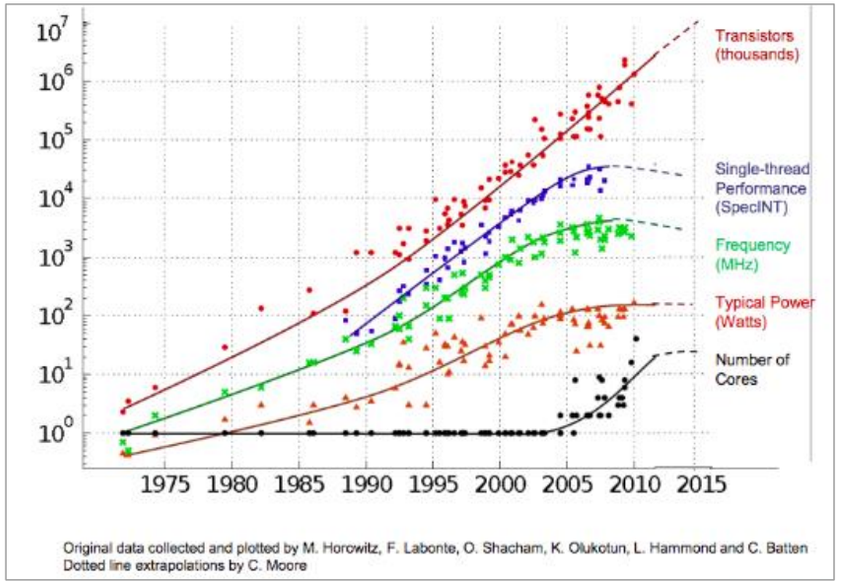
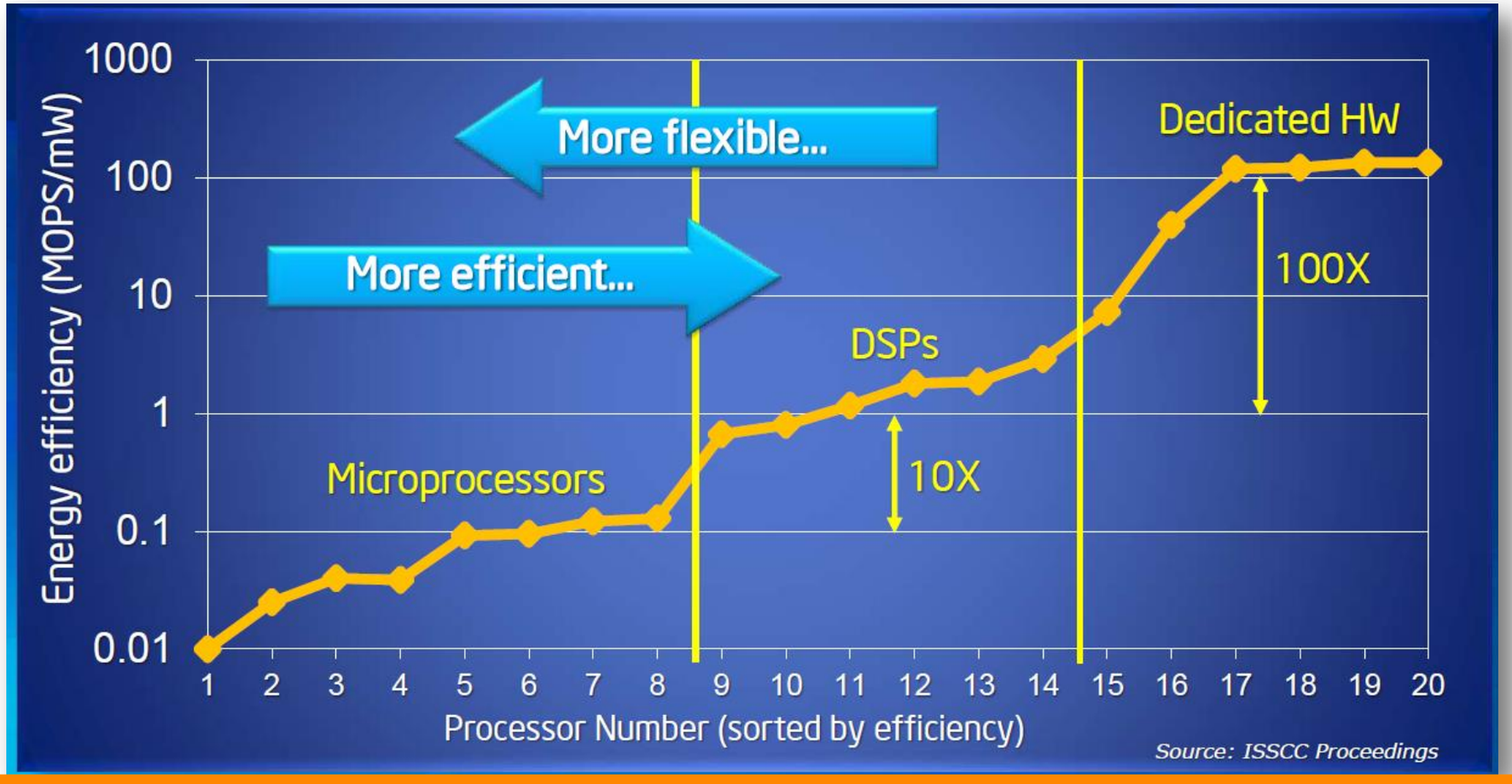

IMAGES VIDEOS MAPS NEWS SEARCH HISTORY MORE MSN OUTLOOK.COM
Make Bing my homepage 2 250 Sign in
bing
Large-Scale Reconfigurable Computing in a Microsoft Datacenter
IPR2018-01694
EXHIBIT
2065
Andrew Putnam – Microsoft
Hot Chips 26 – Aug 12, 2014
© Image Credits
© 2014 Microsoft | Privacy and Cookies | Legal | Advertise | About our ads | Help | Feedback

Microsoft Cloud Services

Windows Azure · bing · Microsoft SQL Server · XBOX LIVE · COSMOS · skype · msn · OneDrive · Outlook.com · Microsoft Office 365



$$\text{Capabilities, Costs} \propto \frac{Performance/Watt}{\$}$$



Increase Efficiency with Hardware Specialization

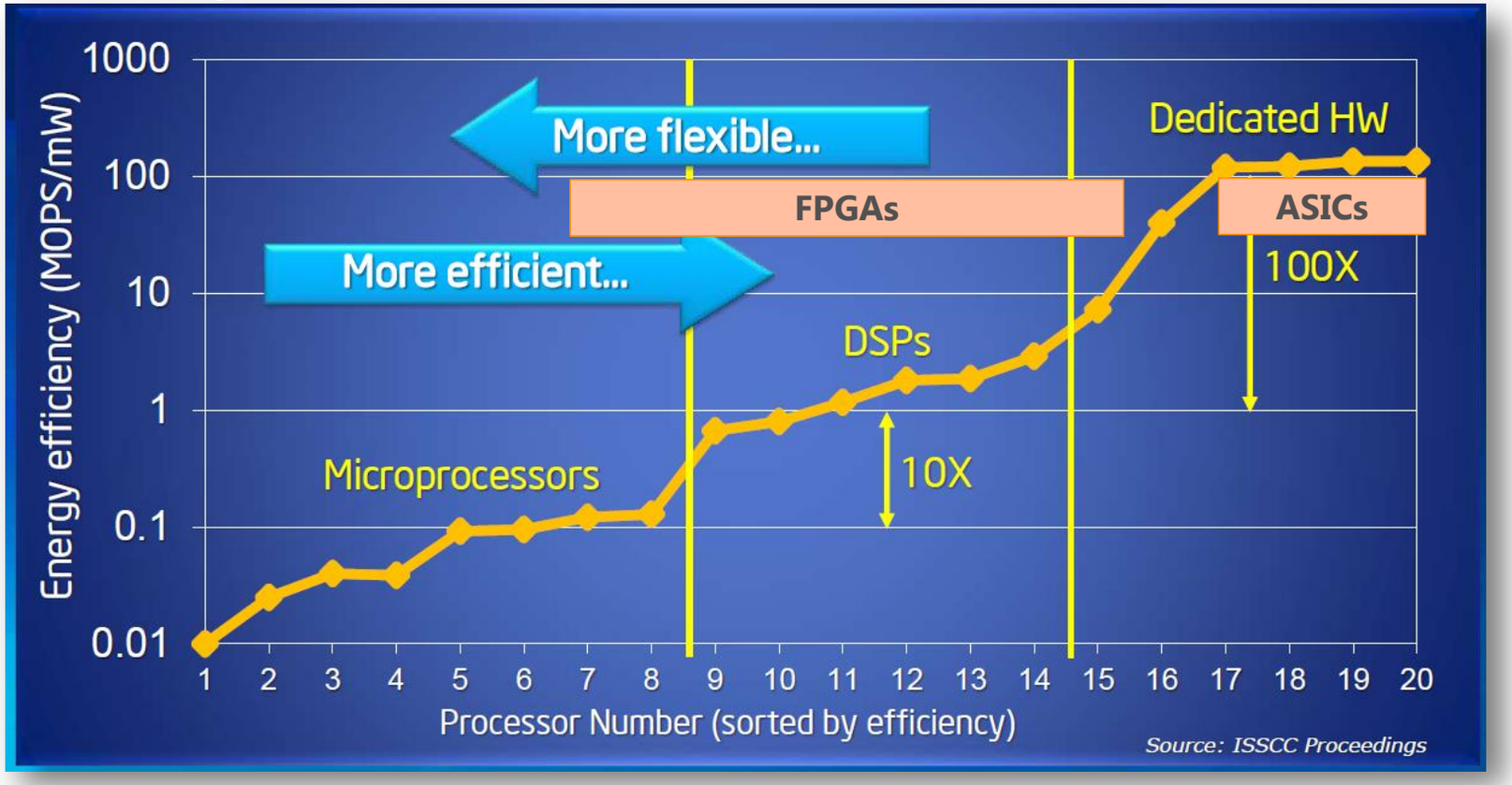
Datacenter Environment

- Software services change monthly
- Machines last 3 years, purchased on a rolling basis
- Machines repurposed ~½ way into lifecycle
- Little/no HW maintenance, no accessibility

- Homogeneity is highly desirable

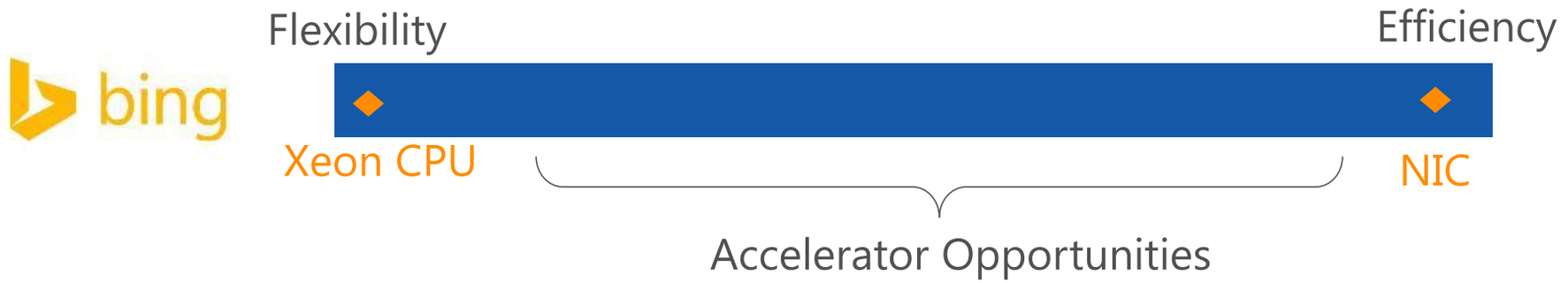
The paradox: Specialization *and* homogeneity

Efficiency via Specialization

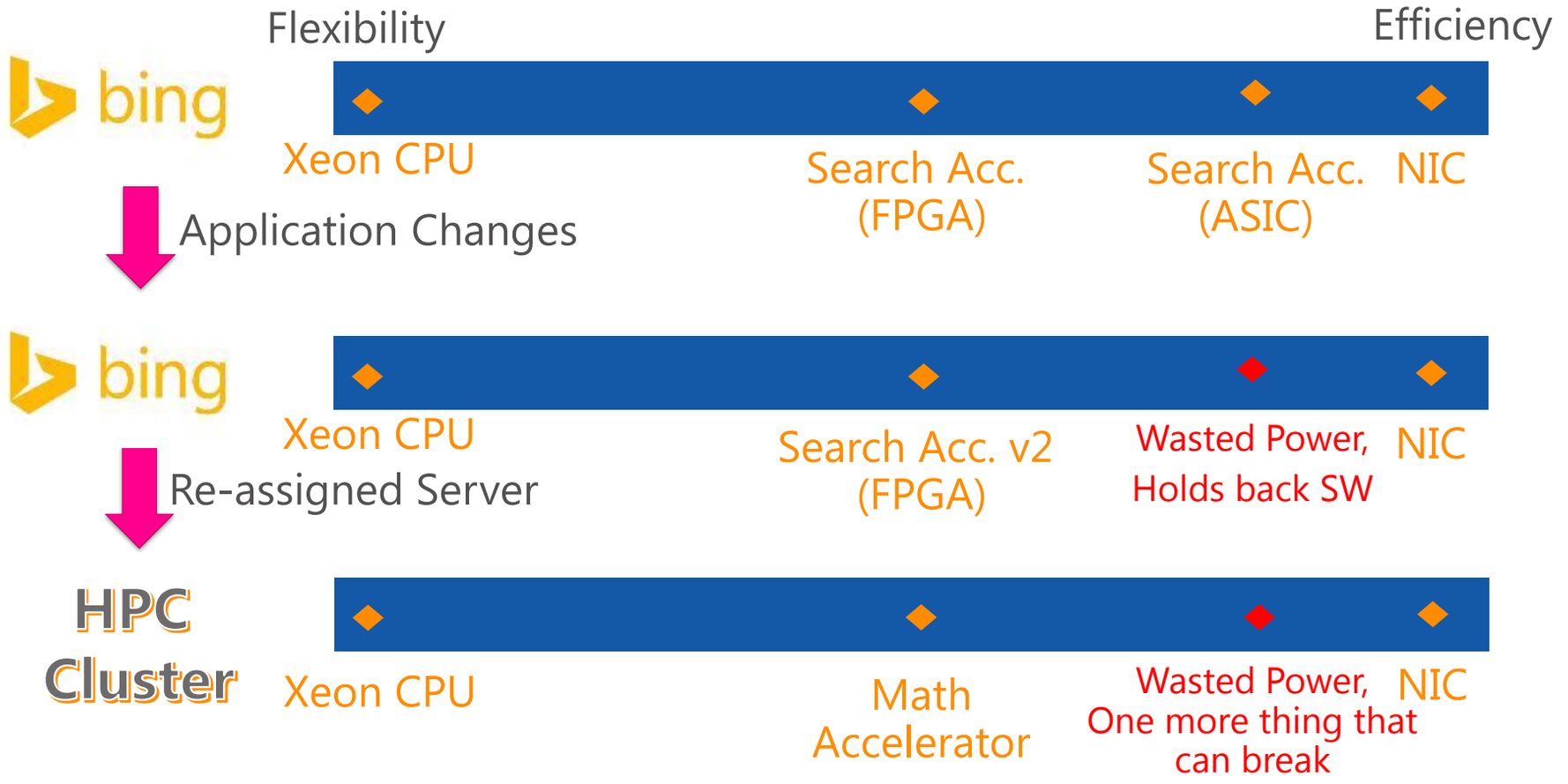


Source: Bob Broderson, Berkeley Wireless group

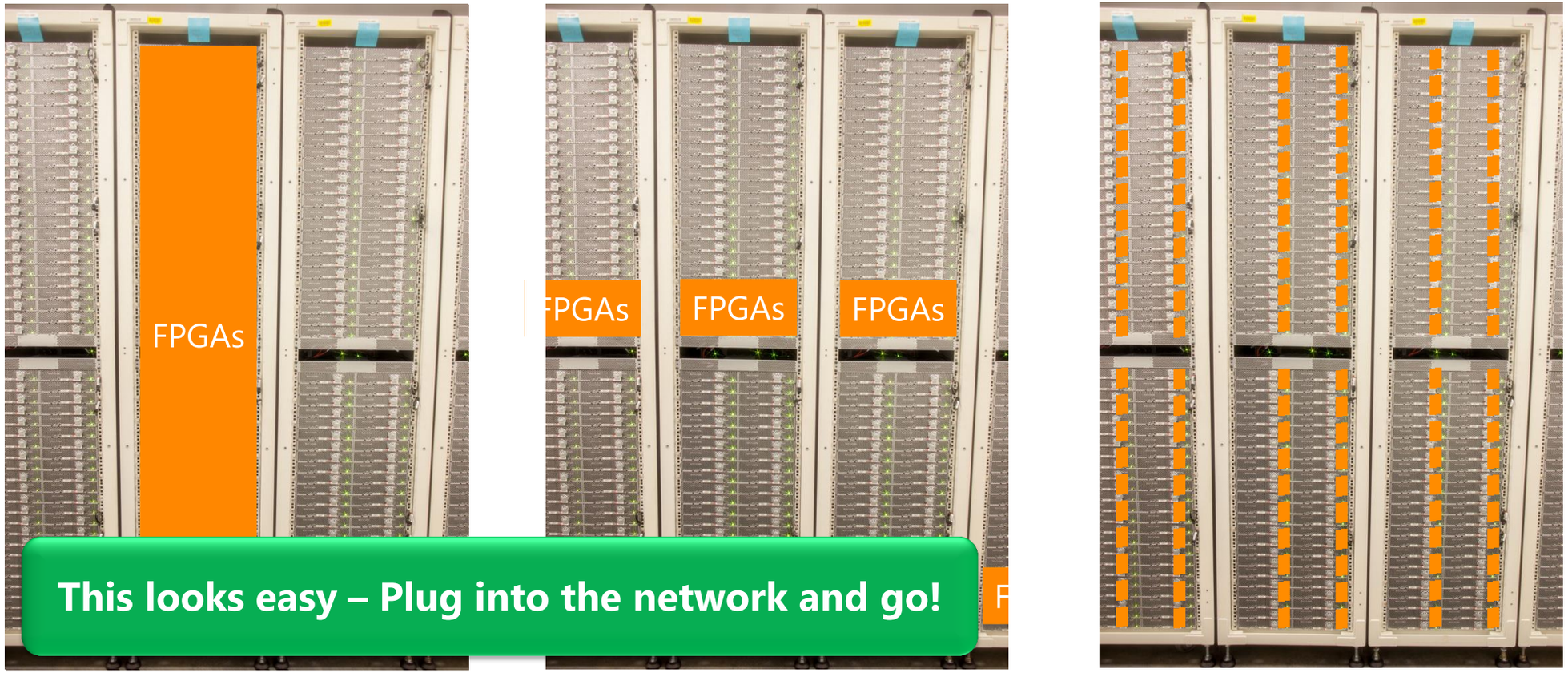
One Application's Accelerator



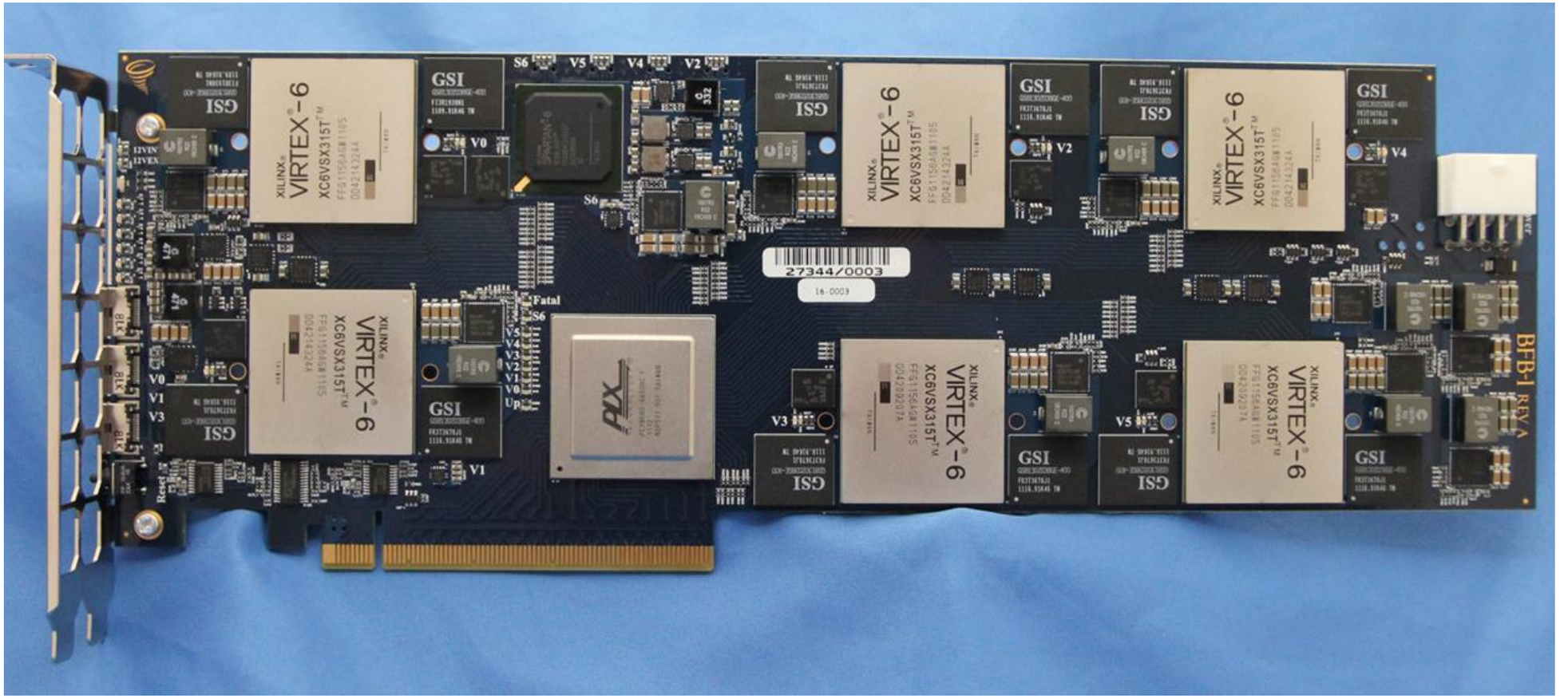
One Application's Accelerator



Integrating FPGAs into the Datacenter

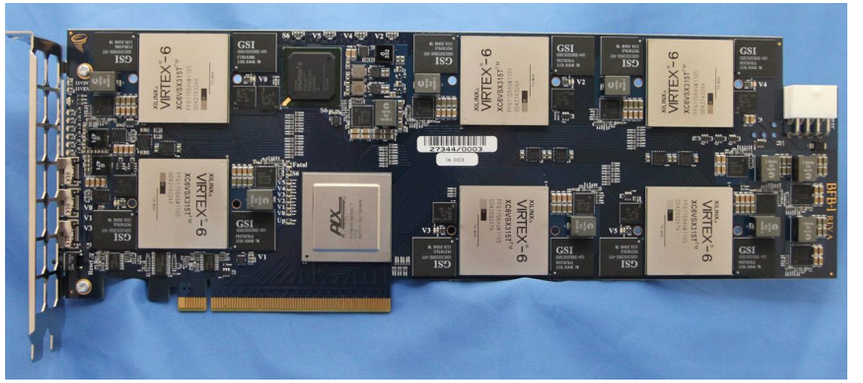


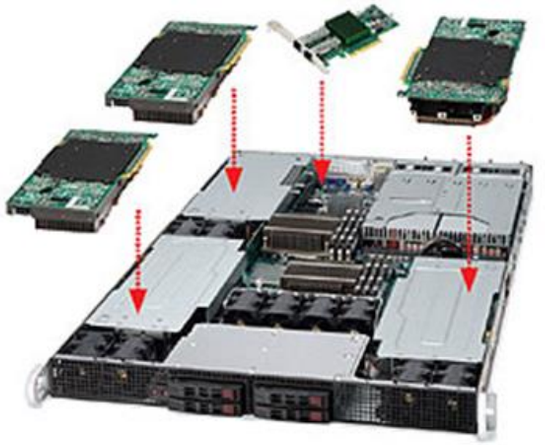
Prototype #1: BFB Board



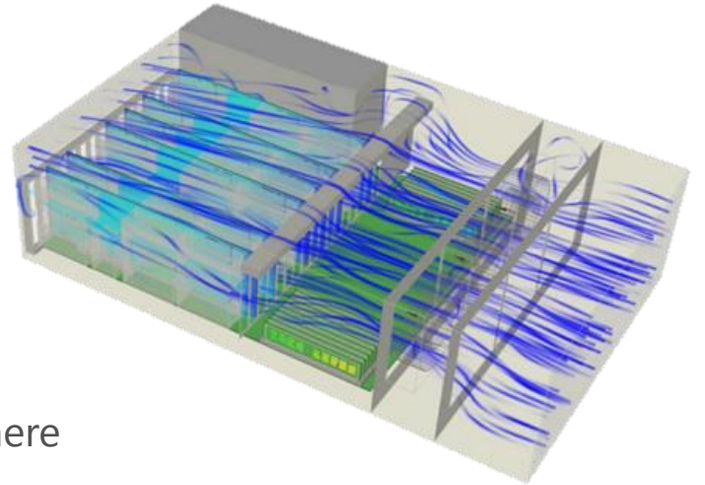
Prototype #1: BFB board

- Prototyped a 6-FPGA board
- 3x2 GPIO mesh
- PCIe connecting all FPGAs, CPU
- Plugs into Supermicro GPU server
- Serves L2 scoring for 48-server pod





- 1U, 2U, or 4U rack-mounted
- 1/2/4 x 10Ge ports
- Up to 4 PCIe x16 slots
- 2 sockets, 6-core Intel Westmere



Centralized Model Unsuitable for Datacenter

- Single point of failure
- Complicates rack design, thermals, maintainability
- Network communication for any use of FPGA
  - Definition of the Network In-cast problem
  - Precludes many latency-sensitive workloads
- Limited elasticity
  - What if you need more than six FPGAs?

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

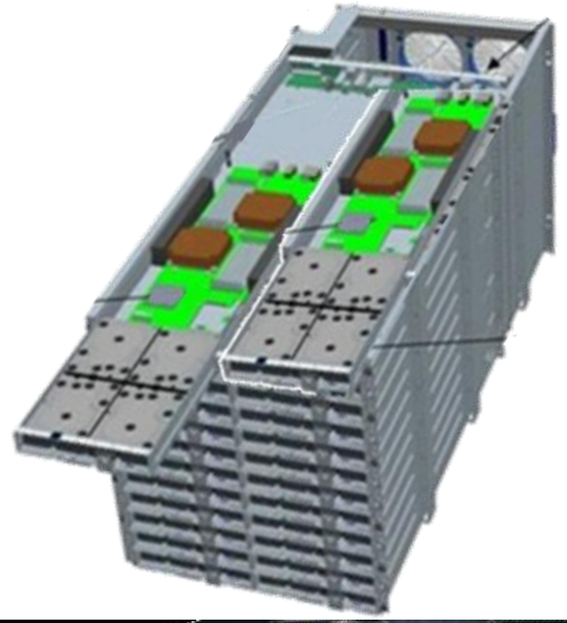
<10% Power Draw
(25W max, all from PCIe)

Don't Break Anything

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

Datacenter Servers

- Microsoft Open Compute Server
- 1U, ½ wide servers
- Enough space & power for ½ height, ½ length PCIe card
- Squeeze in a single FPGA
- Won't fit (or power) GPU



http://www.globalfoundationservices.com/posts/2014/january/27/microsoft-contributes-cloud-server-specification-to-open-compute-project.aspx

Microsoft Open Compute Server

- Two 8-core Xeon 2.1 GHz CPUs
- 64 GB DRAM
- 4 HDDs @ 2 TB, 2 SSDs @ 512 GB
- 10 Gb Ethernet
- No cable attachments to server

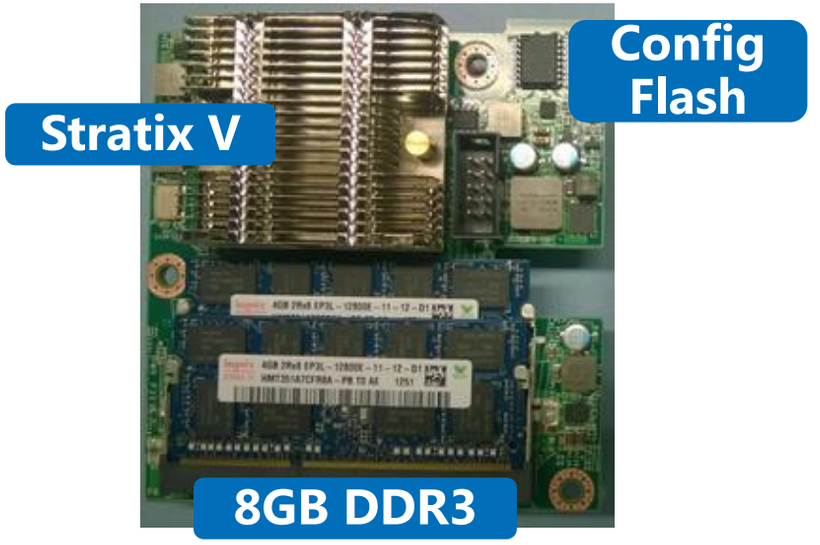
Air flow

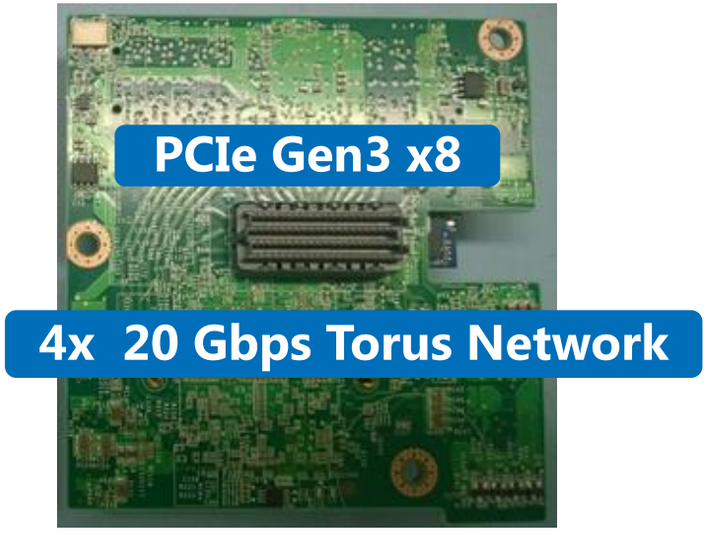
200 LFM

68 ^{0}C Inlet

Catapult FPGA Accelerator Card

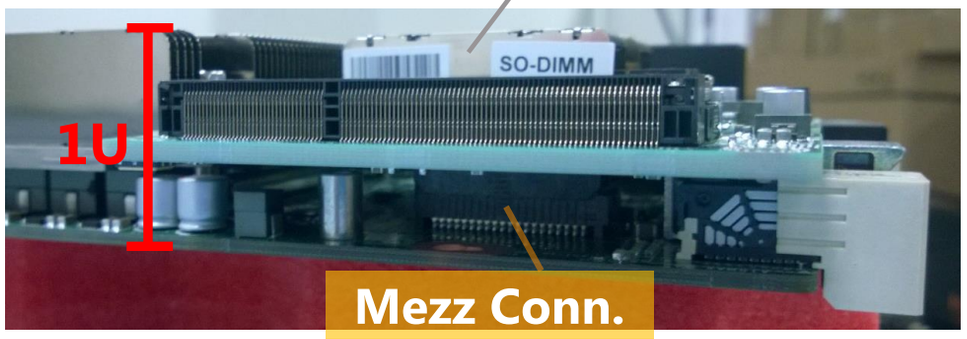
- Altera Stratix V GS D5
  - 172k ALMs, 2,014 M20Ks, 1,590 DSPs
- 8GB DDR3-1333
- 32 MB Configuration Flash
- PCIe Gen 3 x8
- 8 lanes to Mini-SAS SFF-8088 connectors
- Powered by PCIe slot

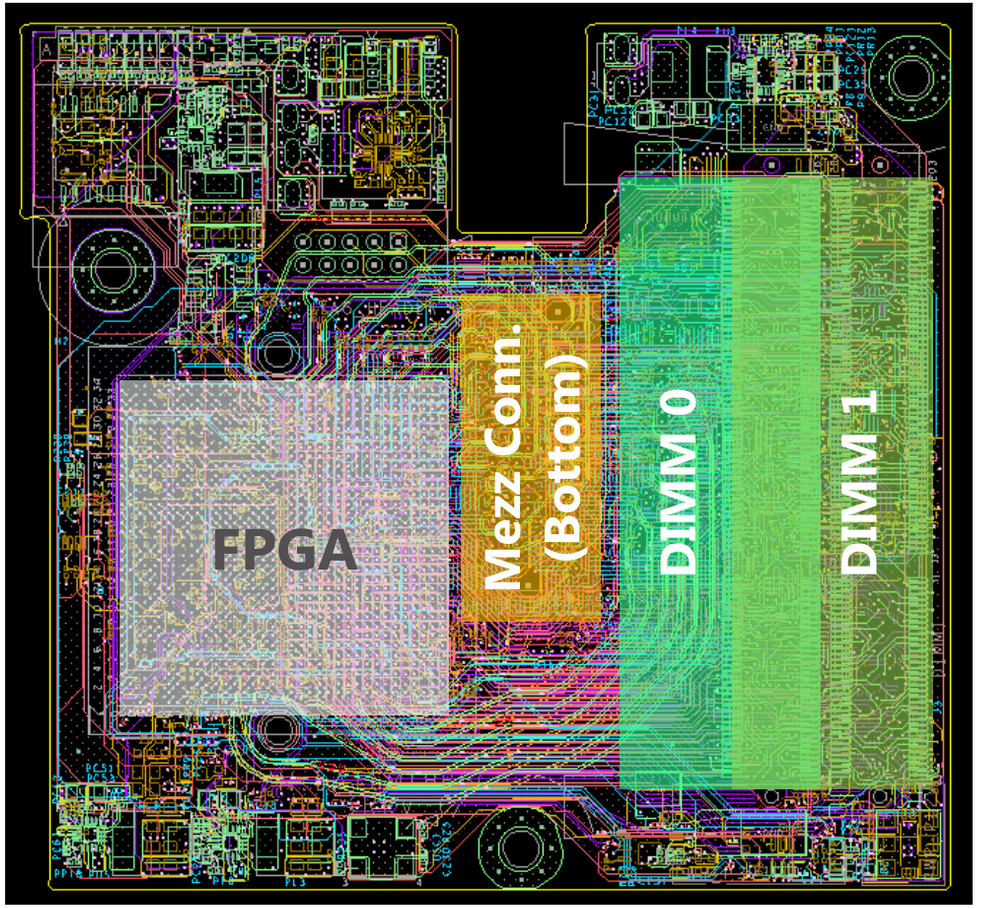




Board Details

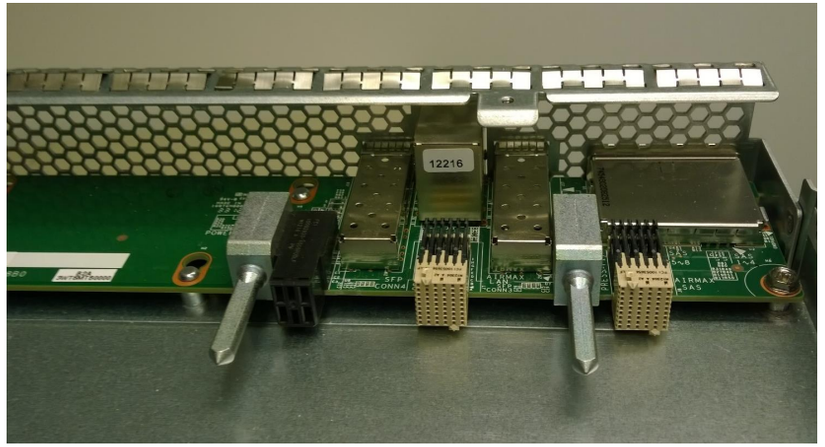
- 16 Layer, FR408
- 9.5cm x 8.8cm x 115.8 mil
- 35mm x 35mm FPGA
- 14.2mm high heatsink



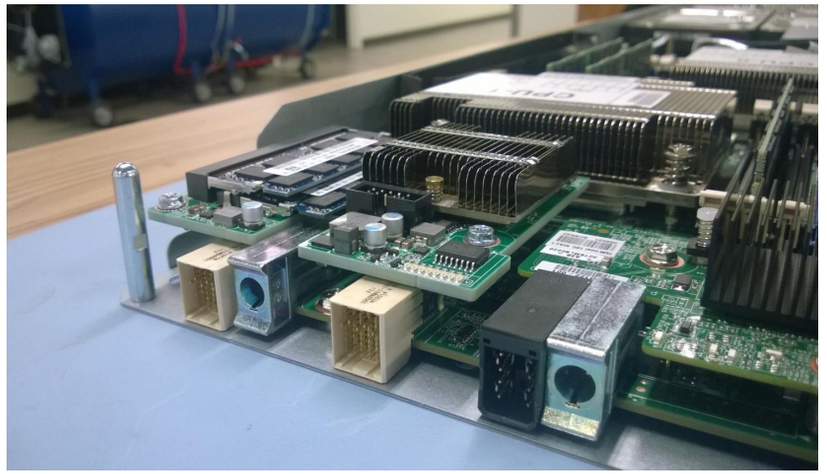


Board / Server Integration

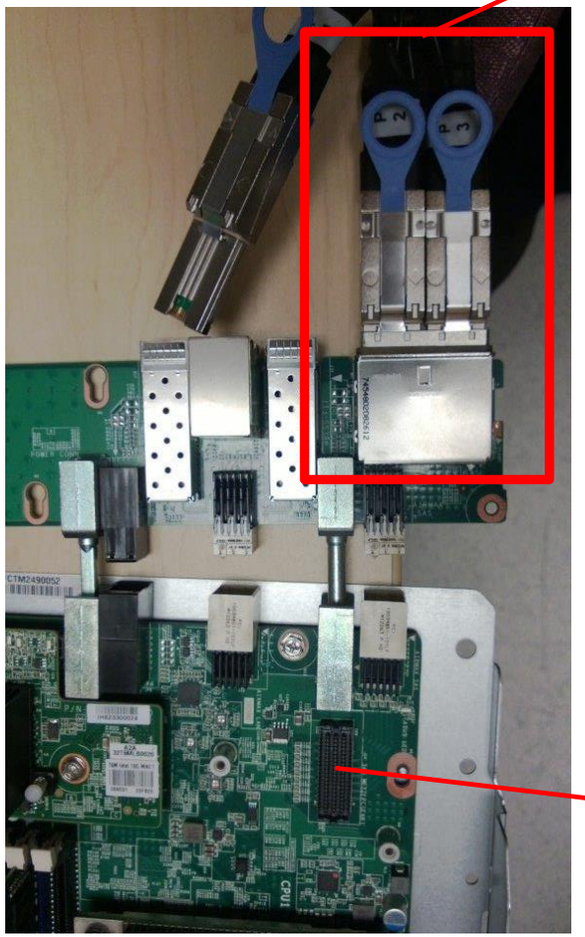
I/O Backplane



Server w/ Catapult Board



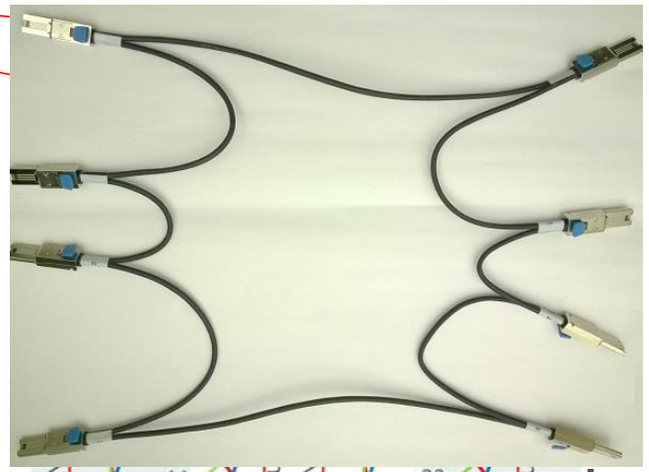
Catapult Network Cables
(Mini-SAS / SFF-8088)

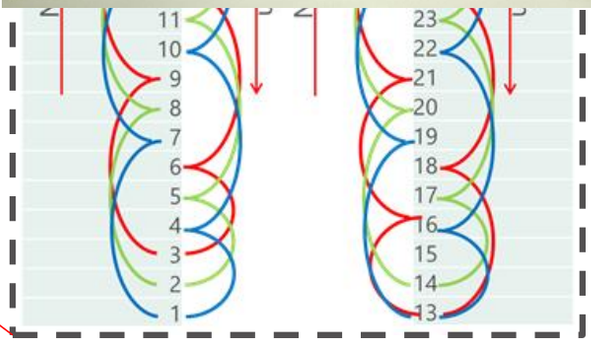


Boards Connected Together

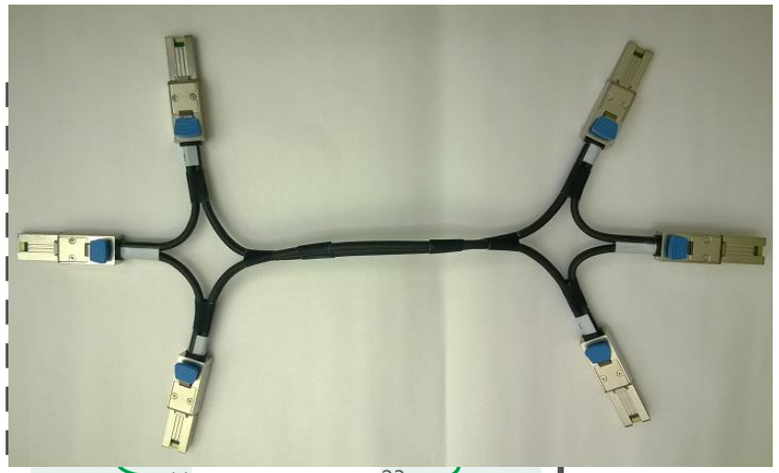
Catapult Board
Mezzanine Slot

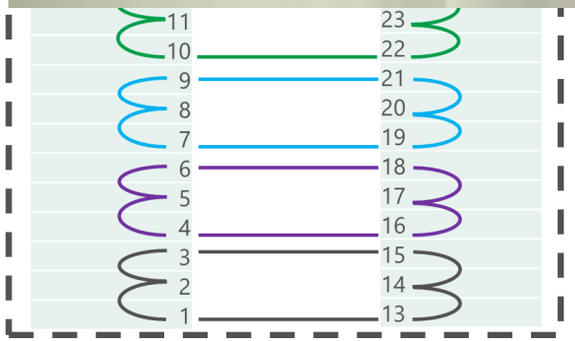
6x8 Torus in a 2x24 Server Layout





8-Shell Cables

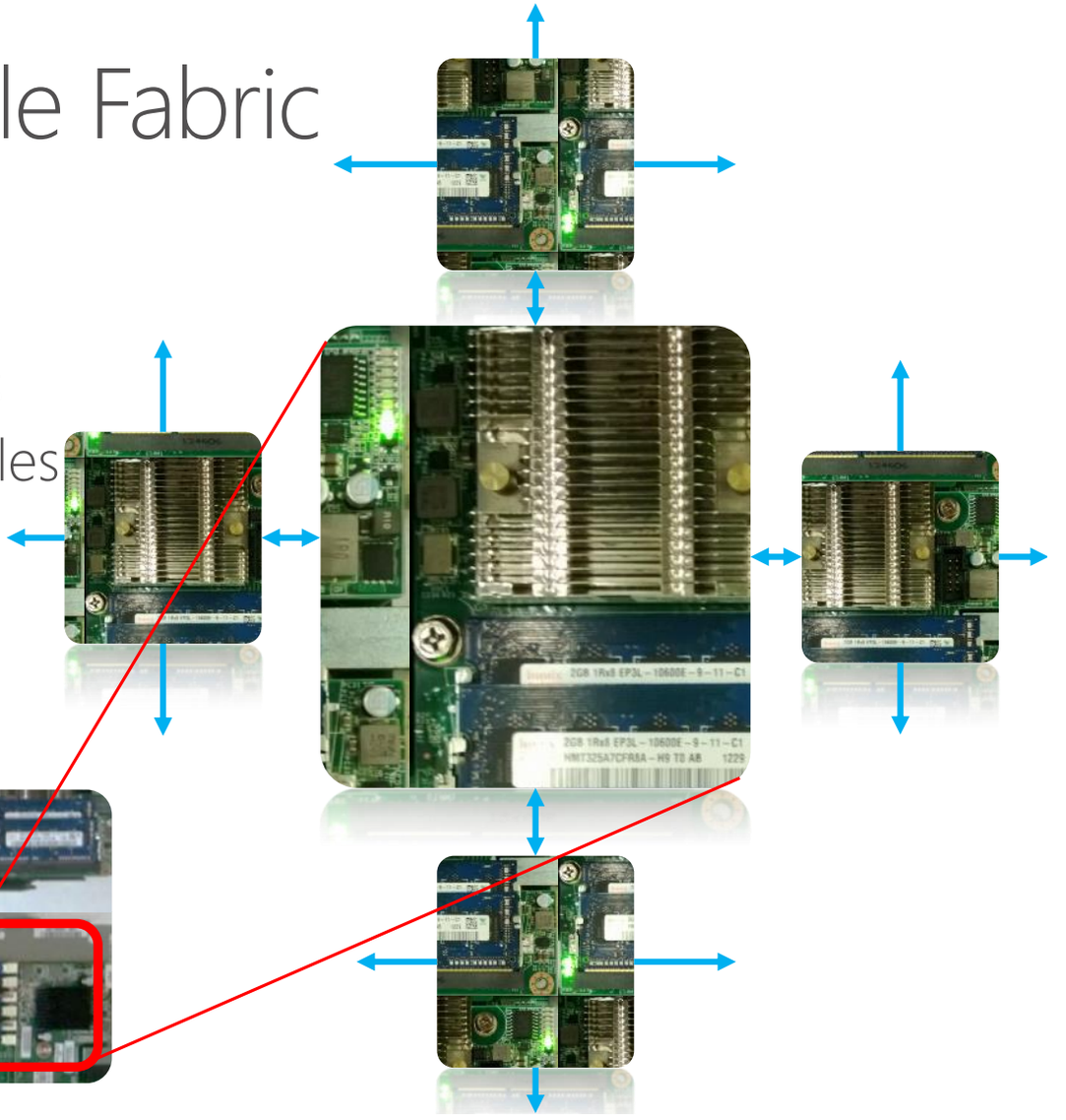




6-Shell Cables

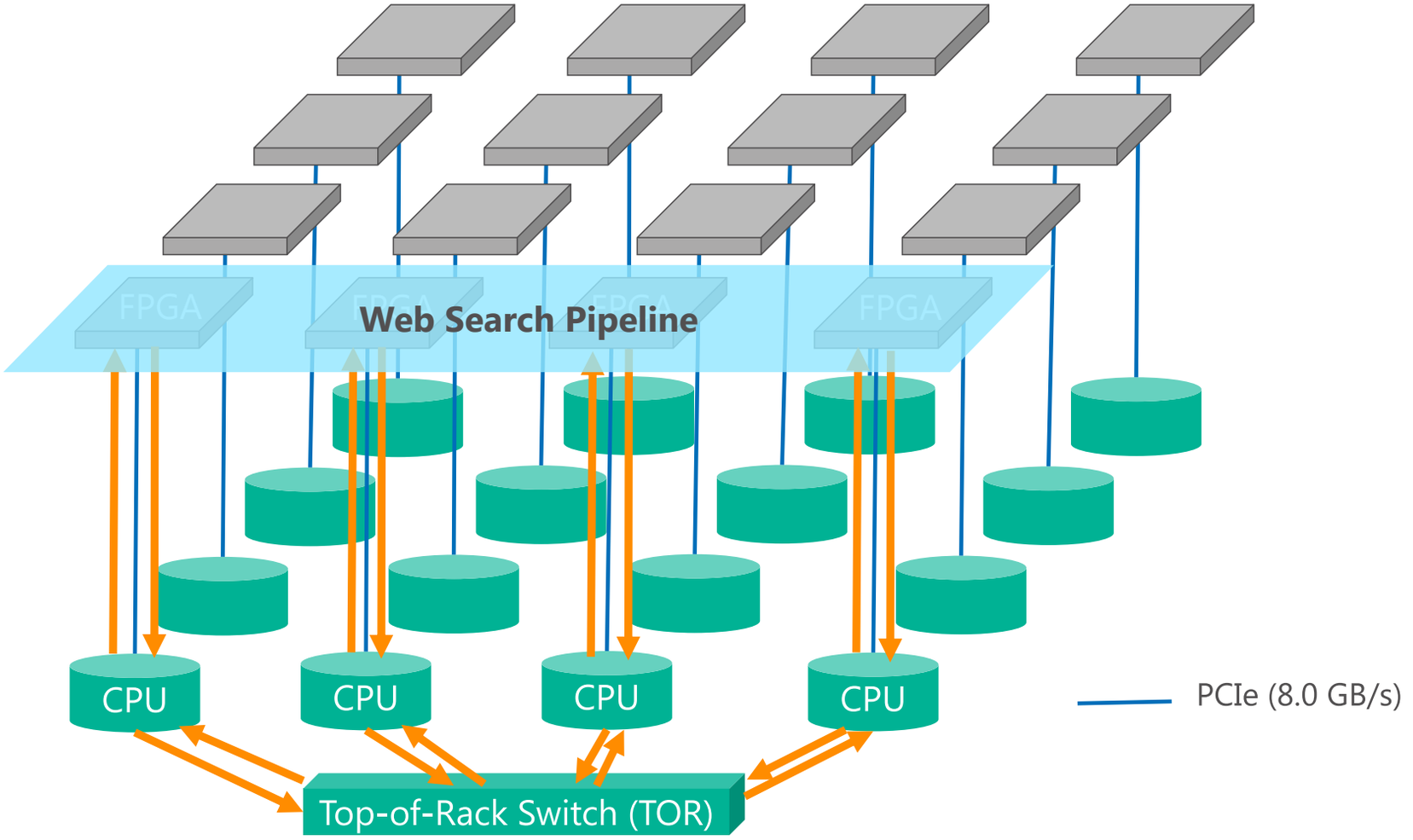
Scalable Reconfigurable Fabric

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

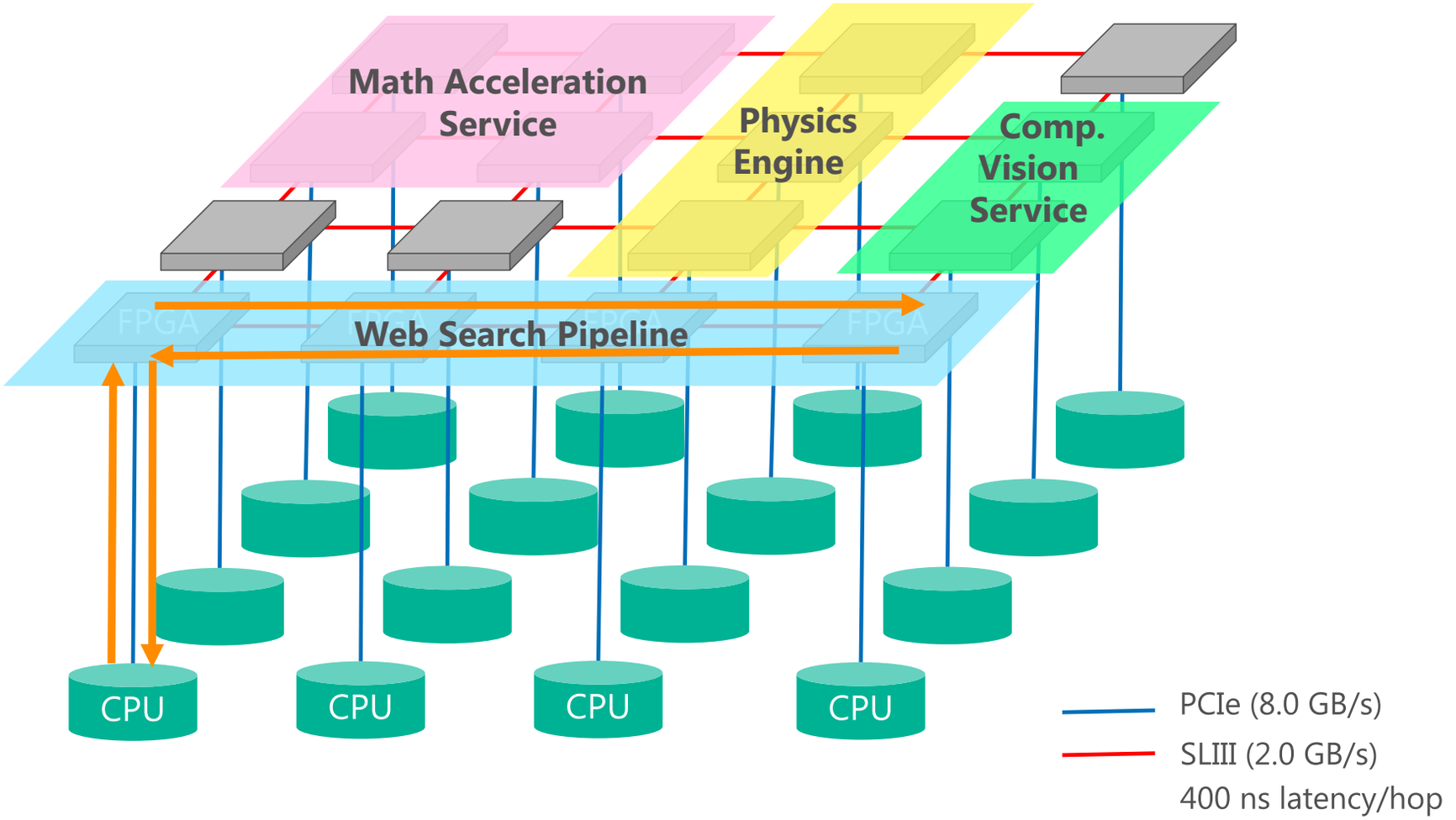


Data Center Server (1U, ½ width)

An Elastic Reconfigurable Fabric

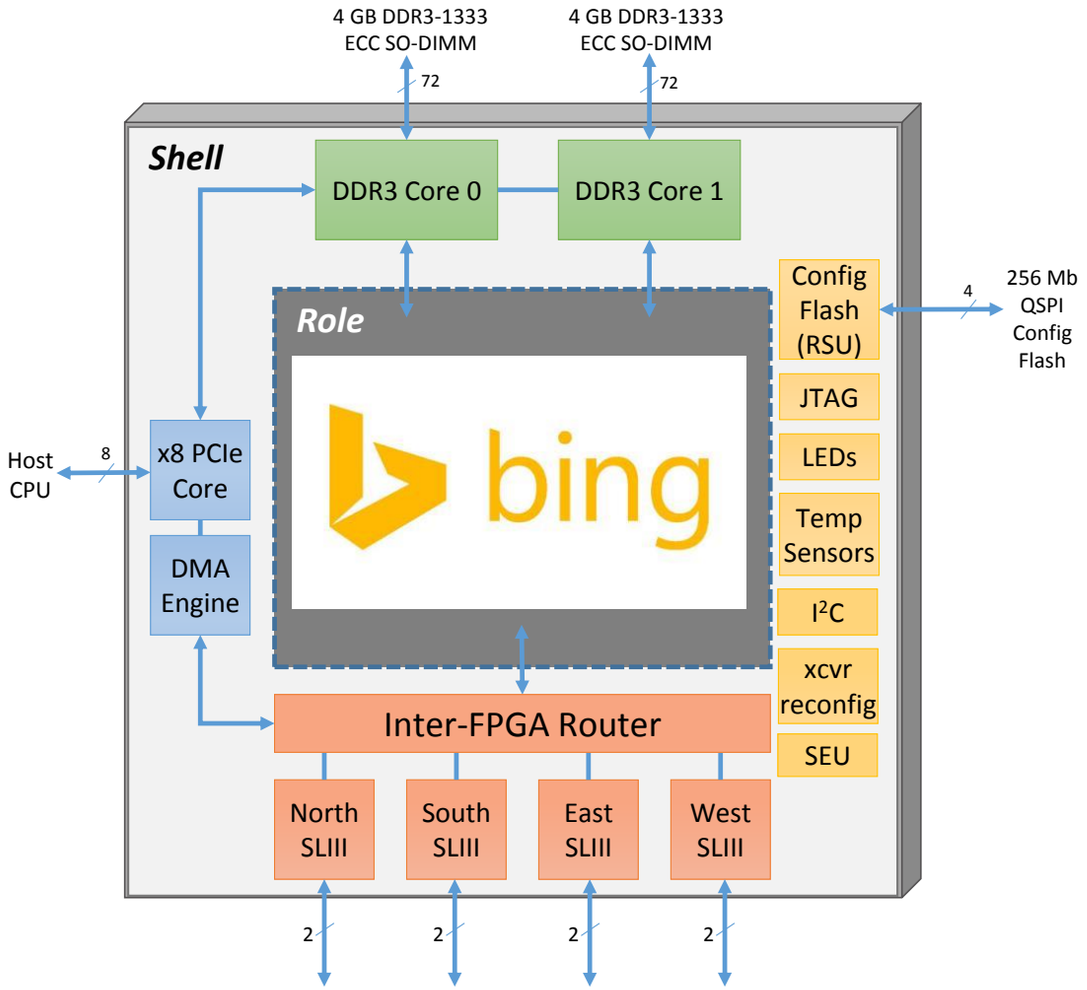


An Elastic Reconfigurable Fabric

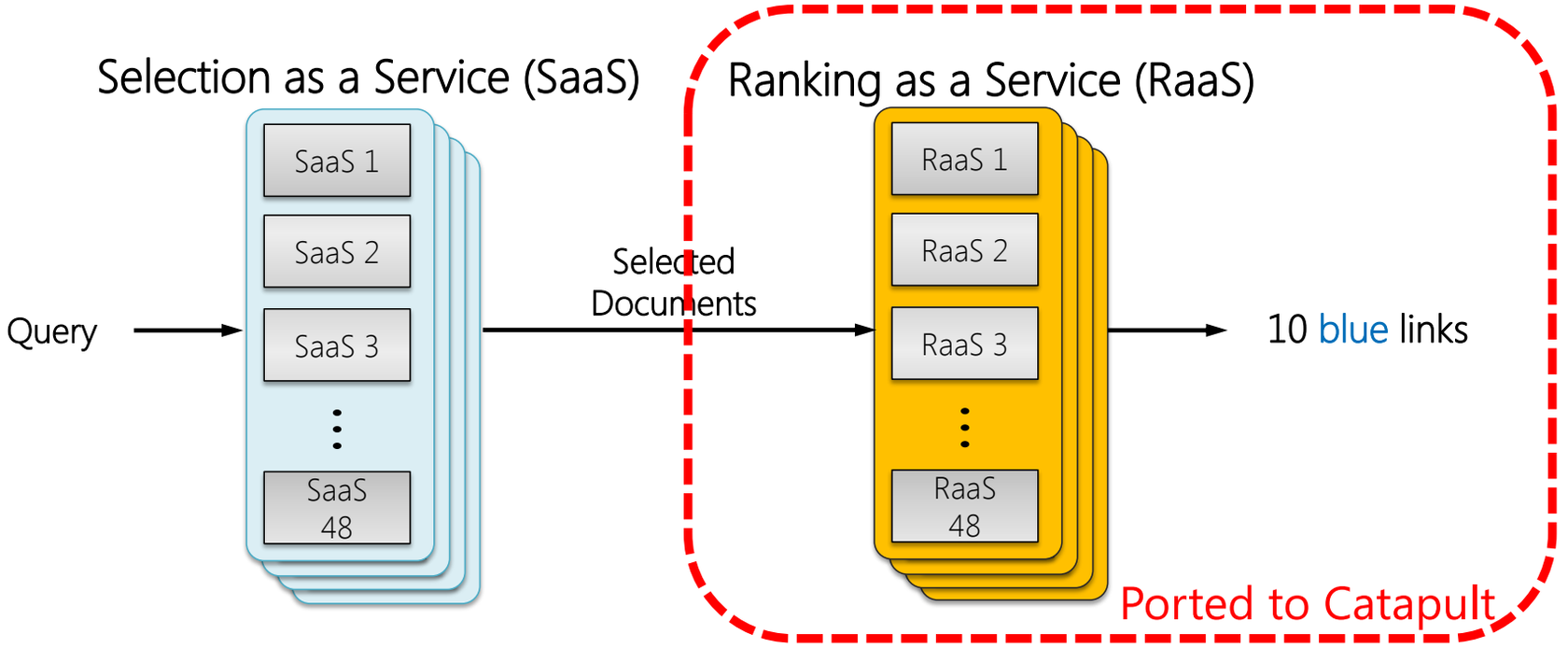


Shell & Role

- *Shell* handles all I/O & management tasks
- *Role* is only application logic
- Shell exposes simple FIFOs
- Flight data recorder for scale-out debug
- Role is Partial Reconfig boundary



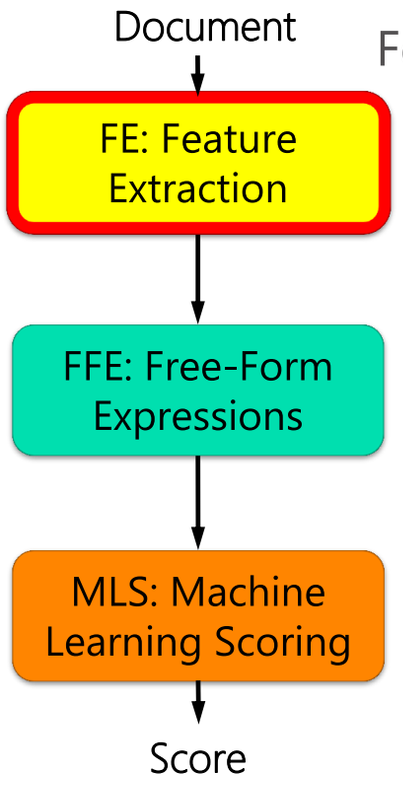
Bing Document Ranking Flow



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

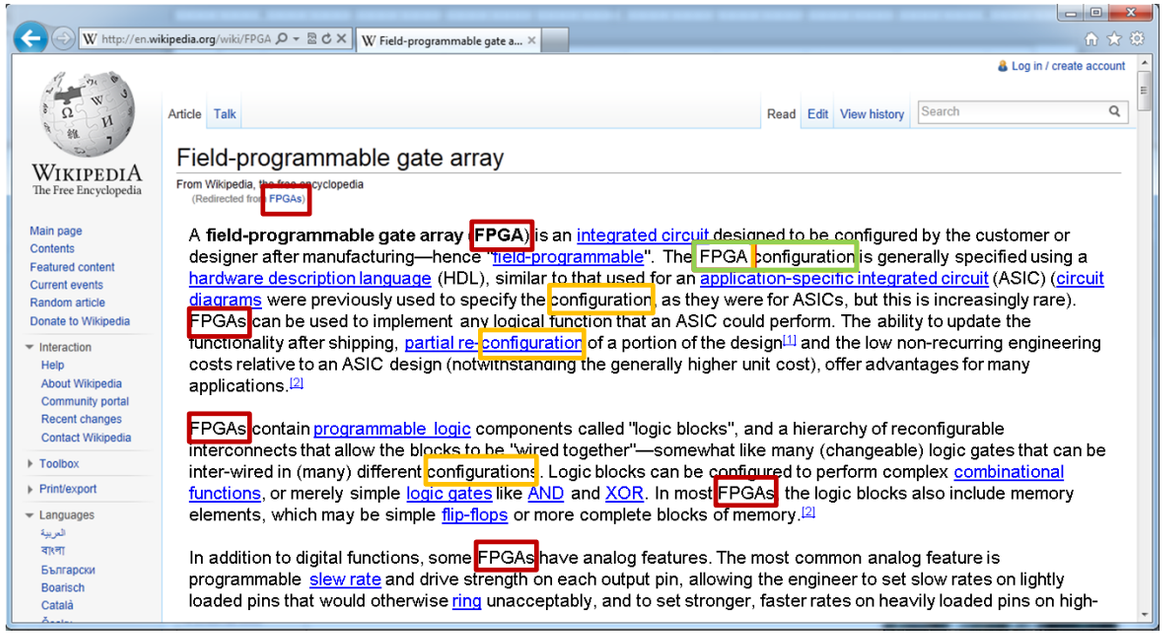
Ranking-as-a-Service (RaaS)
- Compute scores for how relevant each selected document is for the search query
- Sort the scores and return the results

FE: Feature Extraction

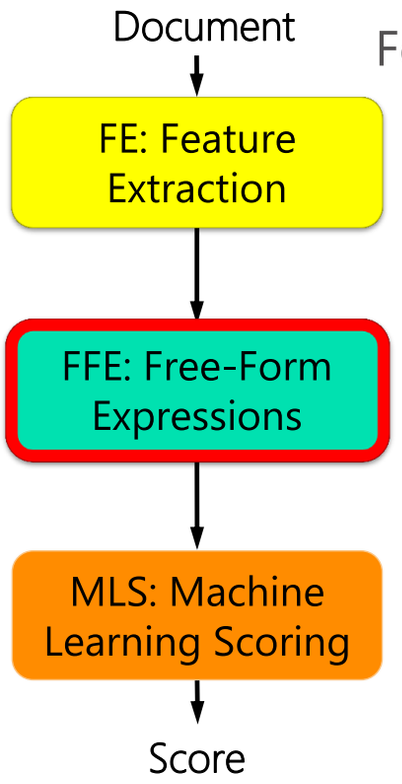


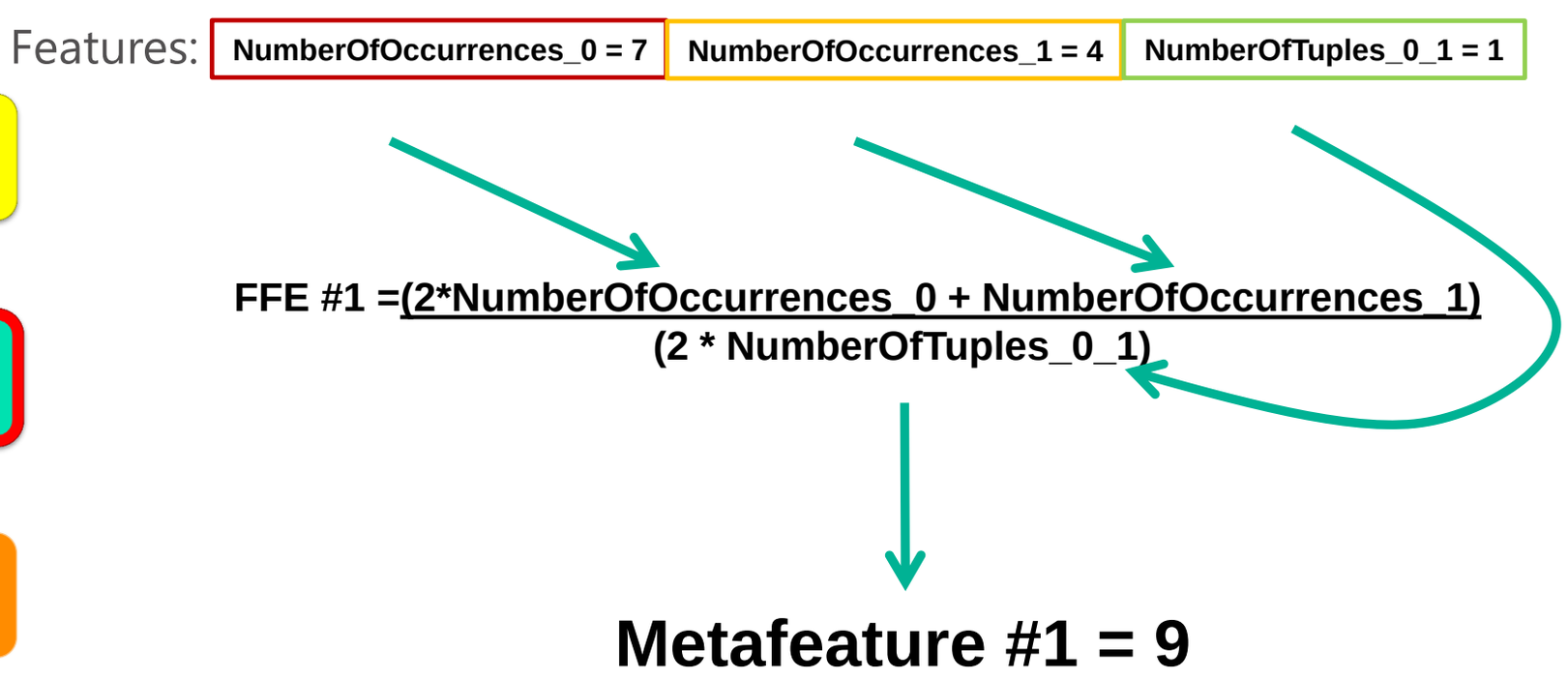
Query: "FPGA Configuration"

Features: **NumberOfOccurrences_0 = 7** | **NumberOfOccurrences_1 = 4** | **NumberOfTuples_0_1 = 1**

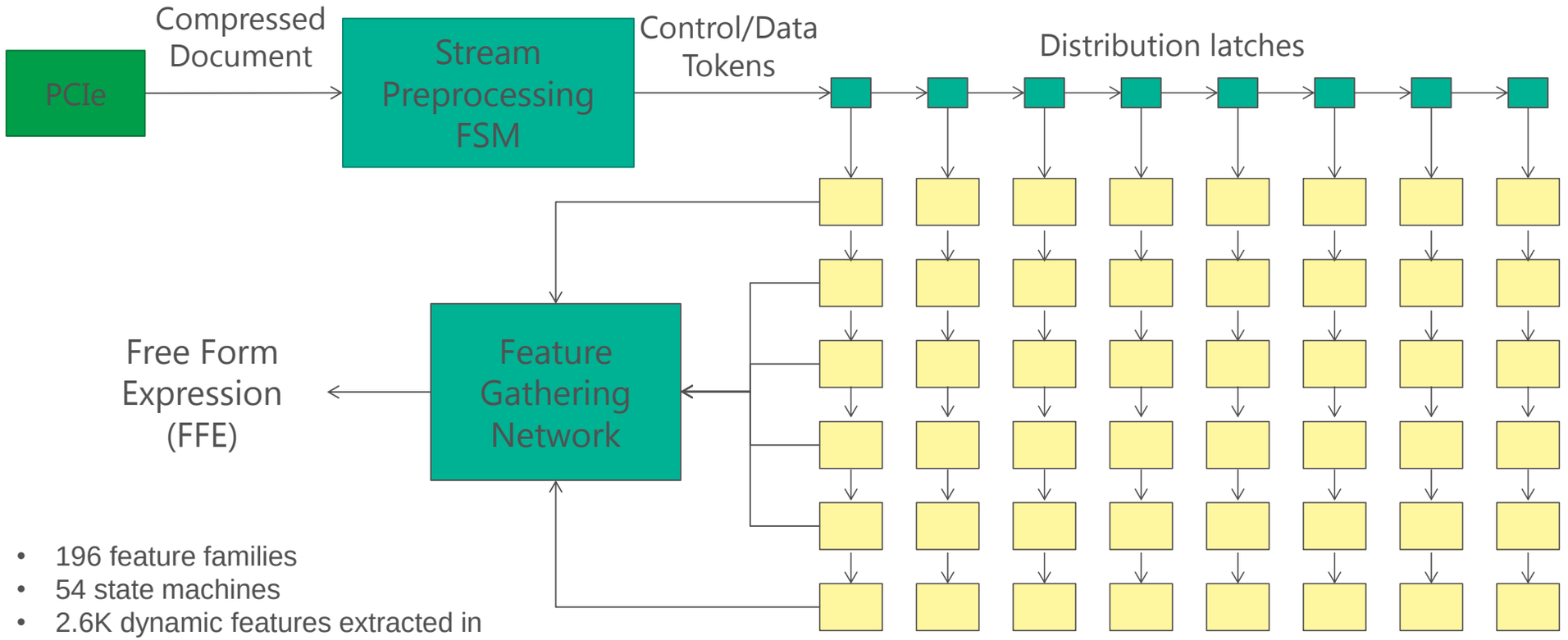


FFE: Free Form Expressions





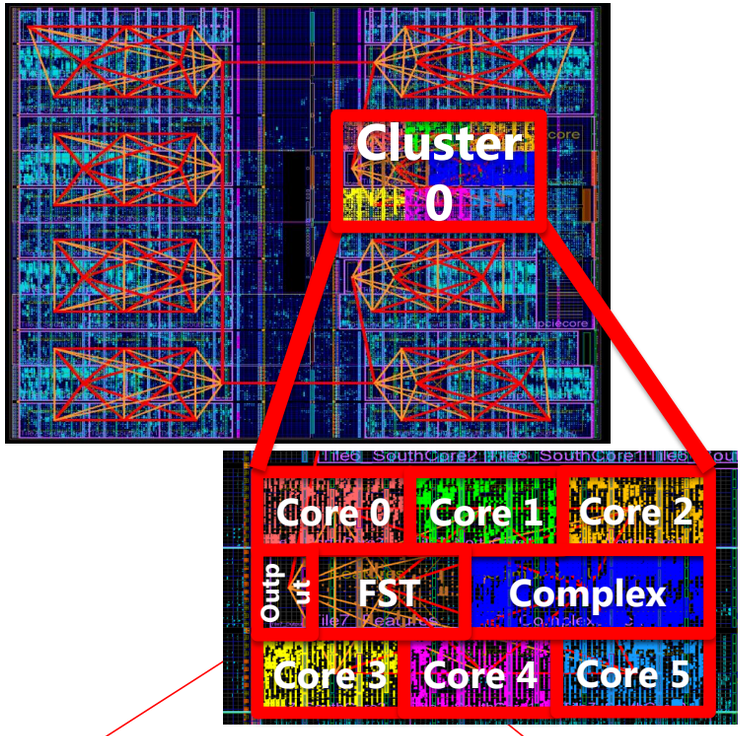
Feature Extraction Accelerator

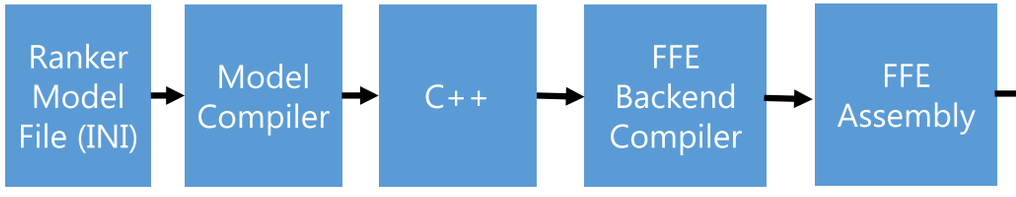


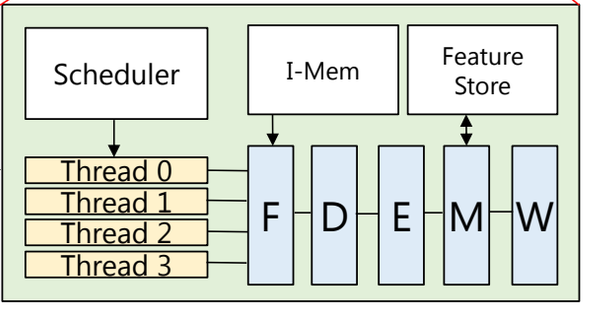
- 196 feature families
- 54 state machines
- 2.6K dynamic features extracted in less than 4us (~600us in SW)

FFE Soft Cores

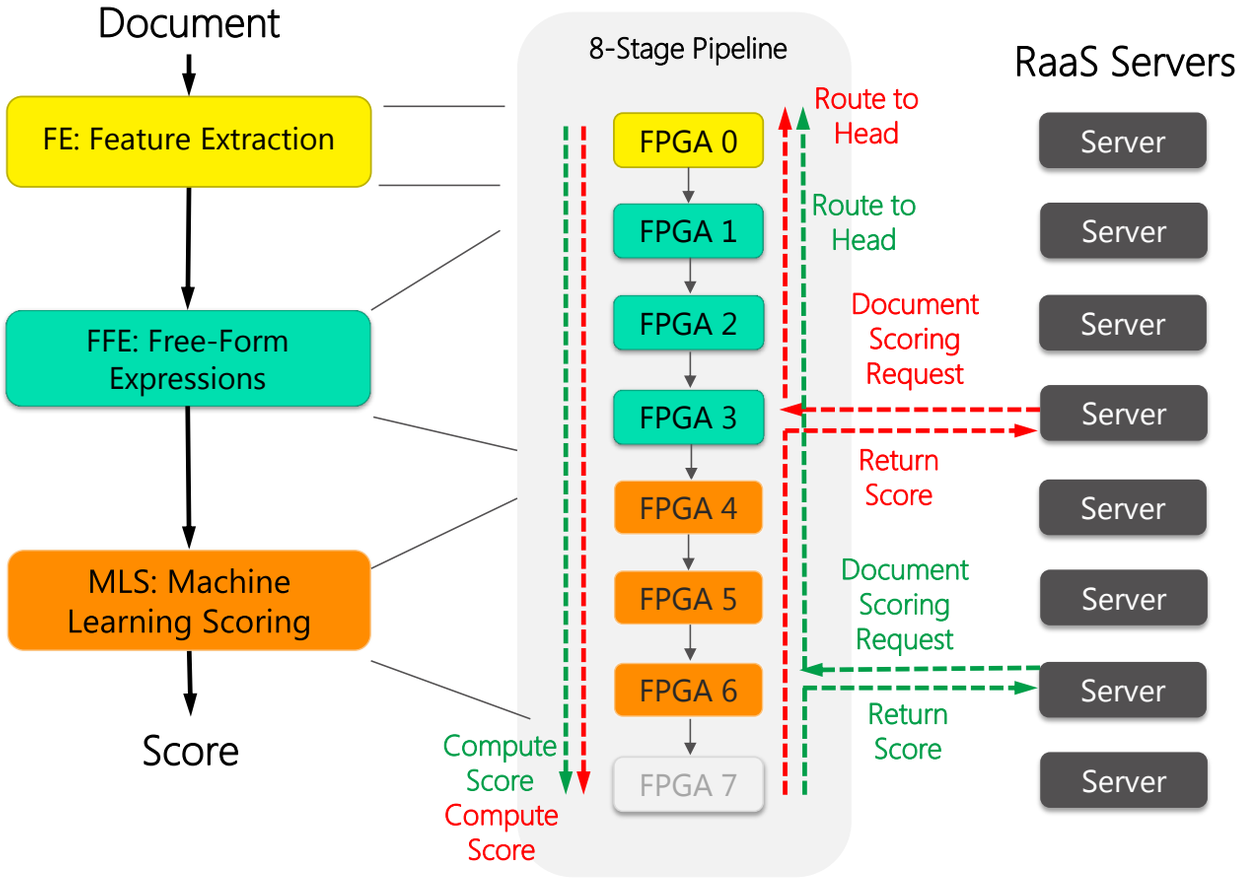
- Soft processor for multi-threaded throughput
- 4 HW threads per core
- 6 cores share a complex ALU
- log, divide, exp, float/int conv.
- 10 clusters (240 HW threads) per FPGA

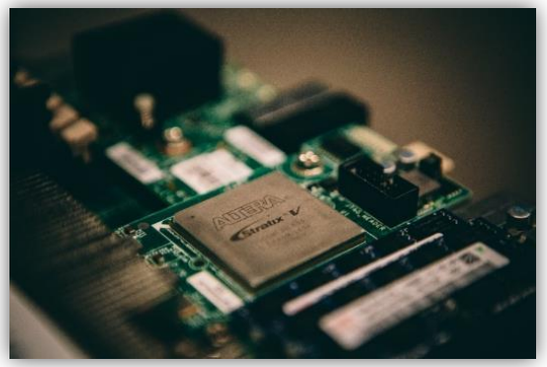


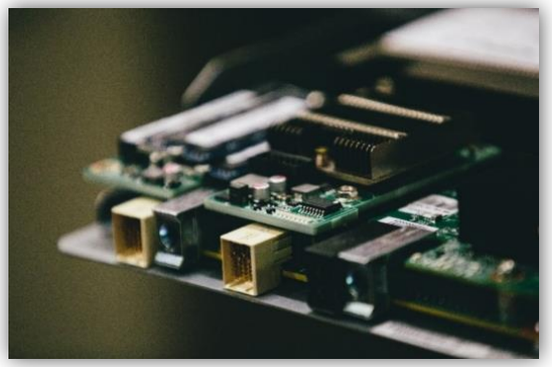




Putting it all together



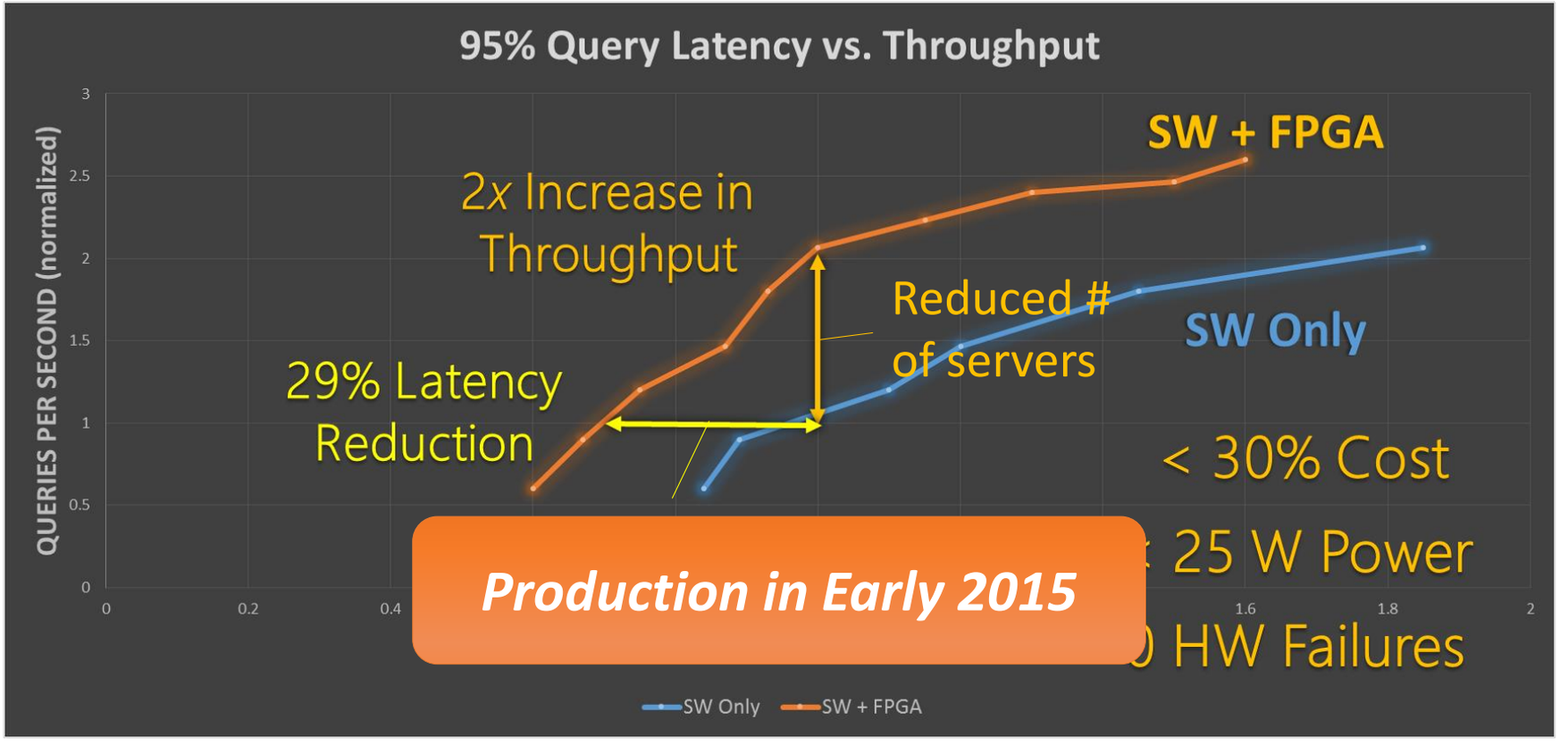




1,632 Server Pilot Deployed in a Production Datacenter

Accelerating Large-Scale Services – Bing Search

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



FPGAs for Application Acceleration

- Hardware is the "easy" part
- Software changes fast
- Services last across hardware generations

Ensuring Maintainability & Sustainability

Software & Language

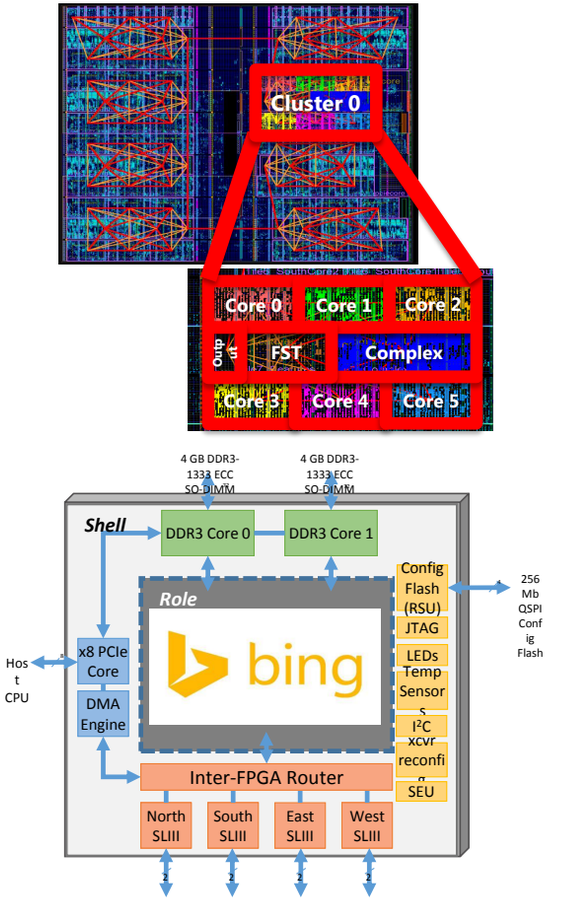
- Structure software for explicit data communication
  - Pass-by-value vs. Pass-by-reference

Compiler & Tools

- Create programmable substrates
  - "Programming" FFE is just writing C++ code
  - Wide spectrum between CPU and full custom RTL

RTL & Hardware

- Shell & Role, Common APIs
  - Abstract away board interfaces & FPGA details from the application



Key Needs for FPGA Computing

Software & Language

- Huge need for high-productivity languages
  - C-to-gates tools did not do well on FE state machines
  - Domain-specific languages, OpenCL, BlueSpec both show promise

Compiler & Tools

- Faster compilation times
- Fewer warnings… NO warnings on IP libraries
- Better debugging integration

RTL & Hardware

- Hardened PCIe, DDR, JTAG debugging
- Faster, more efficient DDR
- Improved floating-point performance

Conclusions

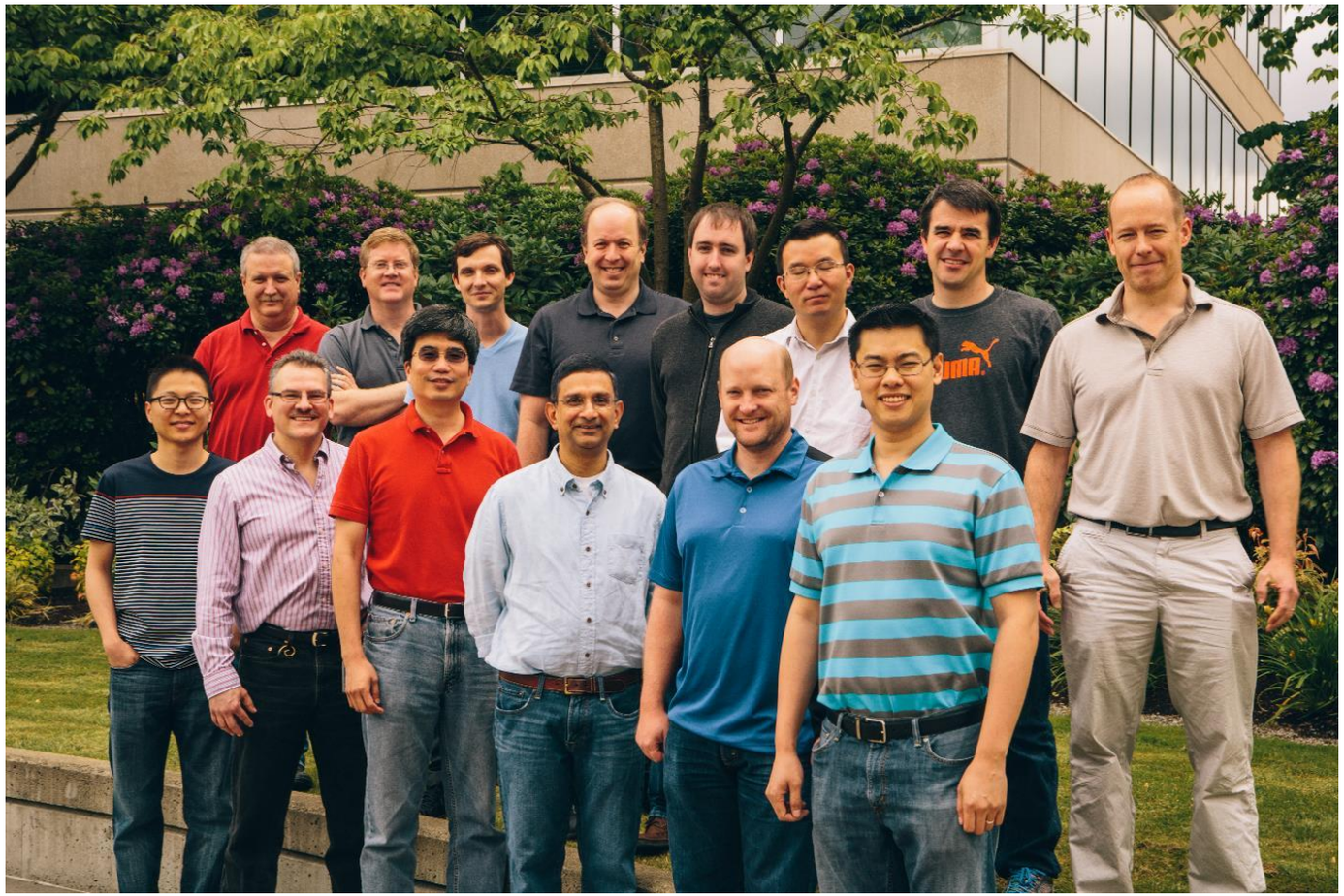
- Hardware specialization is a (the?) way to gain efficiency and performance
- The Catapult reconfigurable fabric offers a flexible, elastic pool of resources to accelerate services
- Results for Bing: ½ the number of ranking servers, lower latency, reduced variance, proven scalability, proven resilience
- Bing going to production in early 2015
- Biggest future problem is programmability

Microsoft

Top Row: Eric Peterson, Scott Hauck, Aaron Smith, Jan Gray, Adrian M. Caulfield, Phillip Yi Xiao, Michael Haselman, Doug Burger

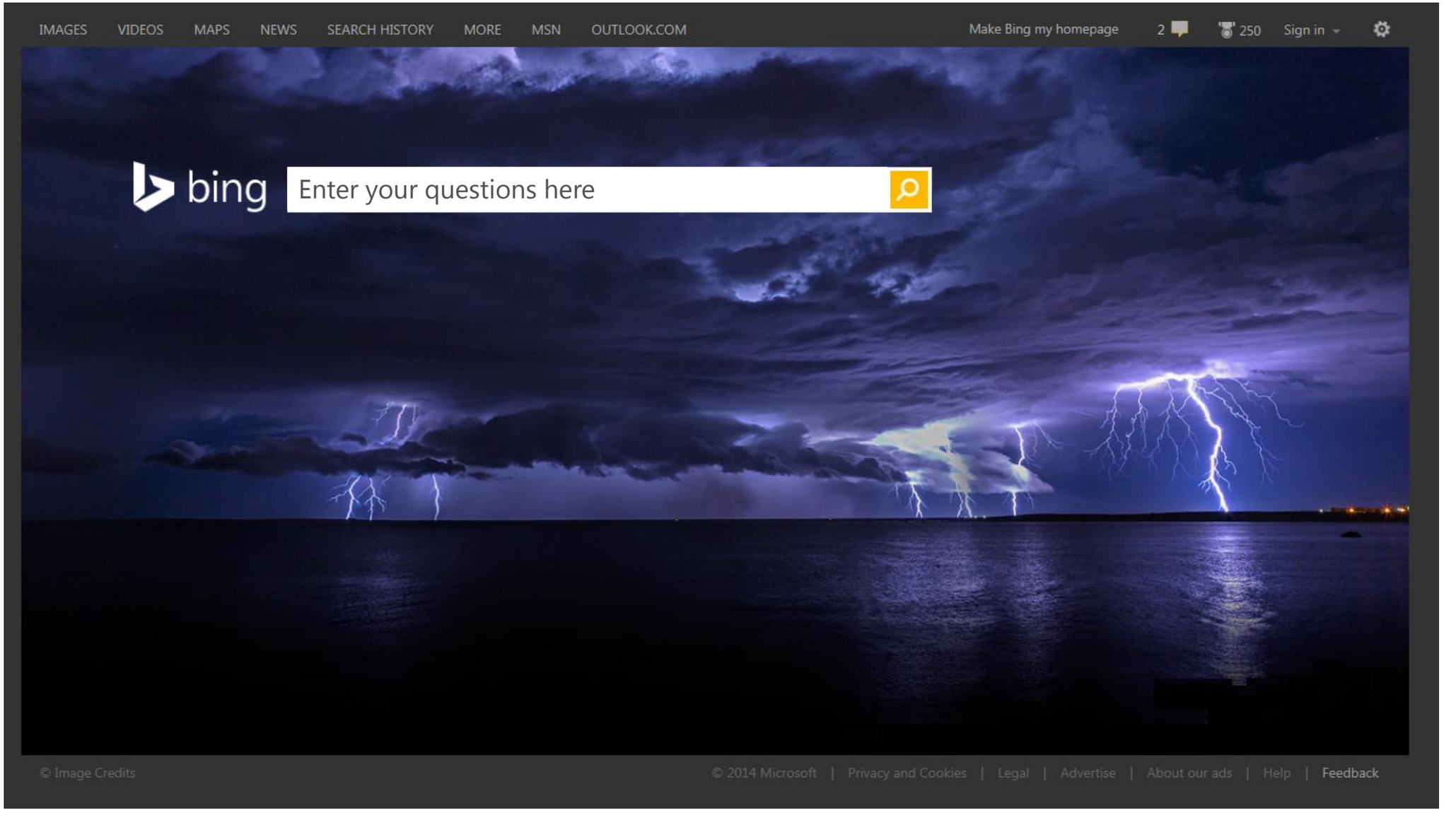
Bottom Row: Joo-Young Kim, Stephen Heil, Derek Chiou, Sitaram Lanka, Andrew Putnam, Eric S. Chung,

Not Pictured: Kypros Constantinides, John Demme, Hadi Esmaeilzadeh, Jeremy Fowers, Gopi Prashanth Gopal, Amir Hormati, James Larus, Simon Pope, Jason Thong



Huge thanks to our partners at

ALTERA


IMAGES
VIDEOS
MAPS
NEWS
SEARCH HISTORY
MORE
MSN
OUTLOOK.COM
Make Bing my homepage
2
250
Sign in
bing
Enter your questions here
© Image Credits
© 2014 Microsoft | Privacy and Cookies | Legal | Advertise | About our ads | Help | Feedback

Microsoft Research