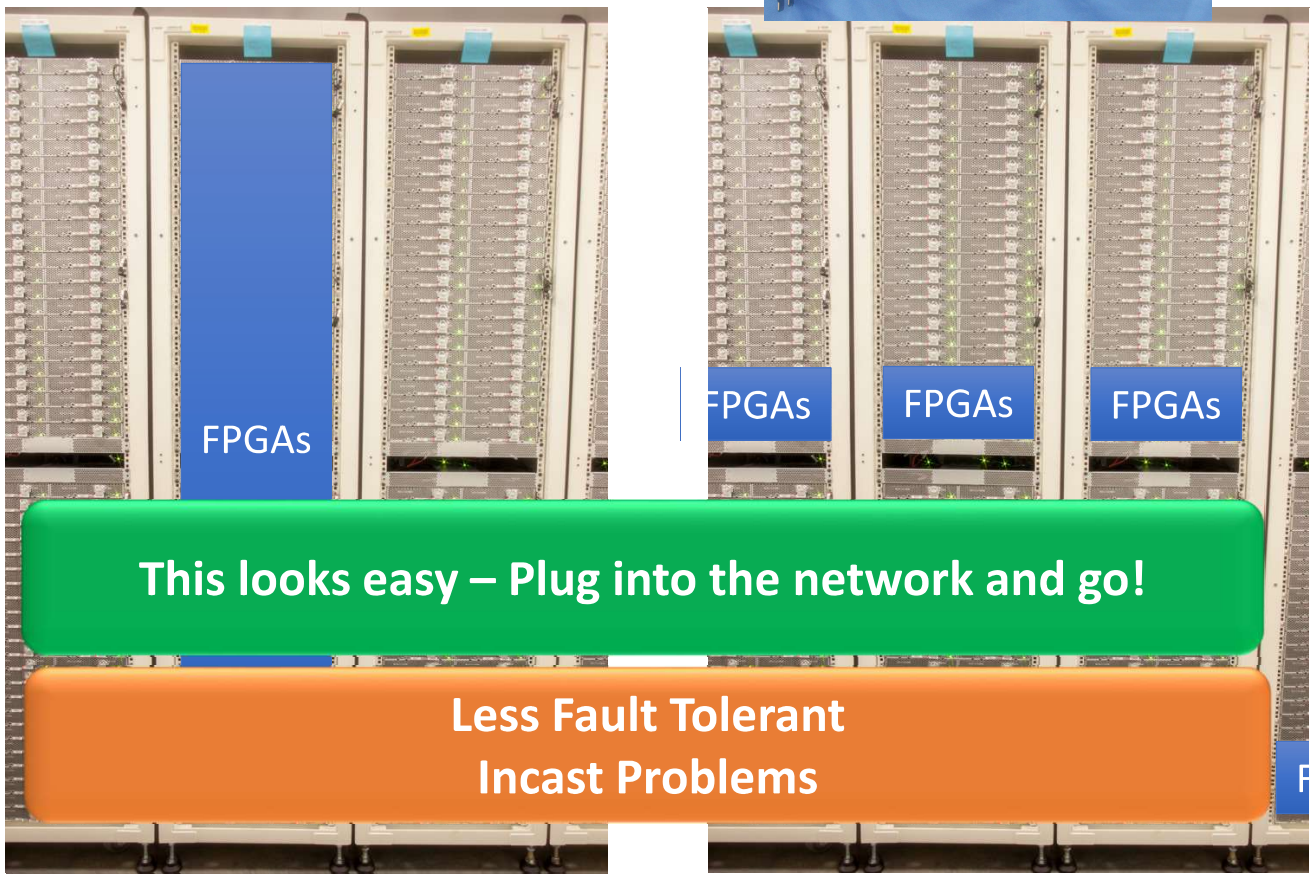
Don't Burn Too Much Power

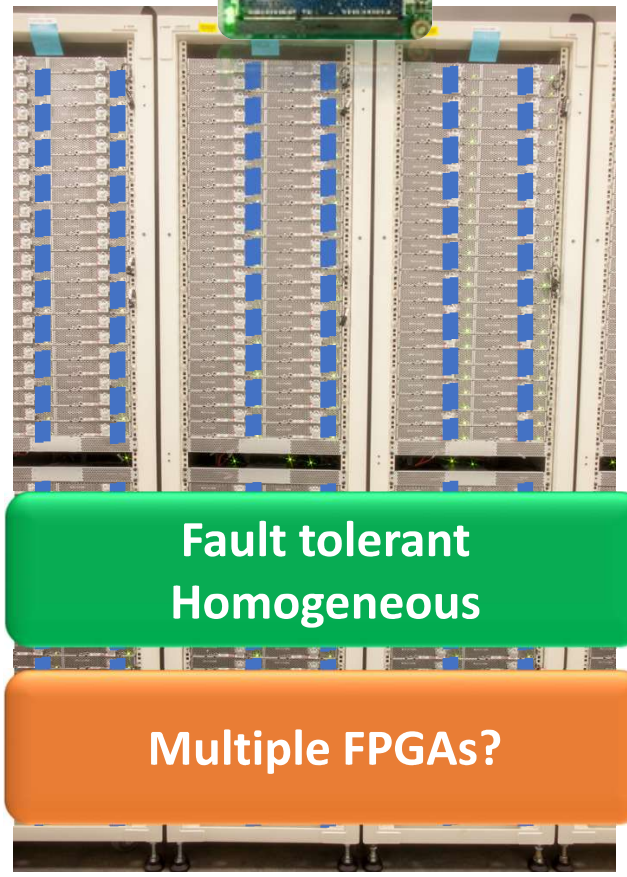
<10% Power Draw
(25W max, all from PCIe)

Don't Break Anything

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

Configuration





This looks easy – Plug into the network and go!

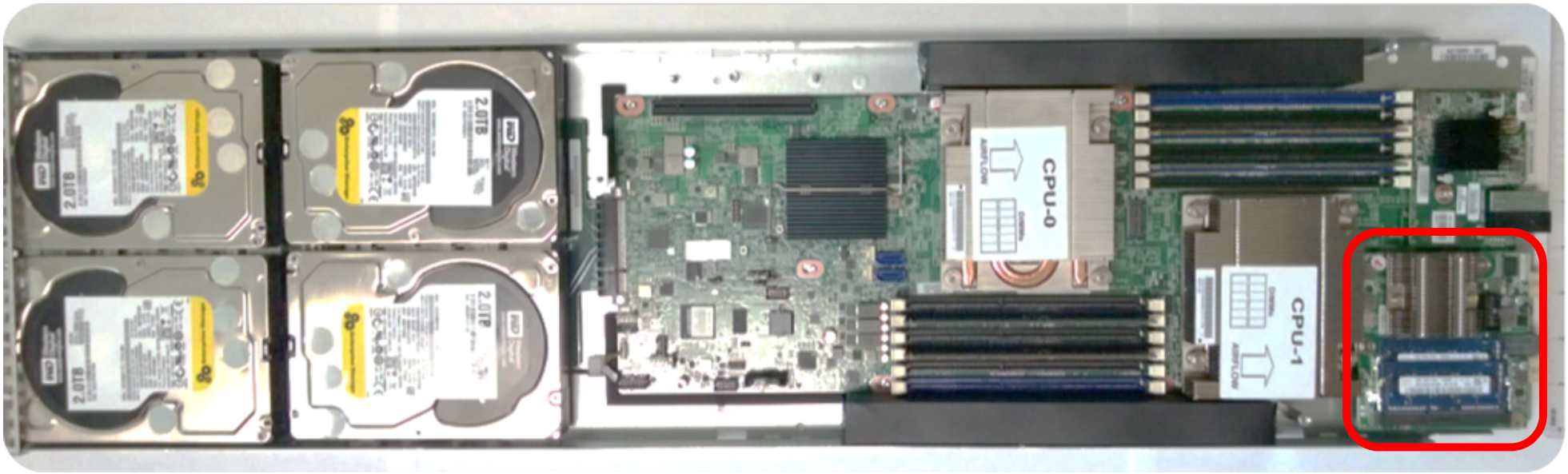
Less Fault Tolerant
Incast Problems

Fault tolerant
Homogeneous

Multiple FPGAs?

Centralized **Distributed**

~2013 Microsoft Open Compute Server



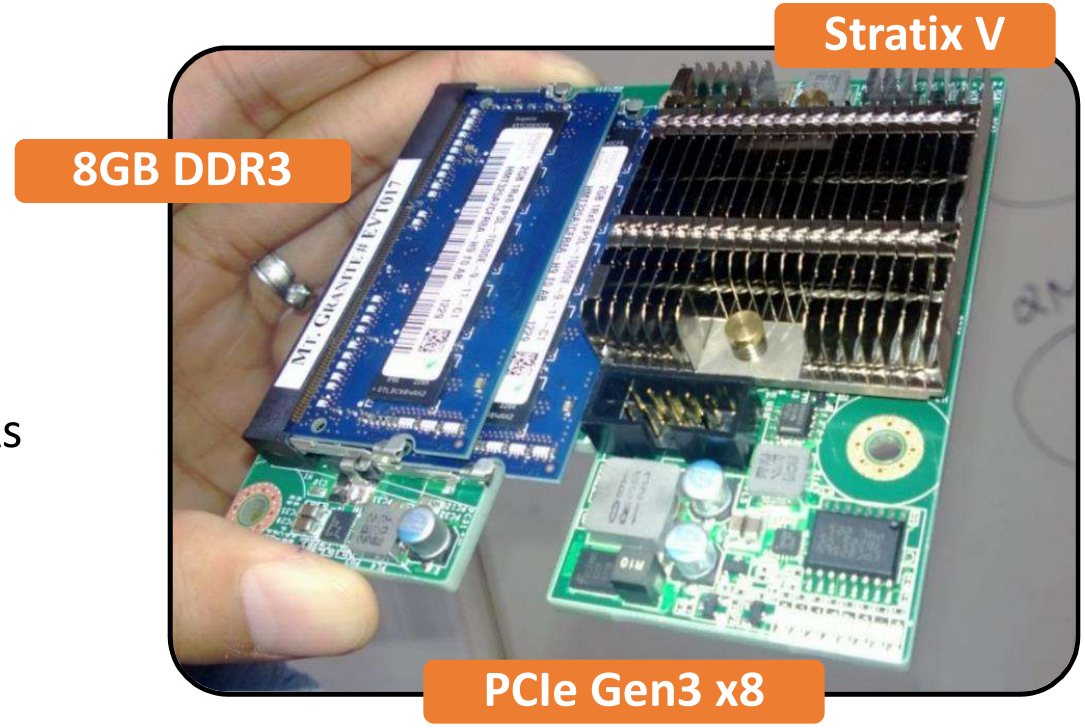
Two 8-core Xeon 2.1 GHz CPUs
64 GB DRAM
4 HDDs, 2 SSDs
10 Gb Ethernet

Air flow

Catapult V1 Accelerator Card

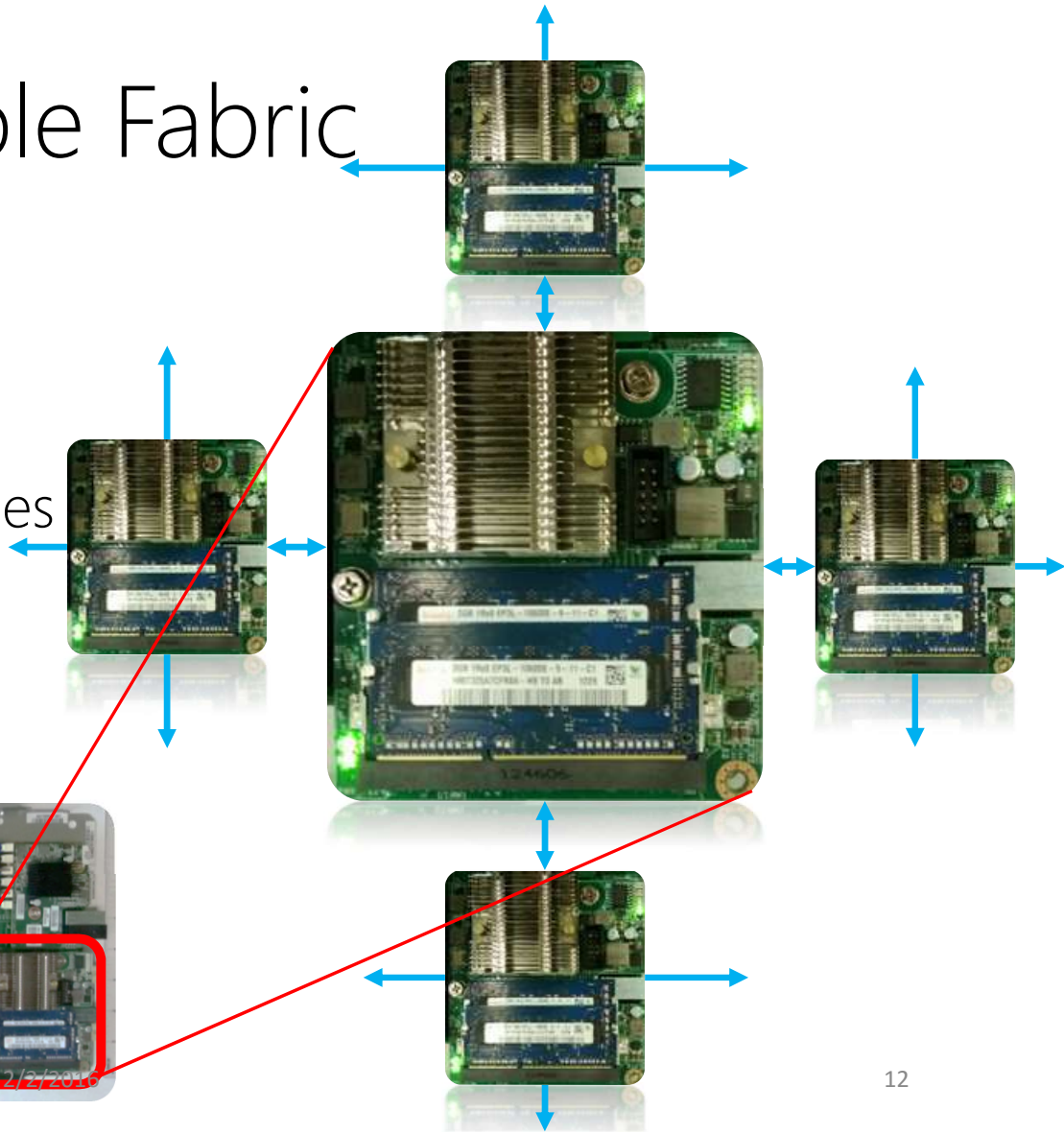
- Altera Stratix V D5
- 172.6K ALMs, 2014 M20Ks
  - 457KLEs
  - 1 KLE == ~12K gates
  - M20K is a 2.5KB SRAM
- PCIe Gen 2 x8, 8GB DDR3
- 20 Gb network among FPGAs

- Single FPGA too small?

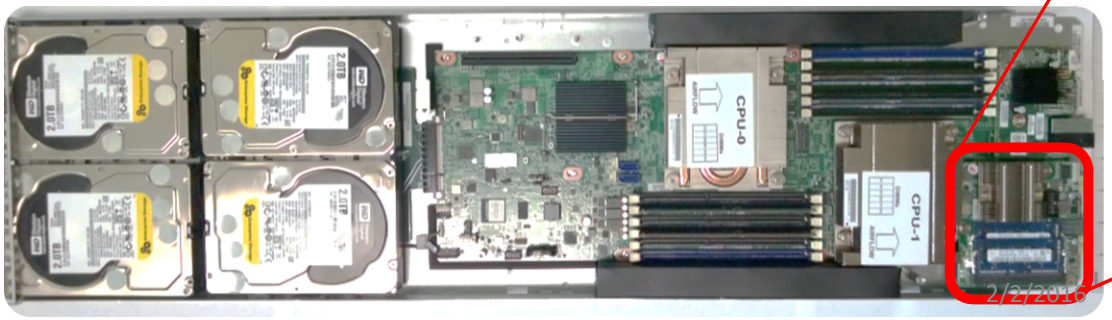


Scalable Reconfigurable Fabric

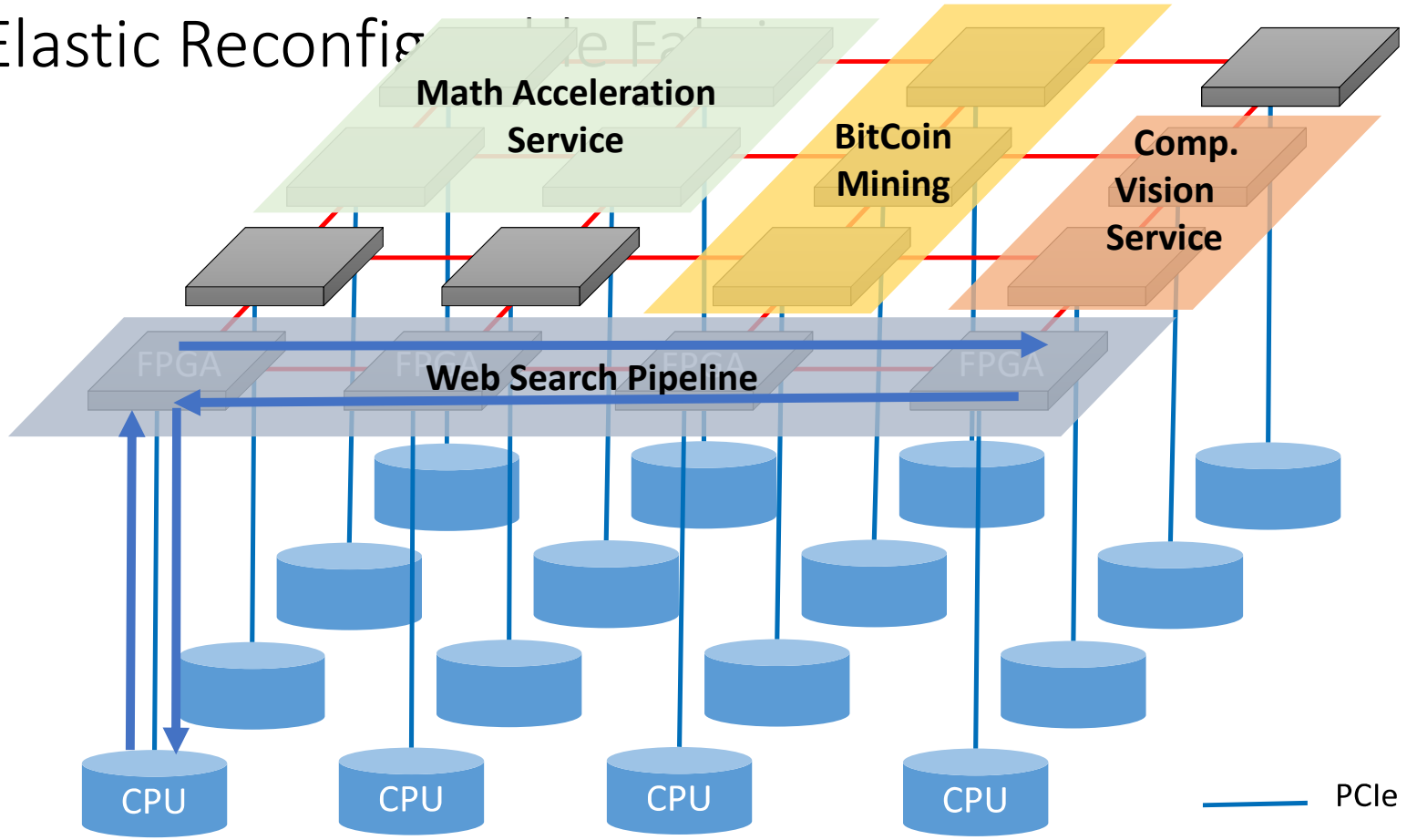
- 1 FPGA board per Server
- 48 Servers per ½ Rack
- 6x8 Torus Network among FPGAs
  - 12.5Gb over SAS SFF-8088 cables



Data Center Server (1U, ½ width)

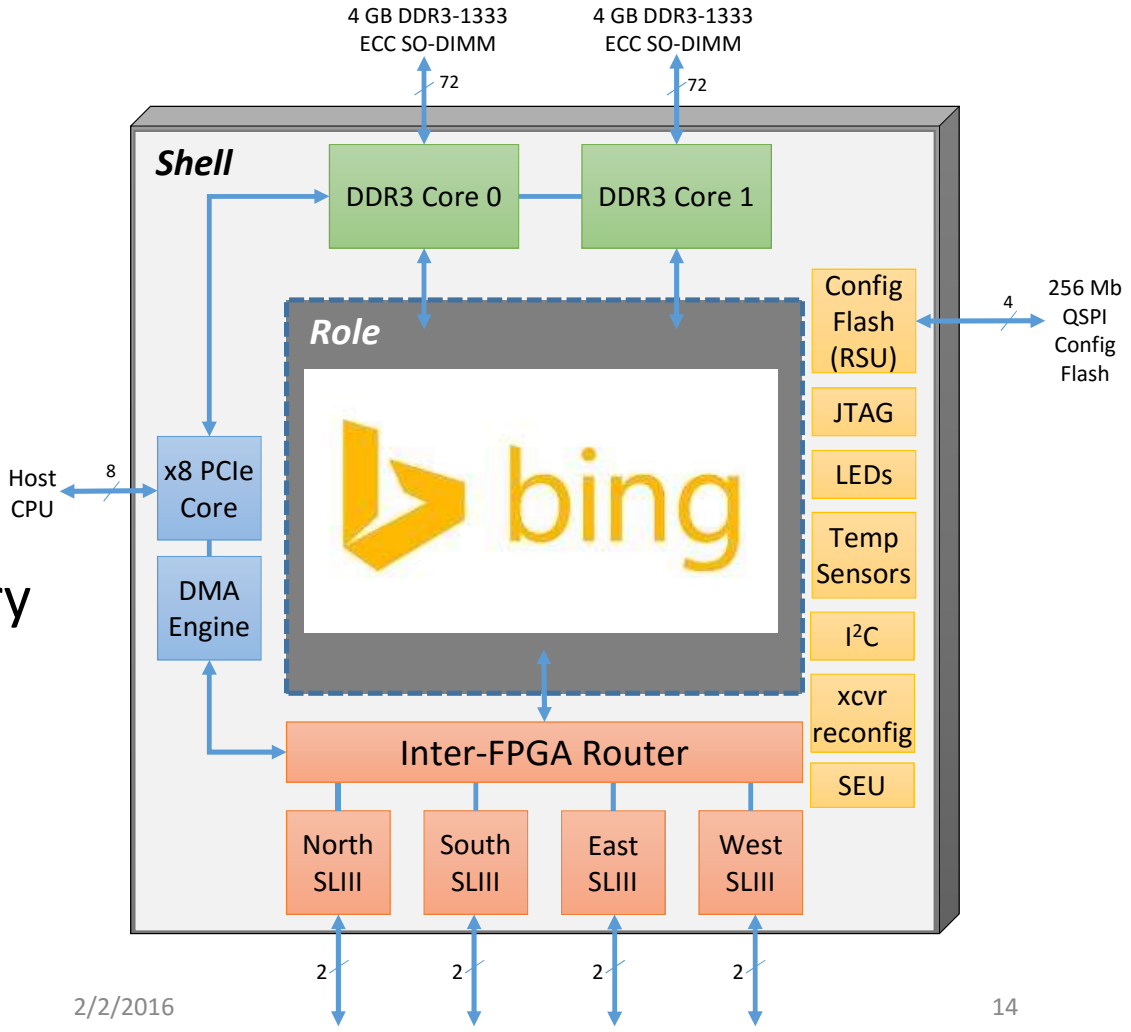


An Elastic Reconfig



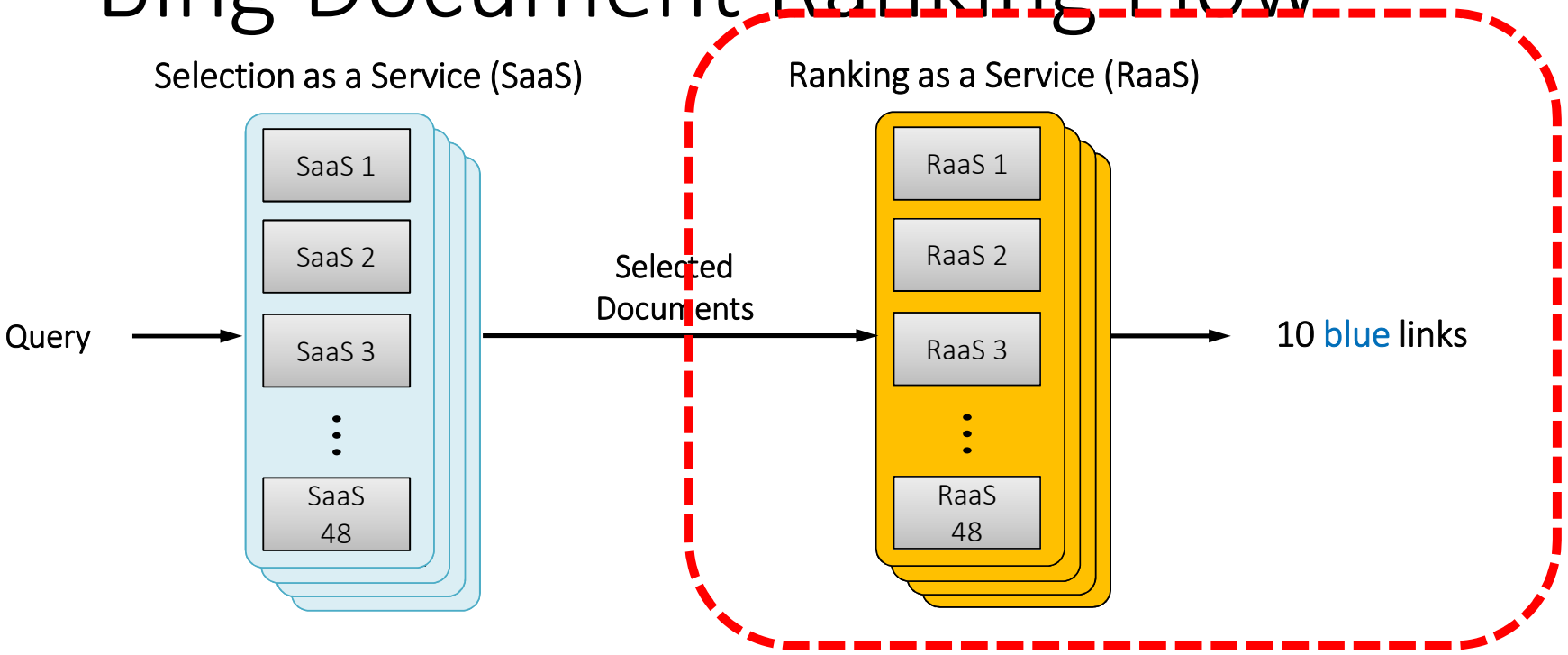
Shell & Role

- *Shell* handles all I/O & management tasks
- *Role* is only application logic
- FIFO access to Shell
- Role is Partial Reconfig boundary



Bing

Bing Document Ranking Flow



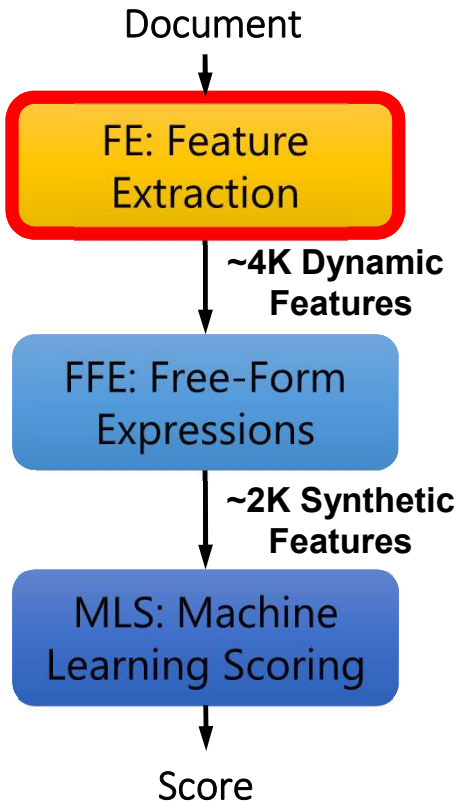
Selection-as-a-Service (SaaS)

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

Ranking-as-a-Service (RaaS)

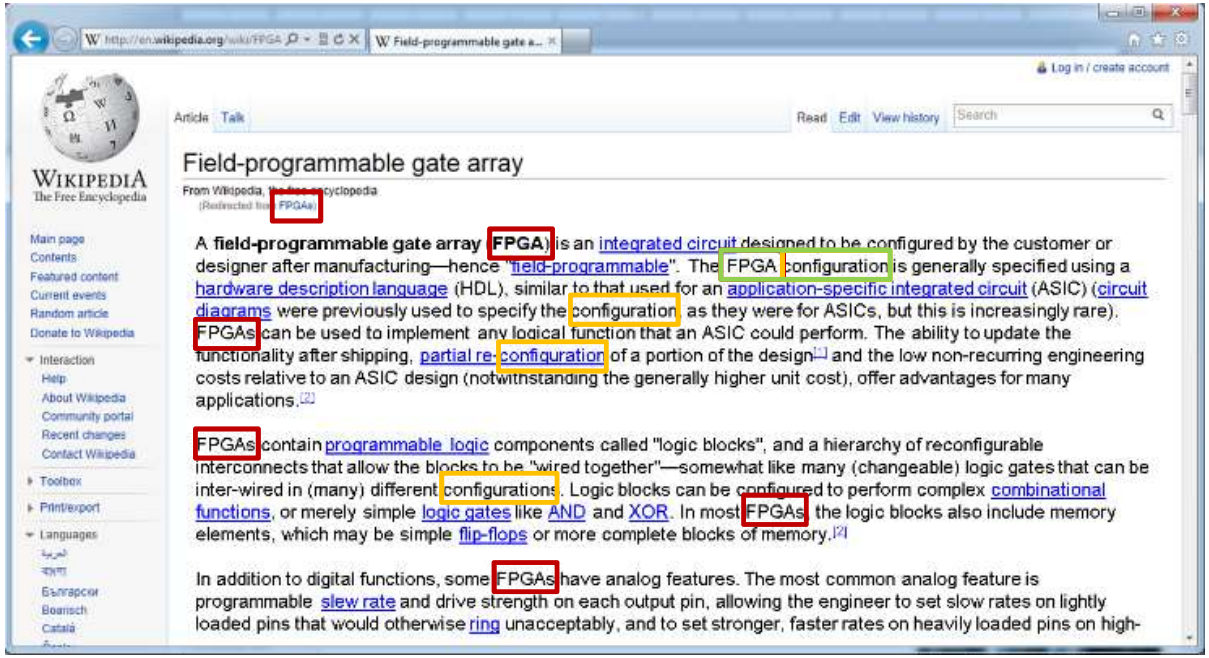
- Compute scores for how relevant each selected document is for the search query
- Sort the scores and return the results

FE: Feature Extraction

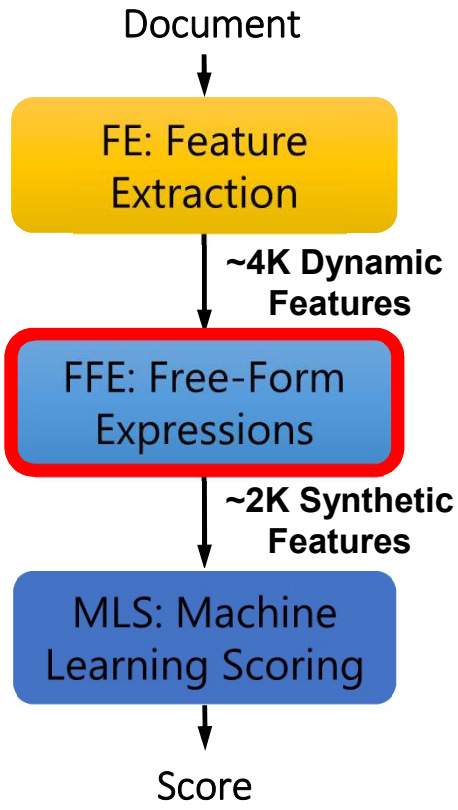


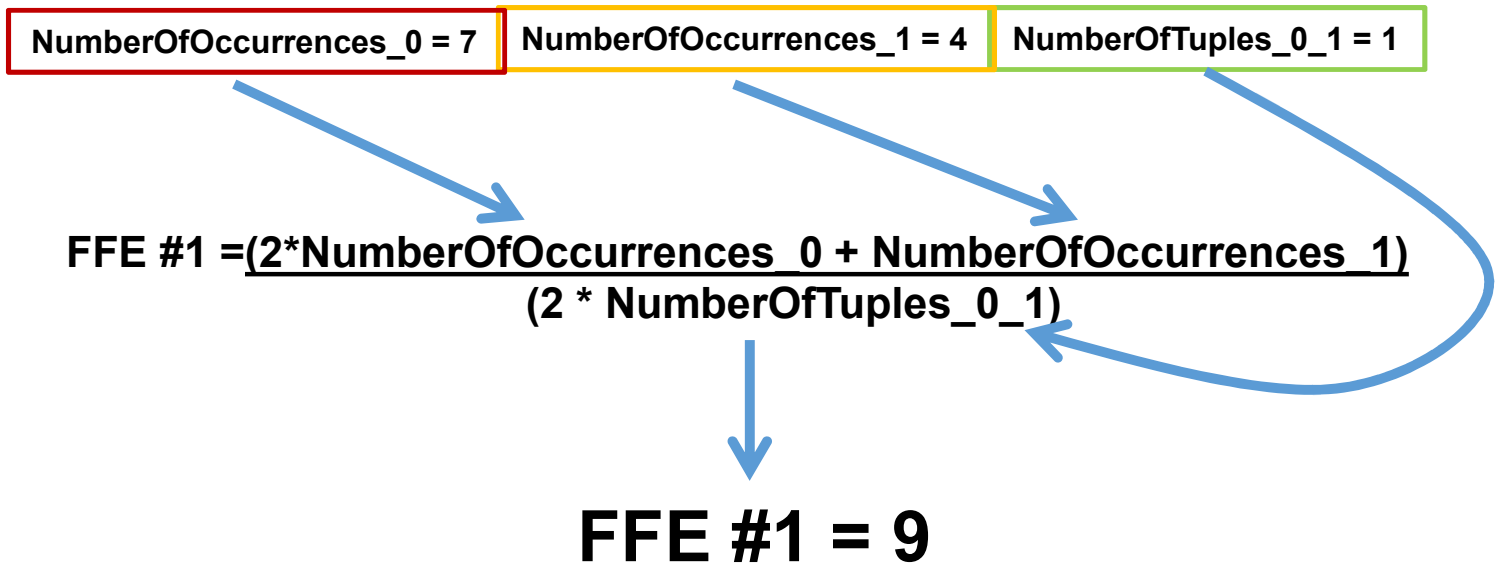
Query: "FPGA Configuration"

NumberOfOccurrences_0 = 7	NumberOfOccurrences_1 = 4	NumberOfTuples_0_1 = 1

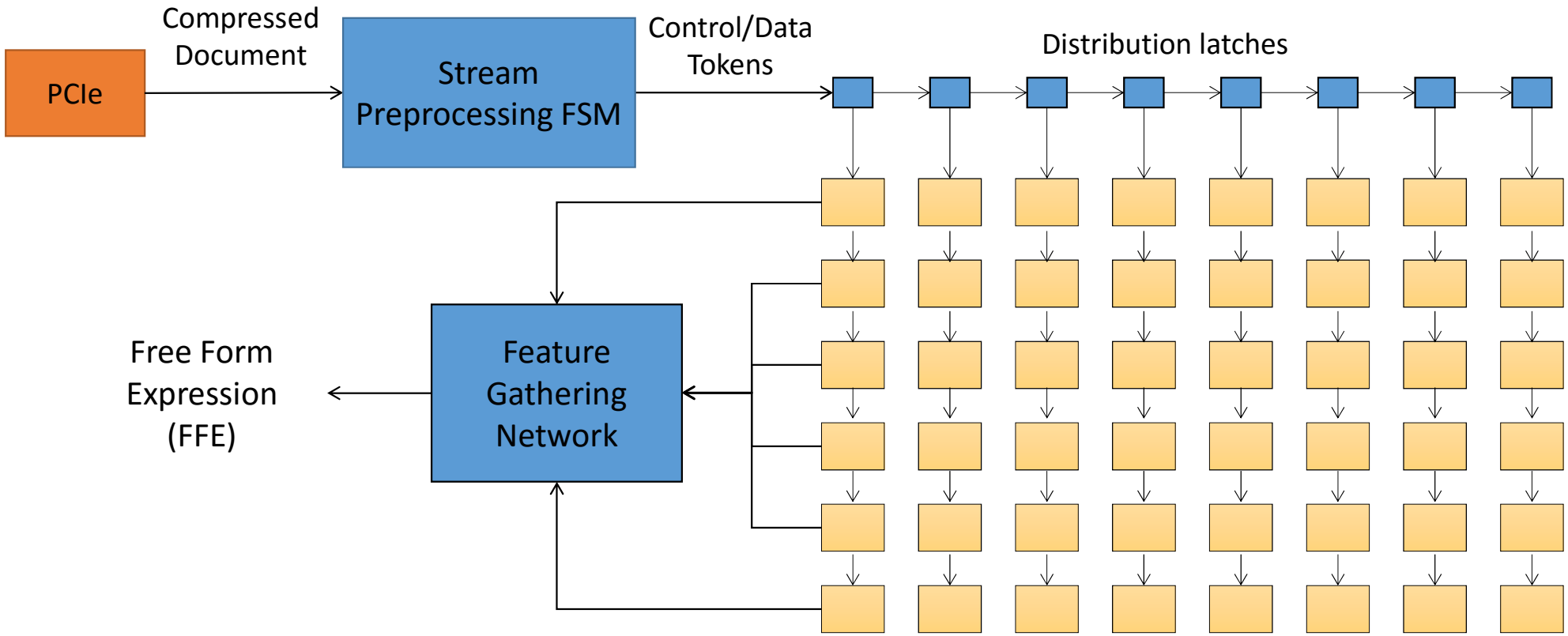


FFE: Free Form Expressions



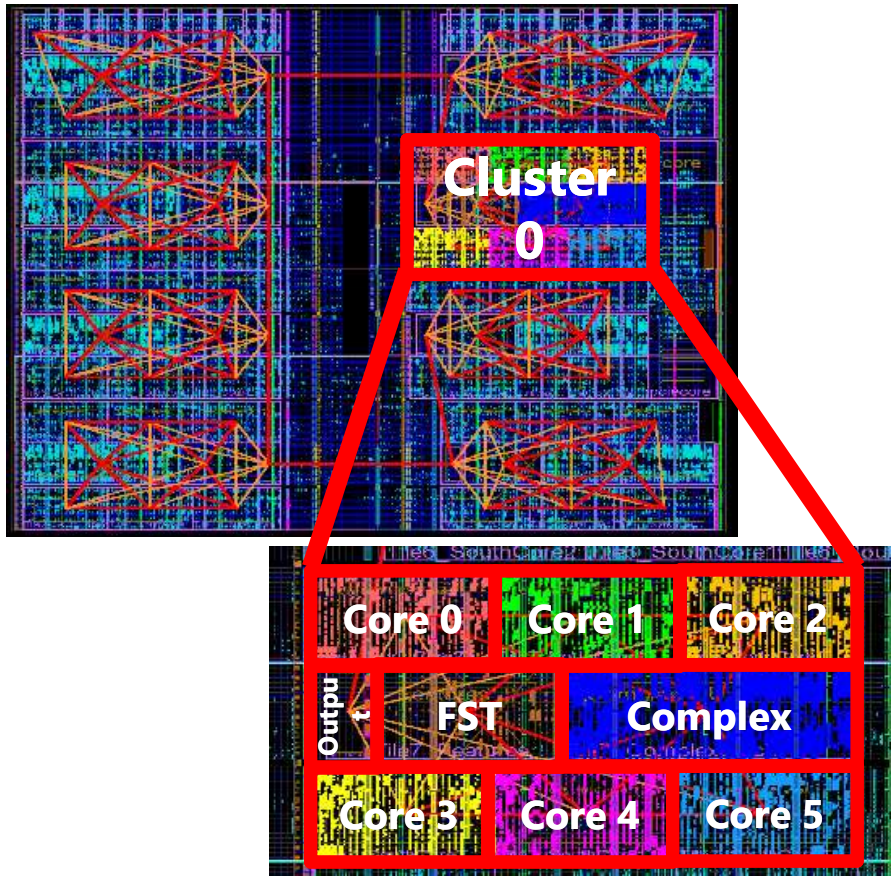


Feature Extraction Accelerator

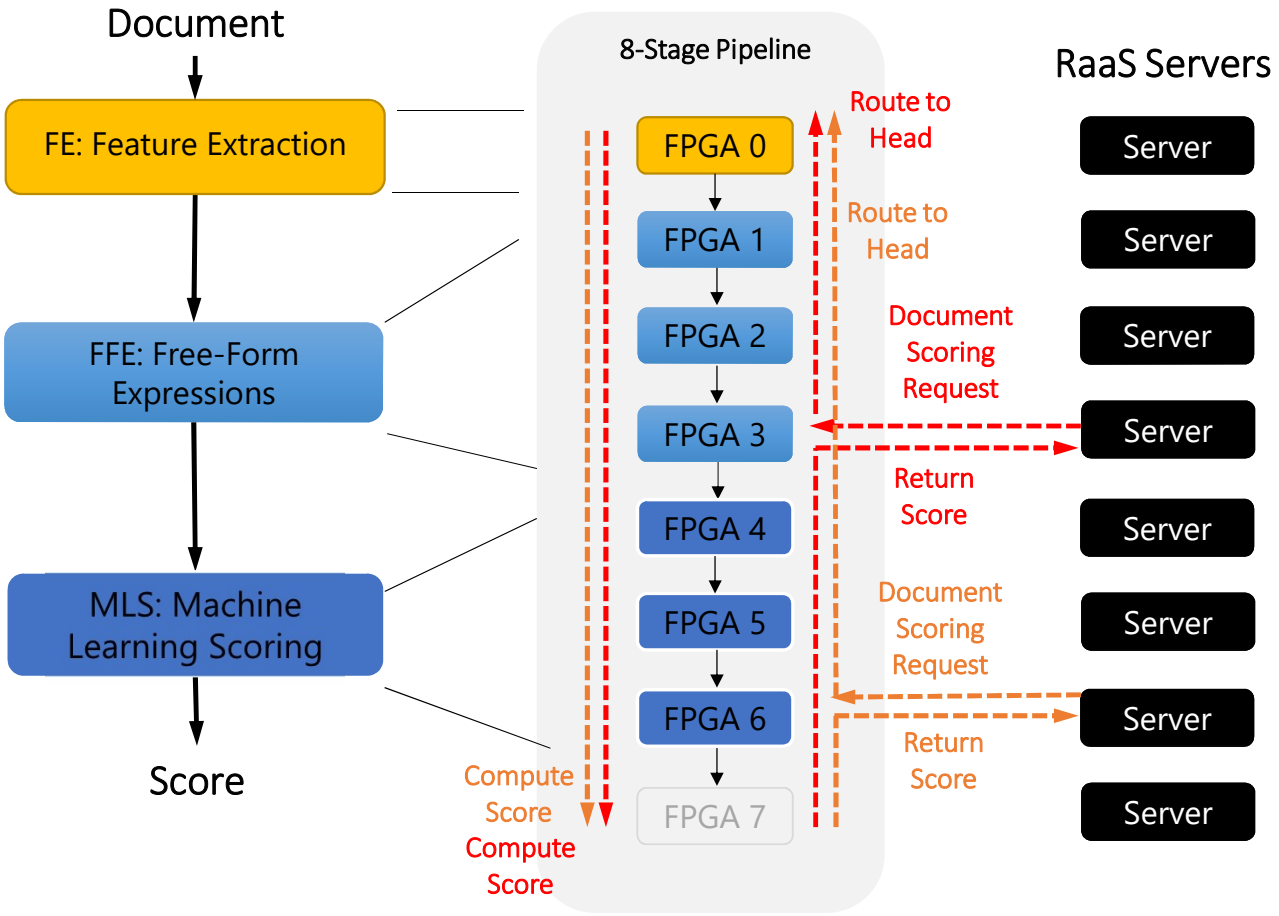


FFE Engines

- Softcore for multi-threaded throughput
- 4 HW threads per core
- 6 cores share a complex ALU
  - log, divide, exp, float/int conversions
- 10 clusters (240 HW threads) per FPGA



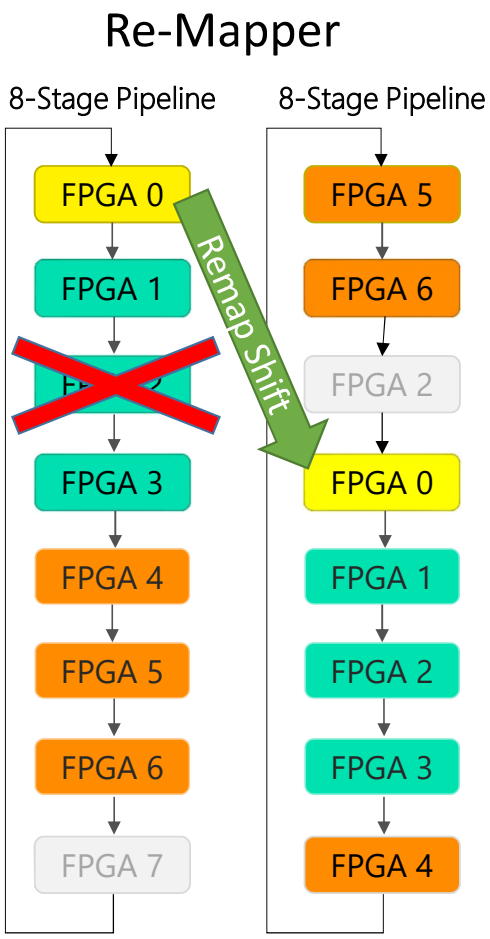
FPGA Accelerator for RaaS

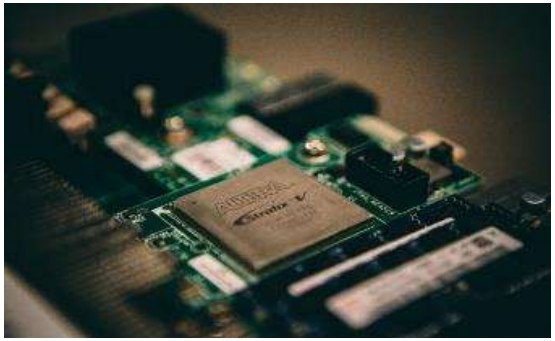


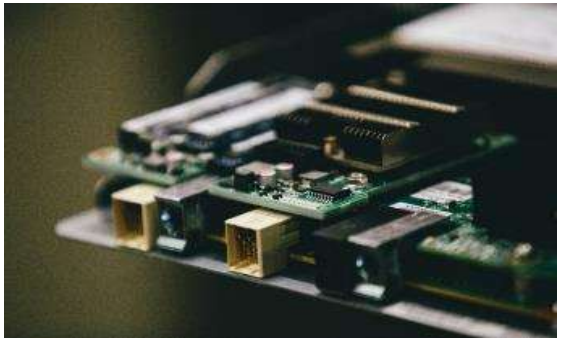
Scalable Deployment Challenges

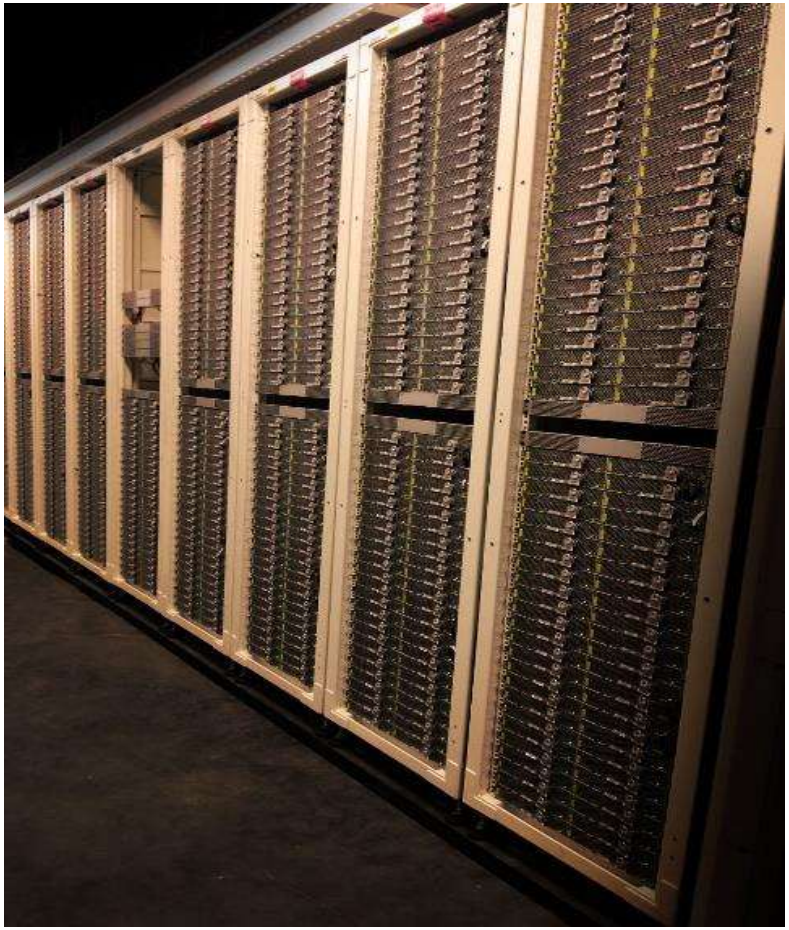
- Issues with Spanning Multiple FPGAs
- Health monitor to detect stalled pipelines
- Reconfiguration protocol to remove lockups
- Re-mapper shifts images on machine failure

- General Issues with an FPGA Fabric
- PCIe driver tuning
- SEU scrubbing of the FPGA
- Wiring and board check at integration
- Bringing down and bringing up links





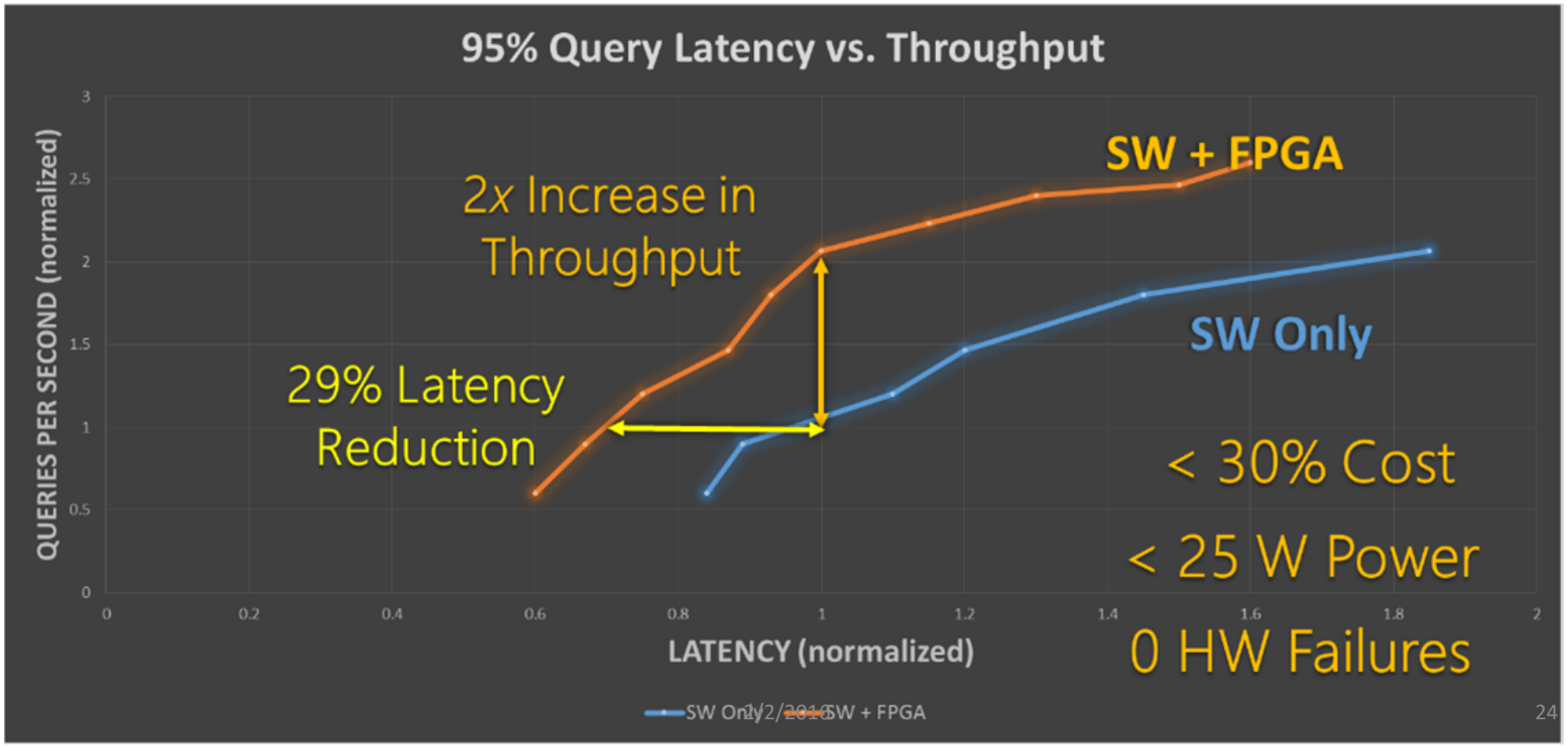




1,632 server pilot deployed in production BN datacenter

Bing Search Pilot Results

1,632 Servers with FPGAs Running Bing Page Ranking Service (~30,000 lines of C++)



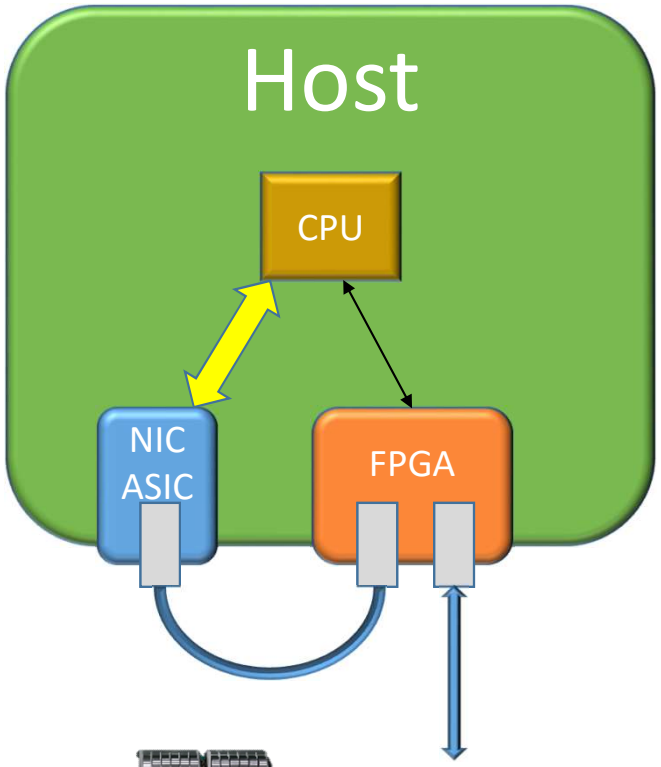
Production issues at scale

- Build system
  - License servers, availability of source, build machines
- Shell/driver/application versioning and deployment
  - Backwards compatibility
- Health monitoring and failure diagnostics
- Debugging/telemetry (esp. on livesite)
- System integrity testing
  - Many servers/vendors
- Verification/validation
- *In situ* updates to drivers, golden image, shell
- Supply chain

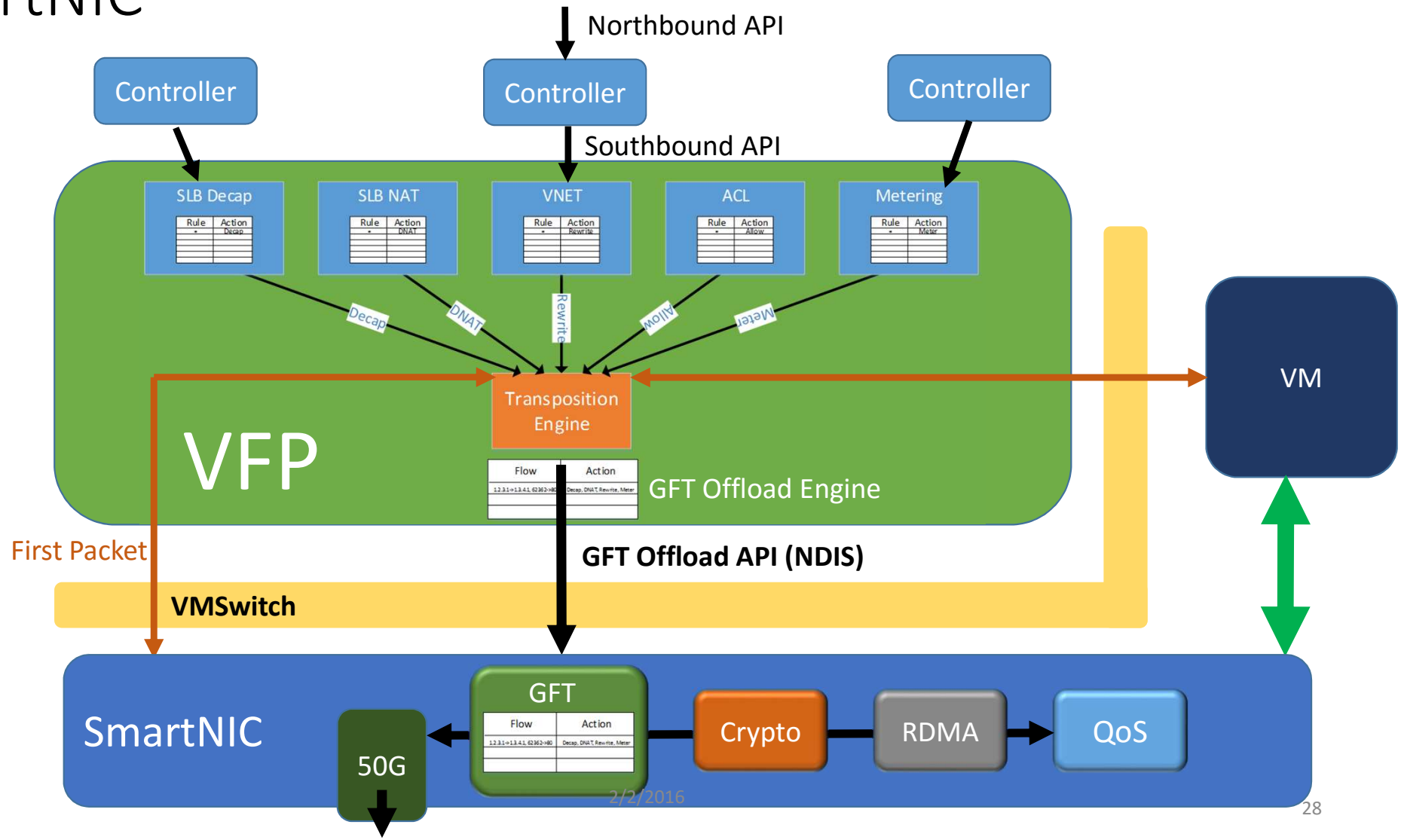
Azure Networking

Azure SmartNIC

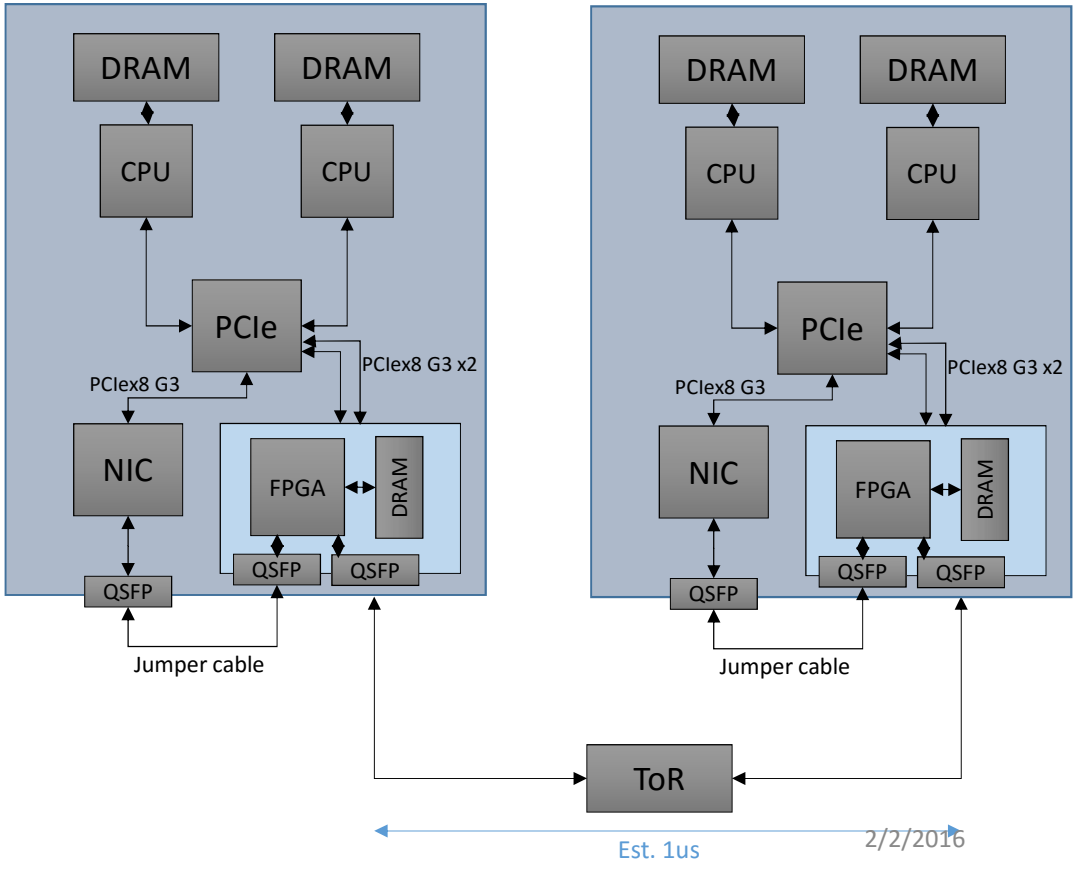
- Use an FPGA for reconfigurable functions
  - FPGAs are already used in Bing (Catapult)
  - Roll out hardware as we do software
- Programmed using Generic Flow Tables (GFT)
  - Language for programming SDN to hardware
  - Uses connections and structured actions as primitives
- SmartNIC can also do Crypto,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



SmartNIC

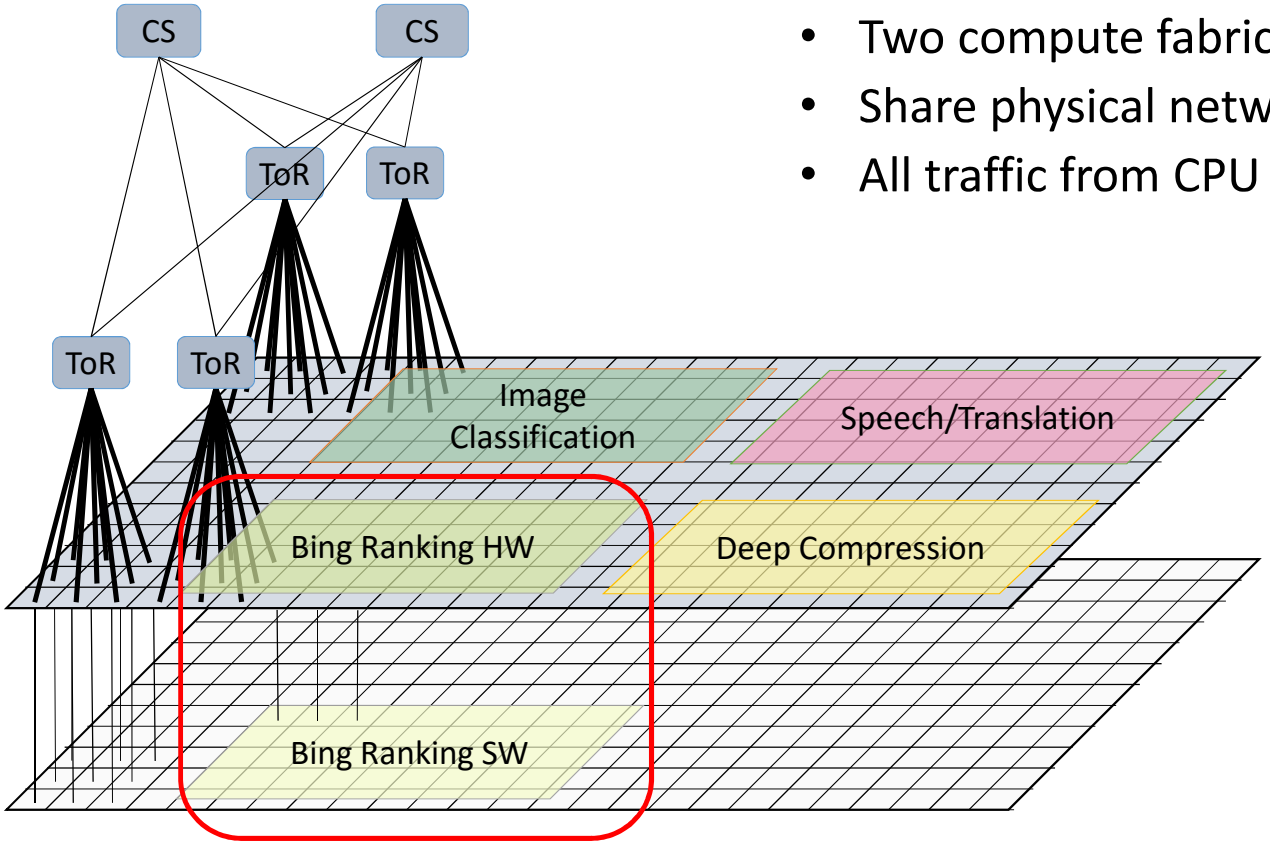

Northbound API
Controller
Controller
Controller
Southbound API
SLB Decap
SLB NAT
VNET
ACL
Metering
Rule
Action
Decap
DNAT
Rewrite
Allow
Meter
Transposition Engine
VFP
Flow
Action
GFT Offload Engine
VM
First Packet
GFT Offload API (NDIS)
VMSwitch
SmartNIC
GFT
Flow
Action
50G
Crypto
RDMA
QoS

Catapult V2



- Microsoft scaling to large volume
- FPGAs can talk directly to one another
  - No software in the loop
  - No system memory in the loop

A Parallel Hardware Acceleration Fabric



- Two compute fabrics
- Share physical network
- All traffic from CPU goes through FPGA

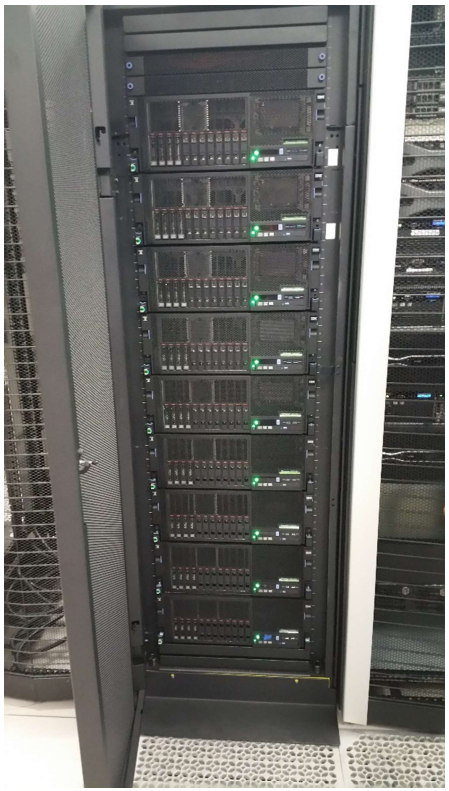
Programmable HW fabric

Programmable SW fabric

FPGA Accelerator Research Infrastructure Cloud (FAbRIC)

Goals and Deployment

- Provide high end/high scale FPGA platforms for open research
  - FPGA systems themselves
  - CAD tools and servers to run on
- Place these systems in The University of Texas supercomputer center
- Enables
  - Amortize overhead of acquisition, installation, running of FPGAs systems
  - Reproducibility of results
  - Remote classes
- Current systems
  - Convey MX-100 (100GB/sec)
  - 9 Power8+CAPI+Xilinx+Altera+Nvidia
- Openfabric.org for instructions on how to get access



Catapult Academic Program

- Jointly funded by Altera and Microsoft
  - Some system administration/tools servers funded by NSF under FAbRIC
- Provide
  - PCIe device driver, shell (initially compiled/encrypted, discussing source access under NDA with lawyers)
  - “Hello, world!” programs
  - Altera tools (including OpenCL), servers to run them on
  - Individual V1 boards sent to you
  - Remote access to 6*48 servers each with V1 board
  - Accessible with one page proposal to catapult@microsoft.com
  - Get a FAbRIC account for tools
- See research.microsoft.com/catapult for details

Conclusions

- We are at the dawn of a new era where programmable logic is playing a central role in systems at massive scale
- “A new kind of computer”
- Deploy hardware before conceiving of application
  - Everything from requirements, to architecture, to implementation, testing, and deployment become much simpler
  - Codesign hardware/software system architecture at server, rack, data center level
- Actively engaging with academia to build an eco-system around reconfigurable data centers
- ***We are hiring!***