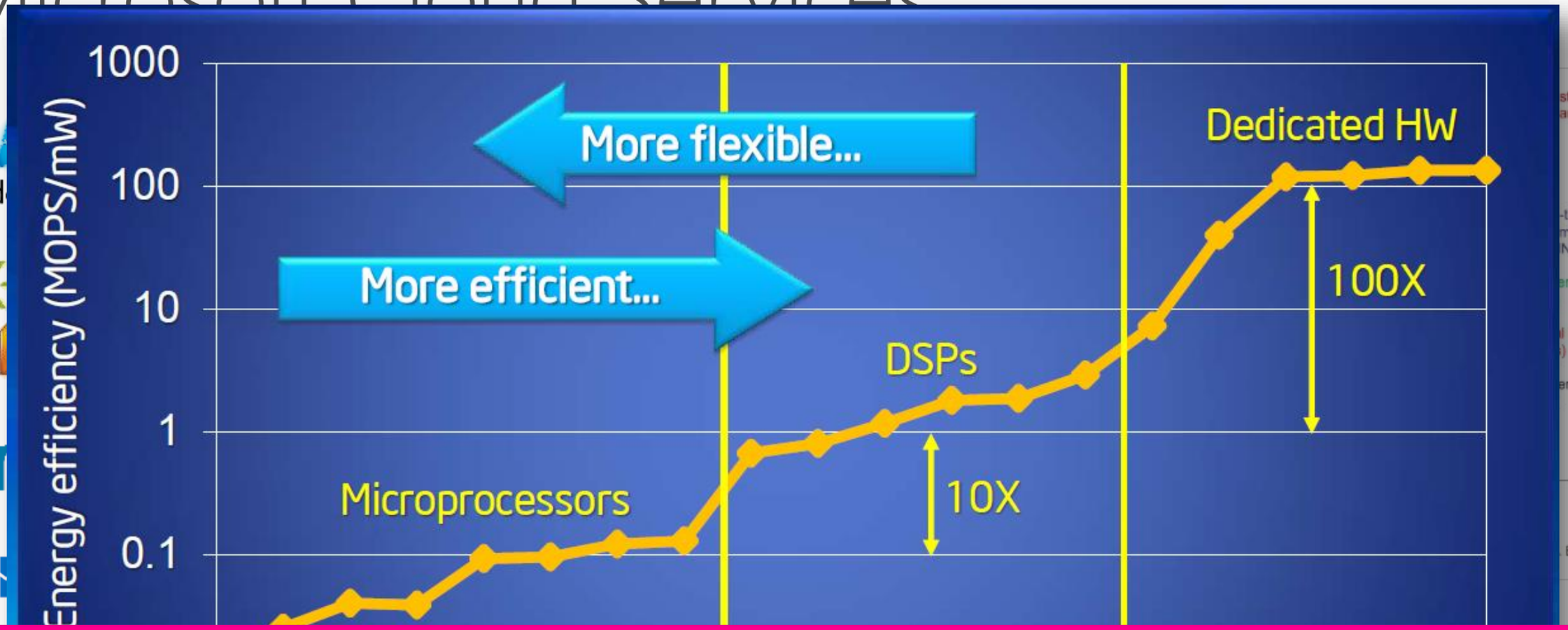
Microsoft Research Faculty Summit 2014

15TH ANNUAL


IMAGES VIDEOS MAPS NEWS SEARCH HISTORY MORE MSN OUTLOOK.COM
Make Bing my homepage
2
250
Sign in
bing
A Reconfigurable Fabric for Accelerating Large-Scale Datacenter Services
© Image Credits
© 2014 Microsoft | Privacy and Cookies | Legal | Advertise | About our ads | Help | Feedback

Microsoft Cloud Services



Increase Efficiency with Hardware Specialization

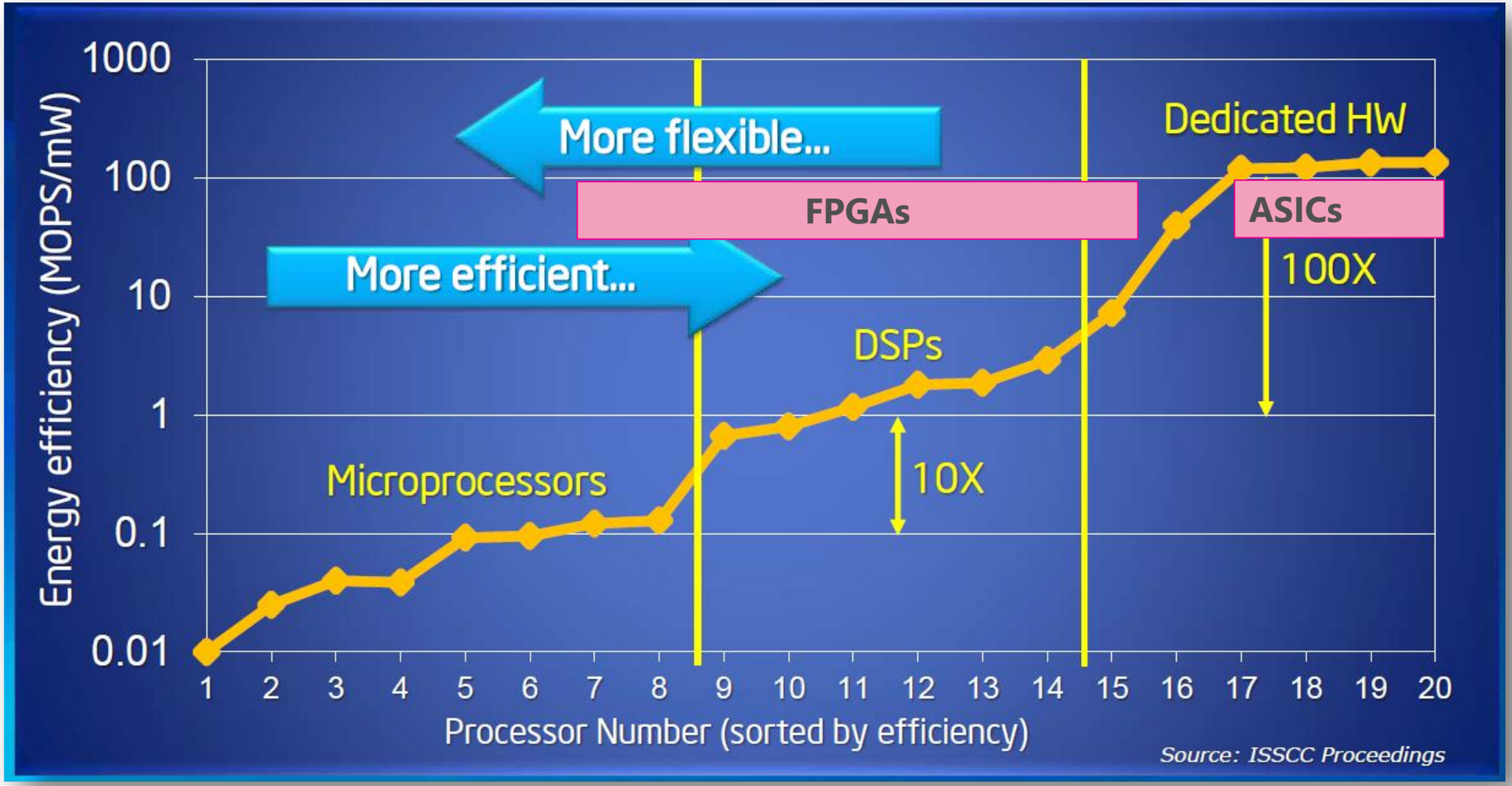
Datacenter Environment

- 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

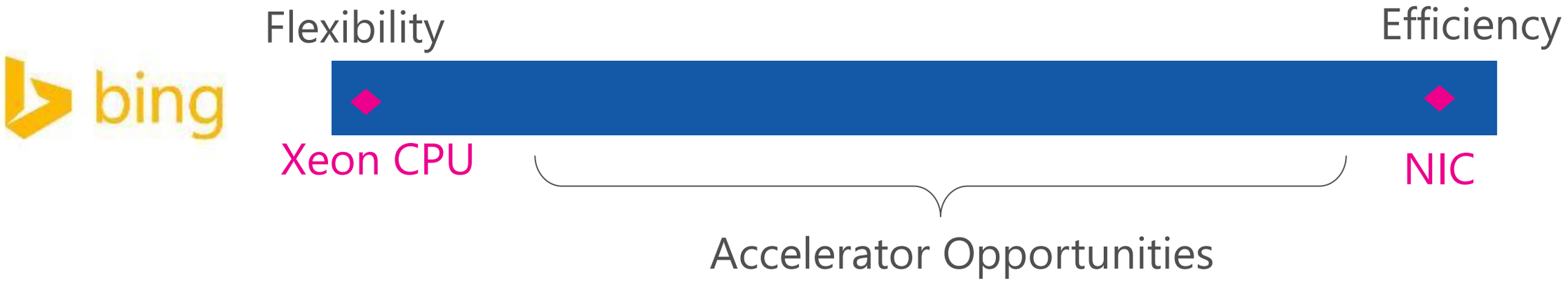
The paradox: Specialization *and* homogeneity

Efficiency via Specialization

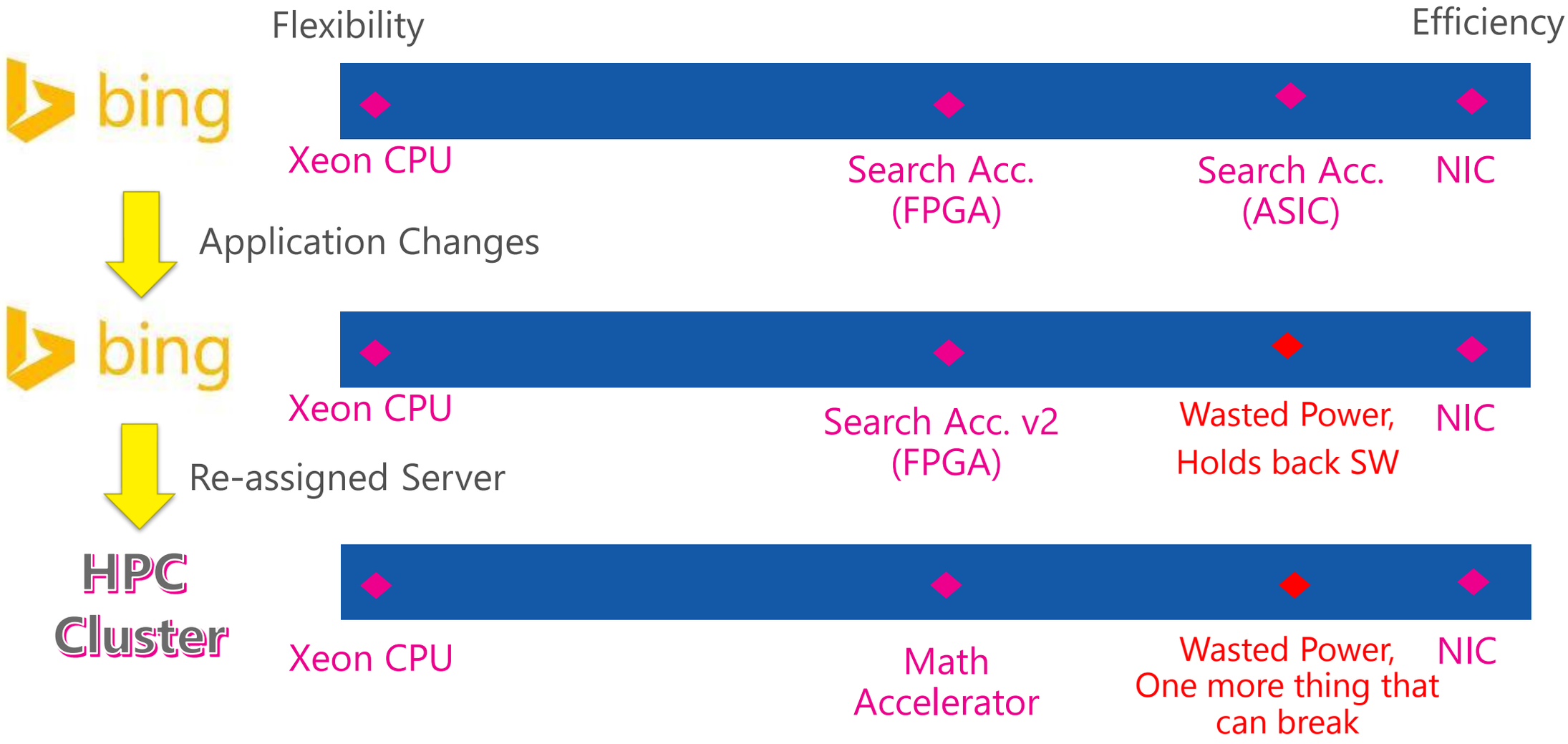


Source: Bob Broderson, Berkeley Wireless group

One Application's Accelerator



One Application's Accelerator



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

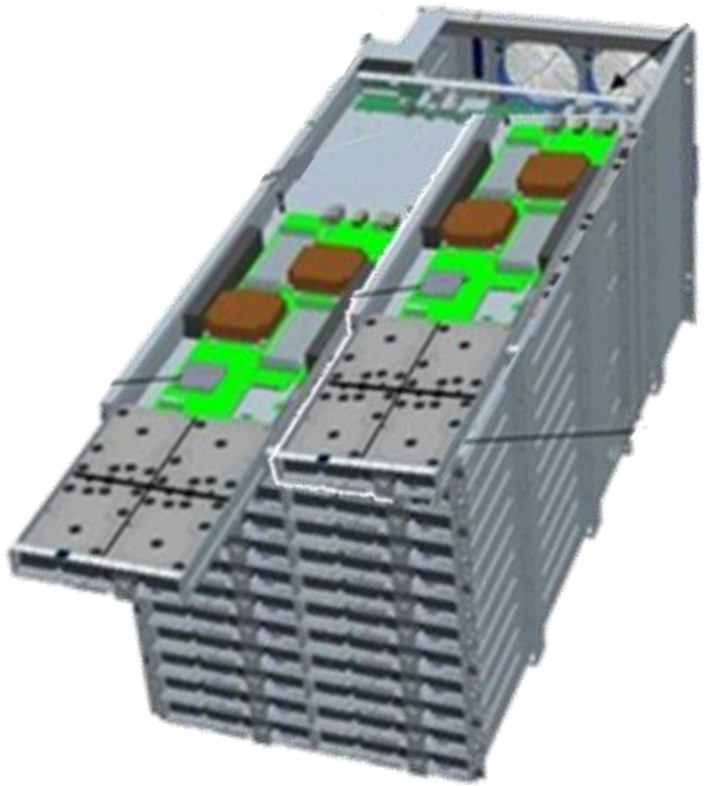
<10% Power Draw
(25W max, all from PCIe)

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



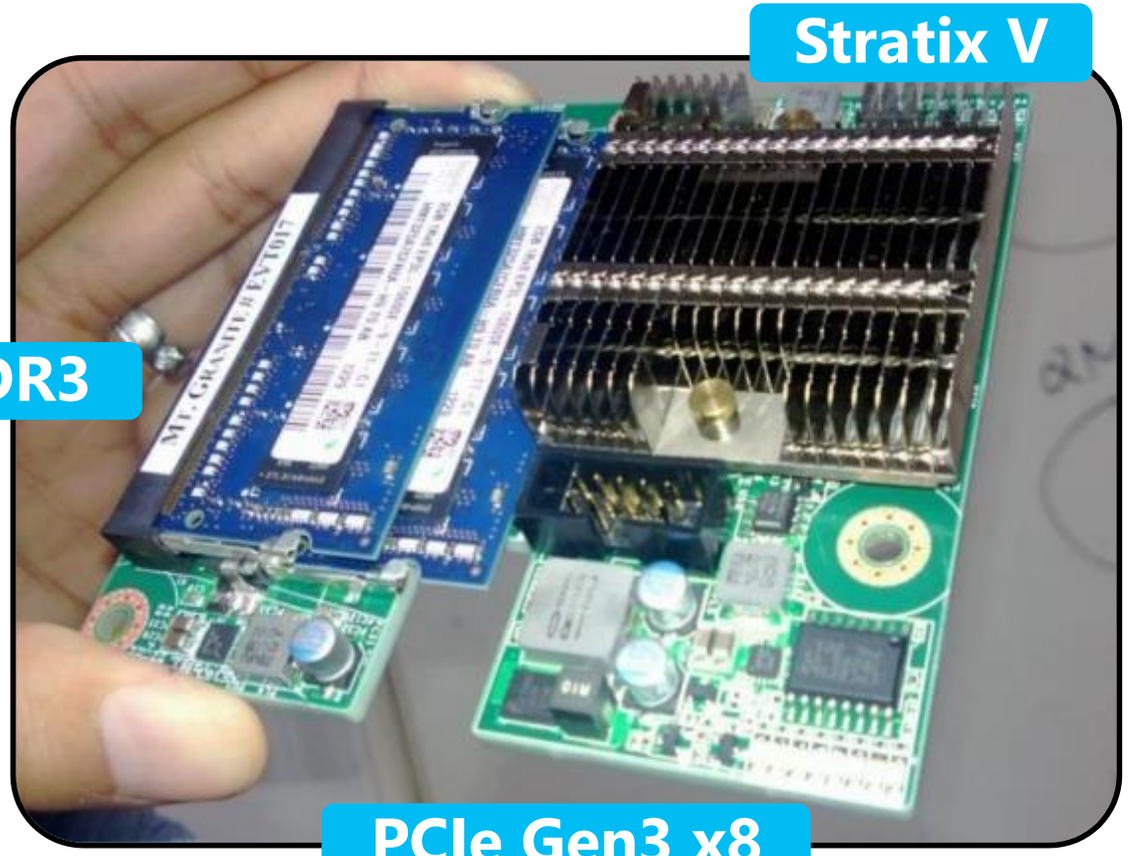
http://www.globalfoundationservices.com/posts/2014/january/27/microsoft-contributes-cloud-server-specification-to-open-compute-project.aspx

Microsoft Open Compute Server

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

Catapult FPGA Accelerator Card

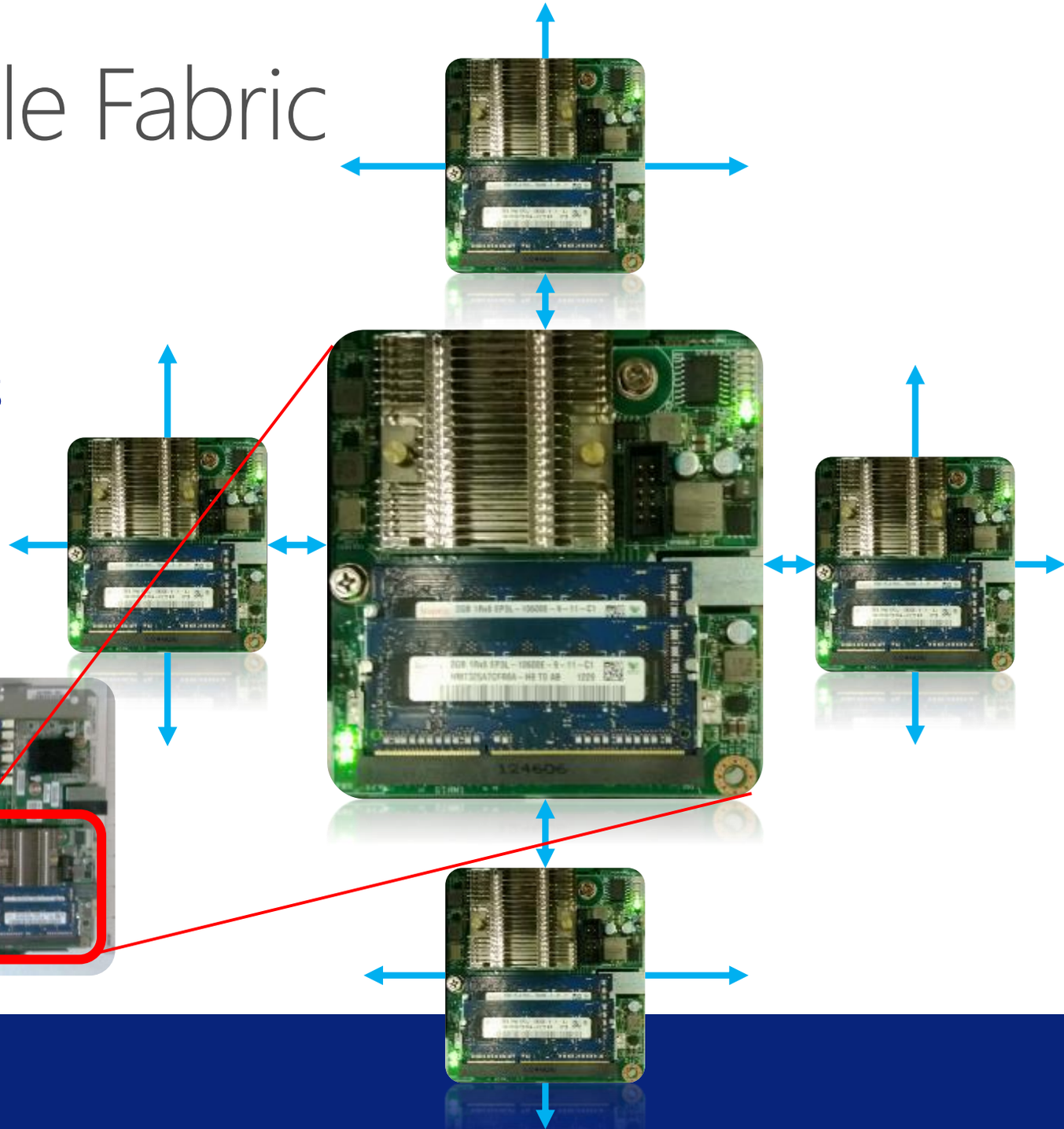
- Altera Stratix V D5
  - 172,600 ALMs
  - 2,014 M20Ks
  - 1,590 DSPs
- PCIe Gen 3 x8
- 8GB DDR3-1333
- Powered by PCIe slot
- Torus Network



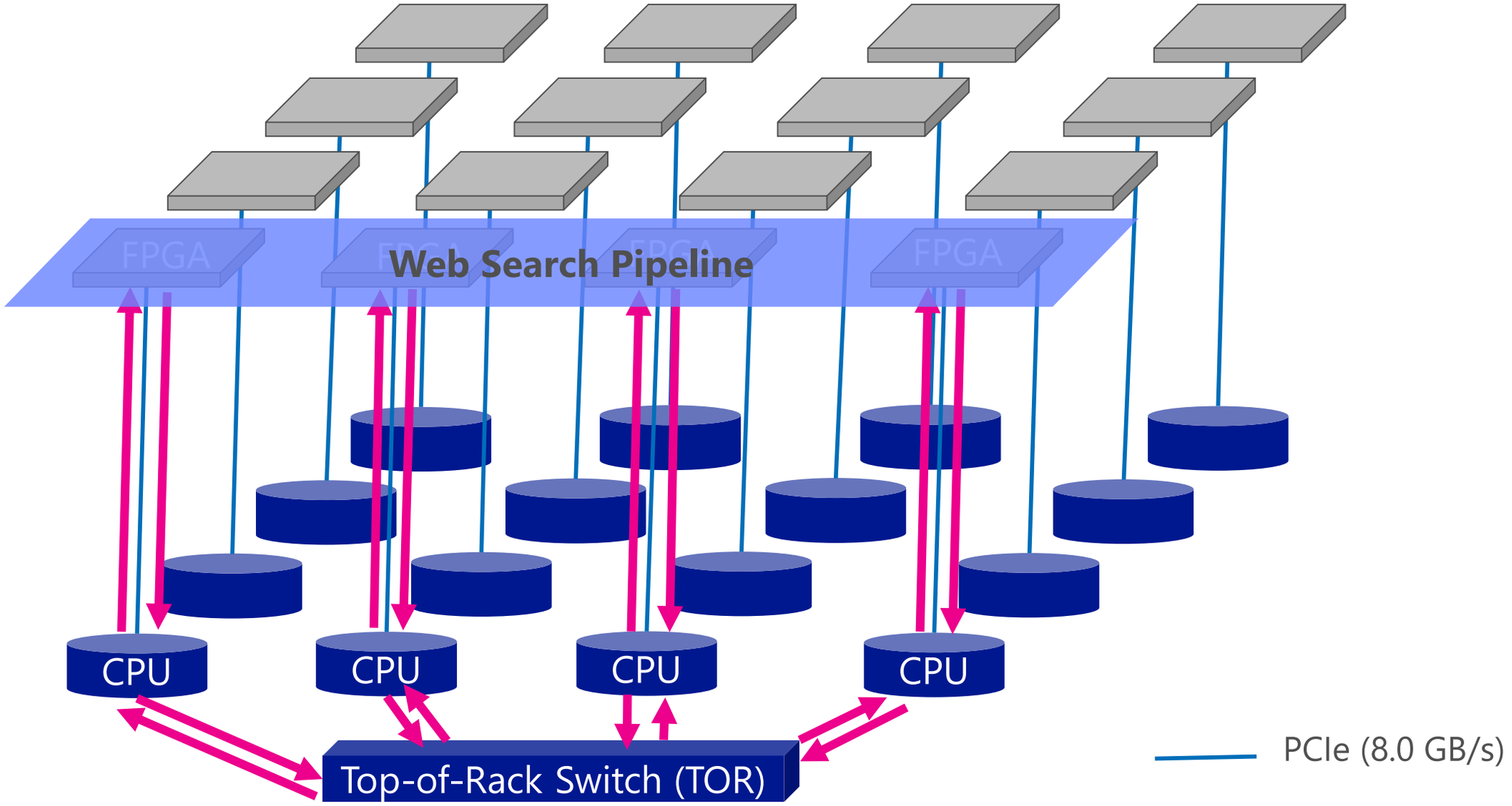
Scalable Reconfigurable Fabric

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

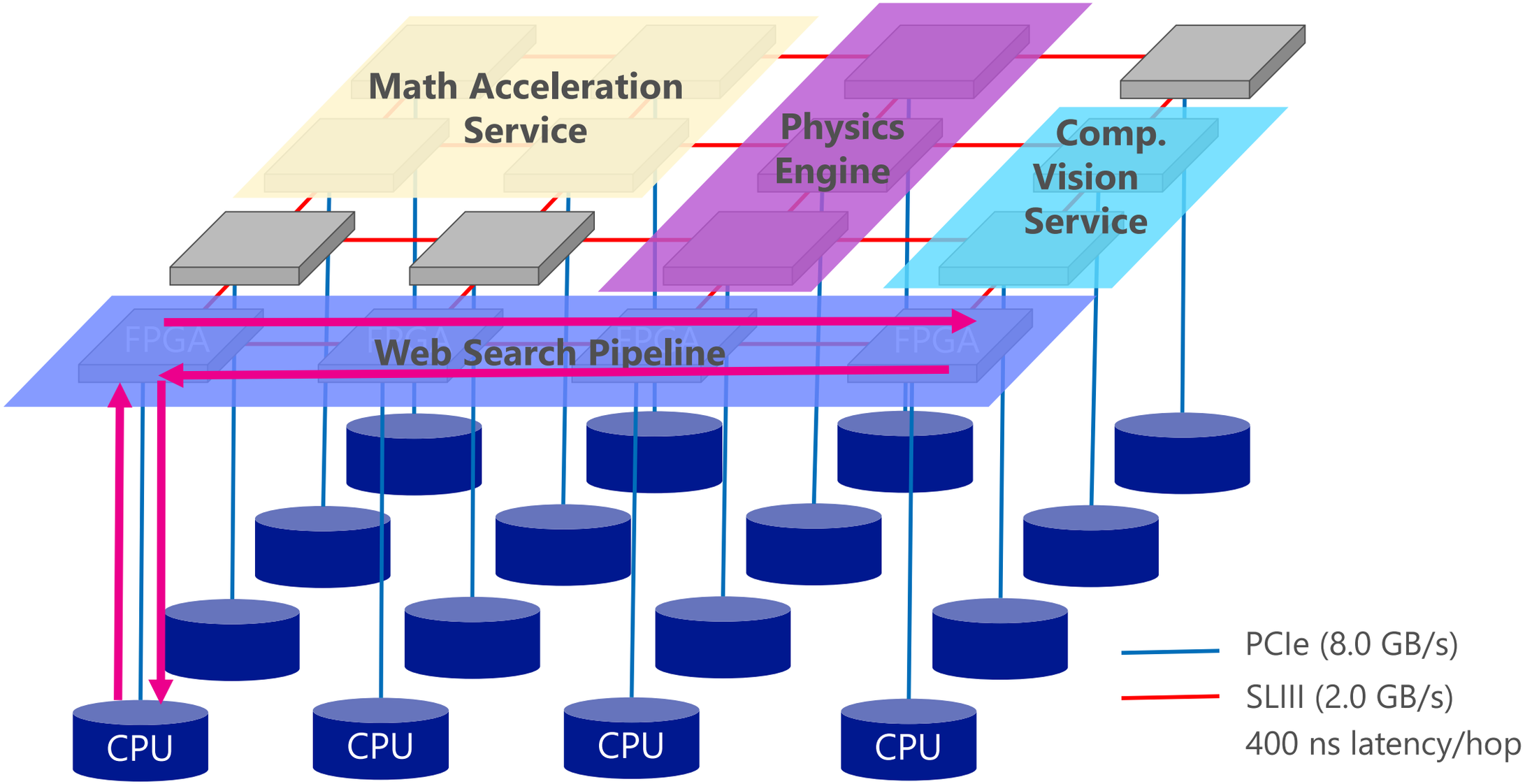
Data Center Server (1U, ½ width)



An Elastic Reconfigurable Fabric

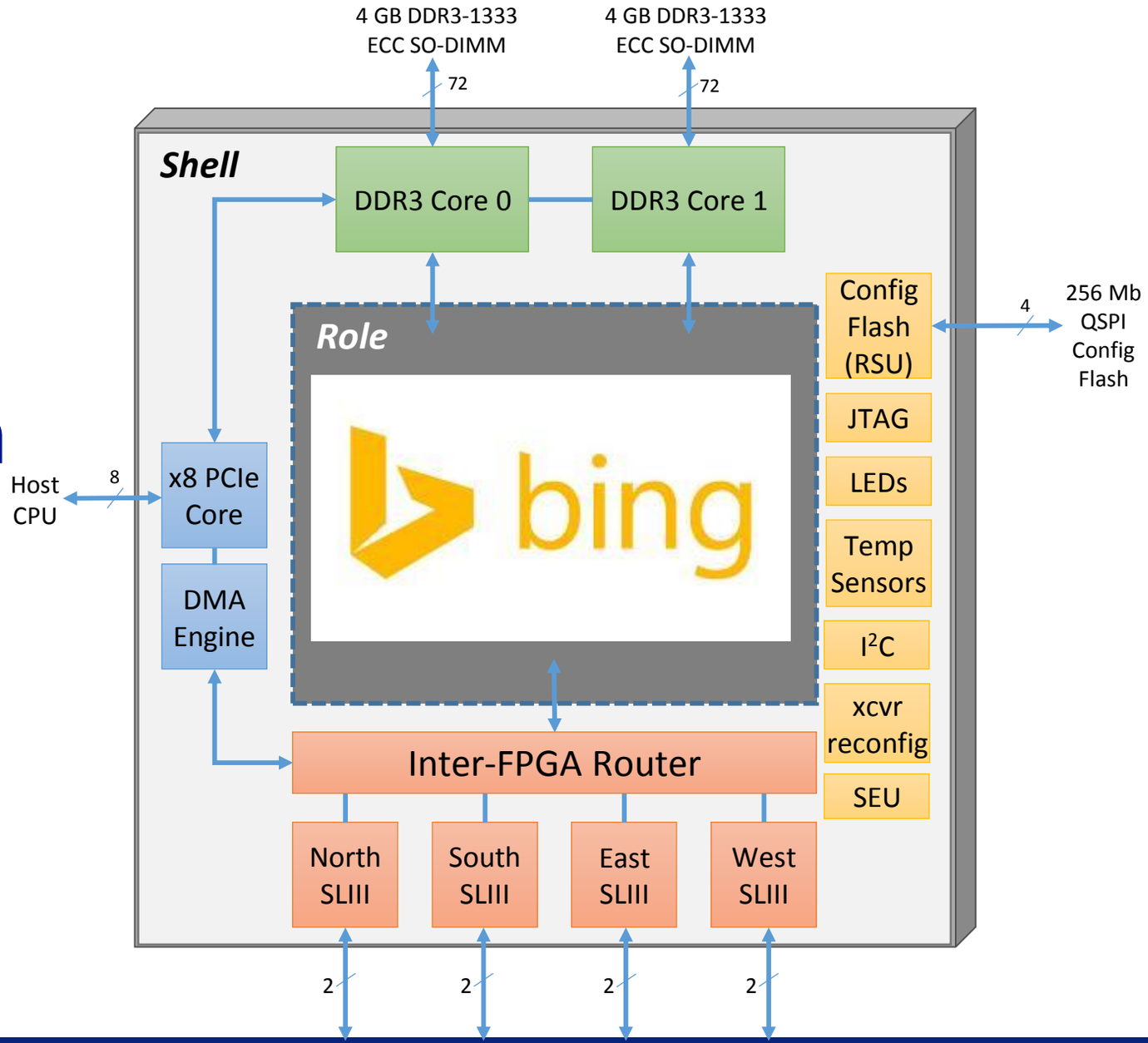


An Elastic Reconfigurable Fabric

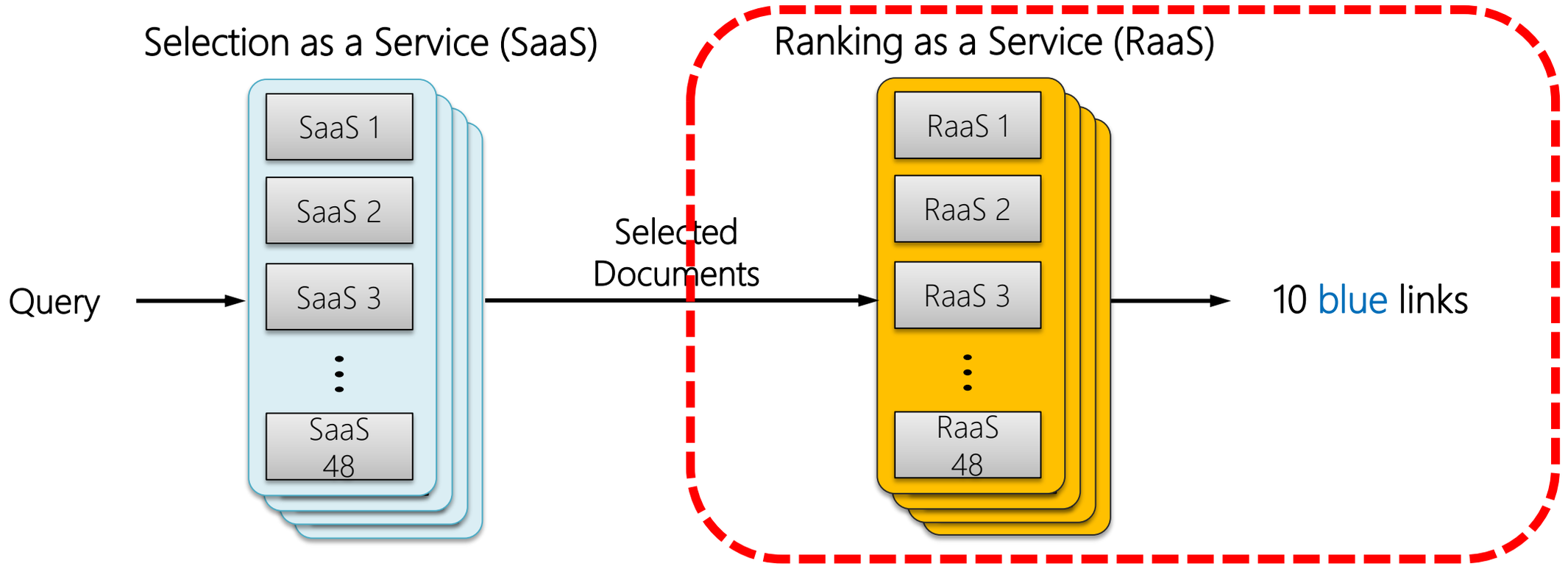


Shell & Role

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



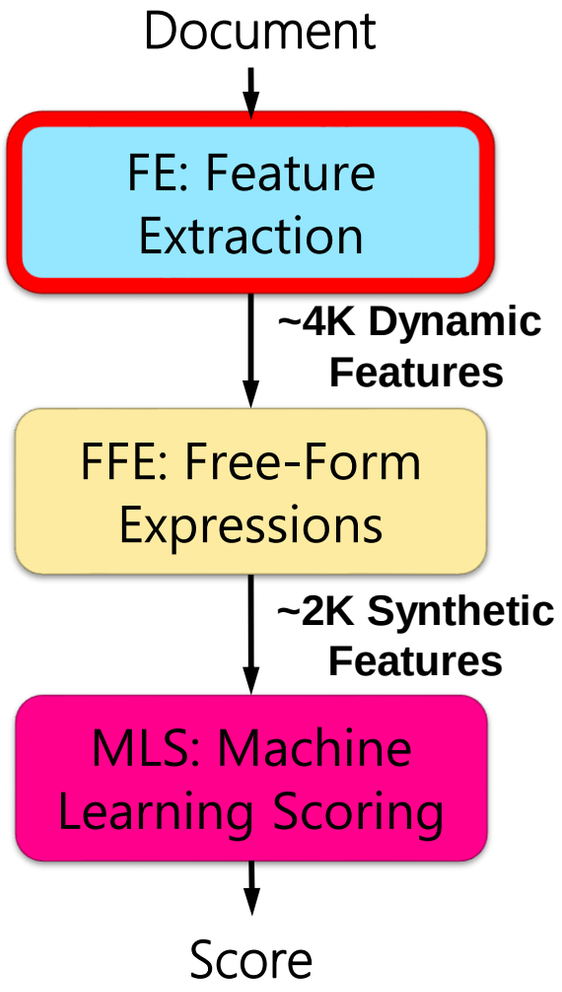
Bing Document Ranking Flow



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

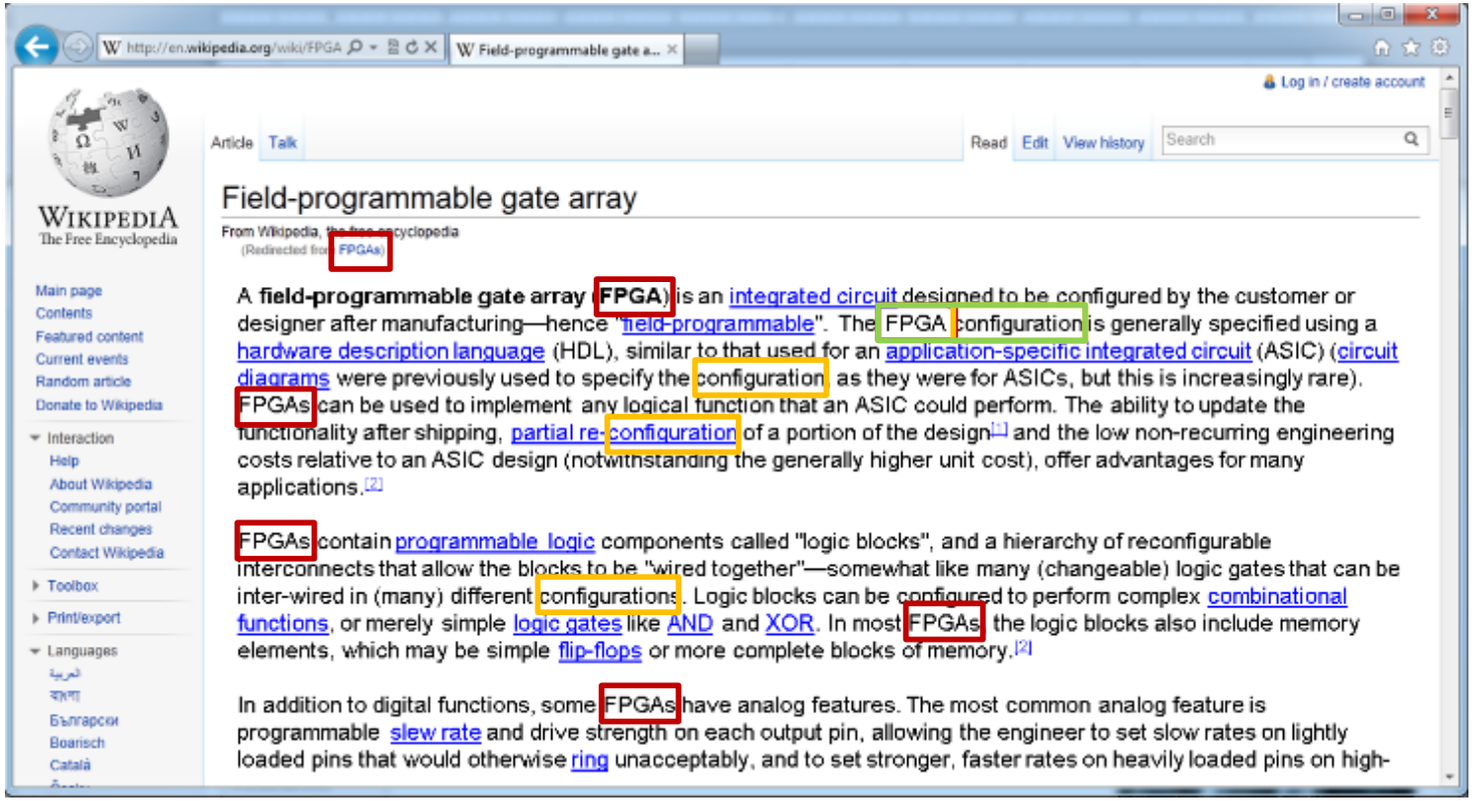
Ranking-as-a-Service (RaaS)
- Compute scores for how relevant each selected document is for the search query
- Sort the scores and return the results

FE: Feature Extraction

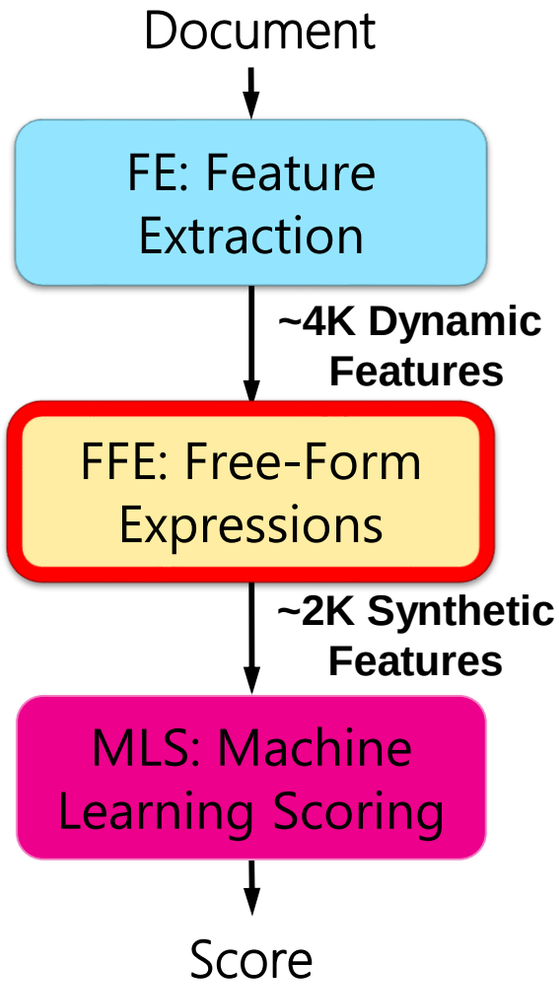


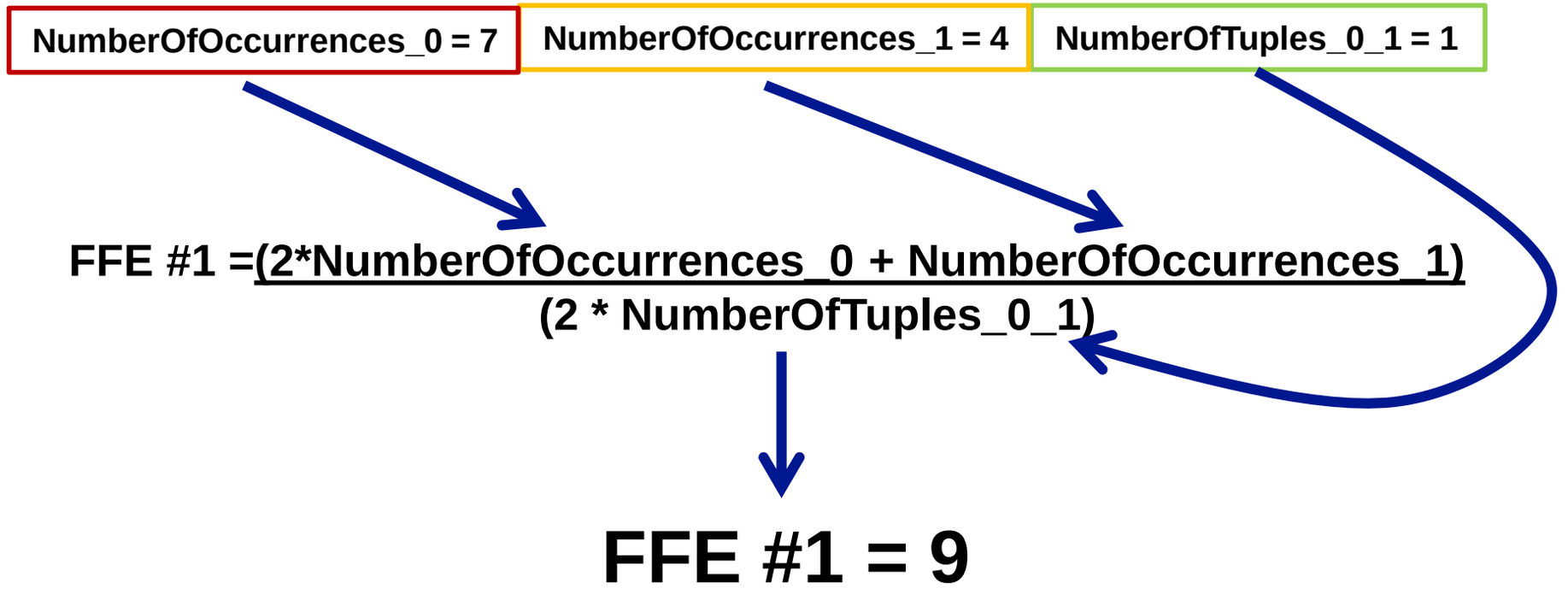
Query: "FPGA Configuration"

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

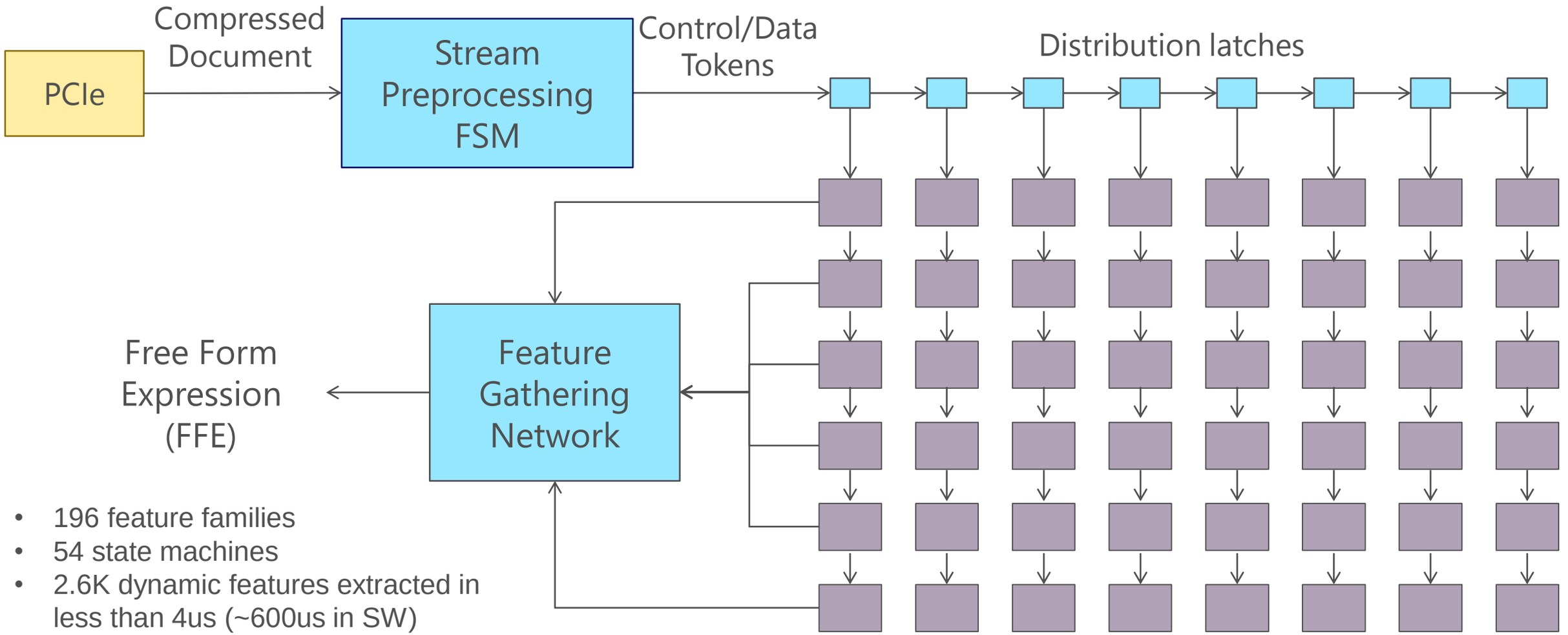


FFE: Free Form Expressions





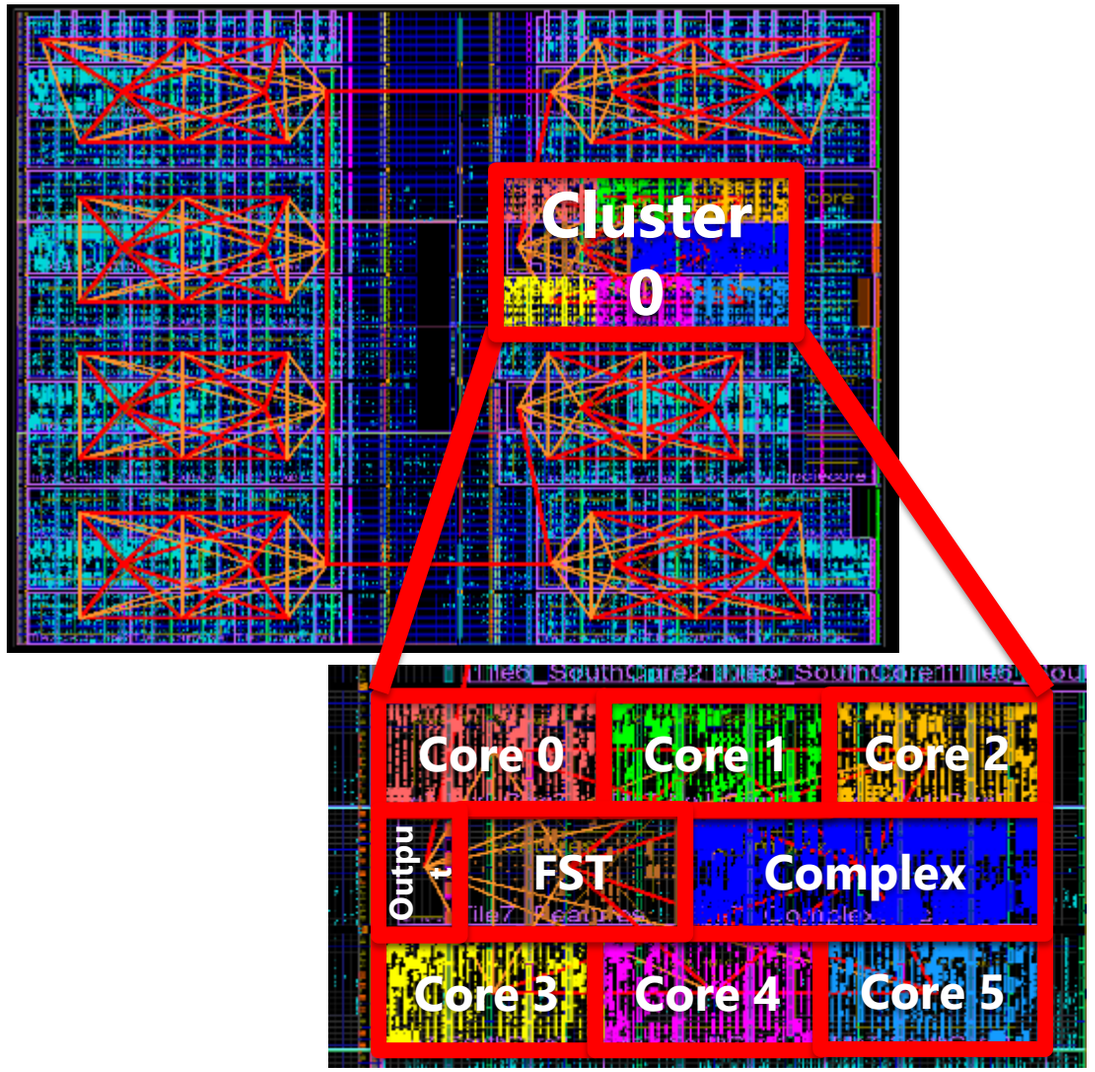
Feature Extraction Accelerator



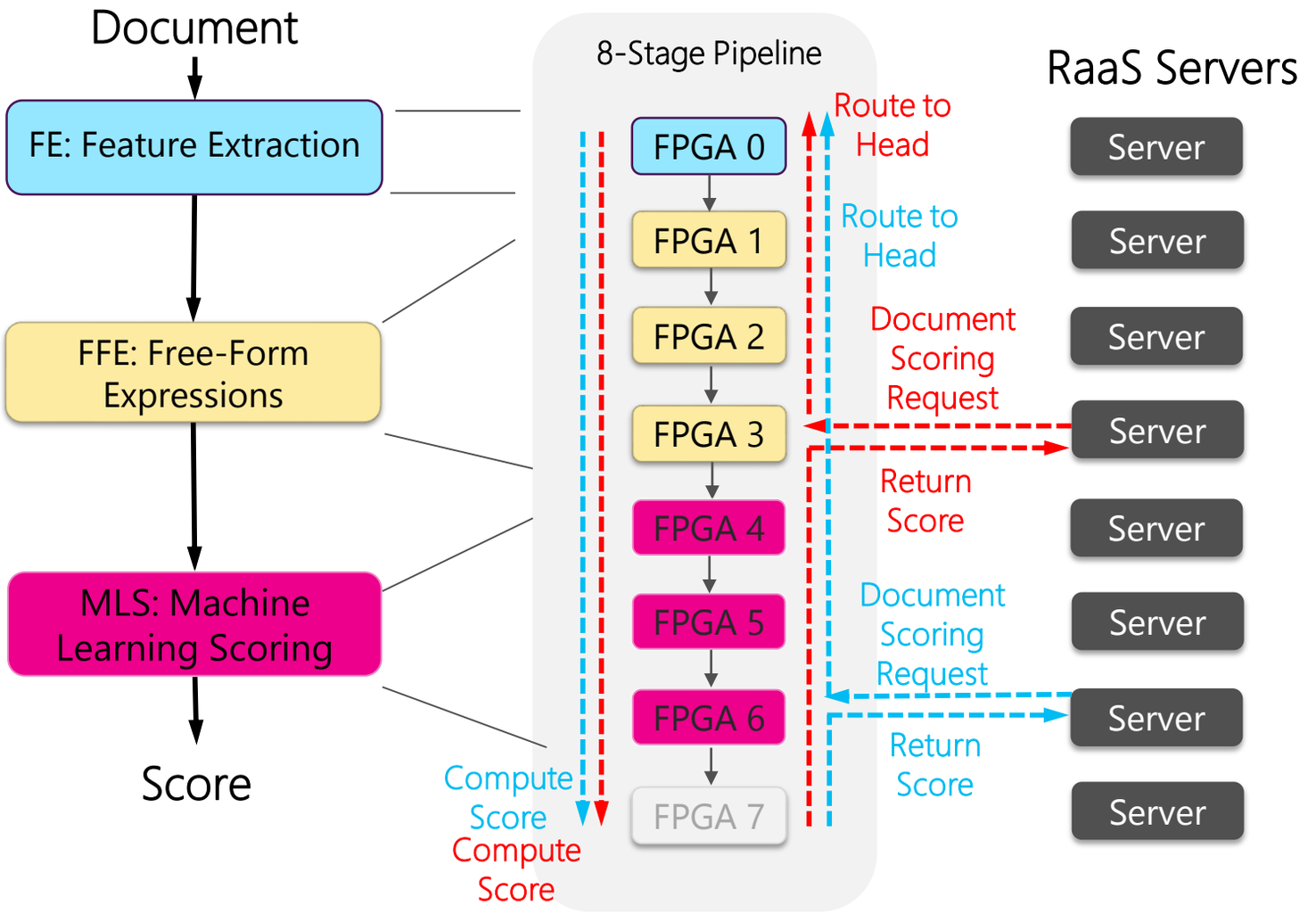
- 196 feature families
- 54 state machines
- 2.6K dynamic features extracted in less than 4us (~600us in SW)

FFE 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



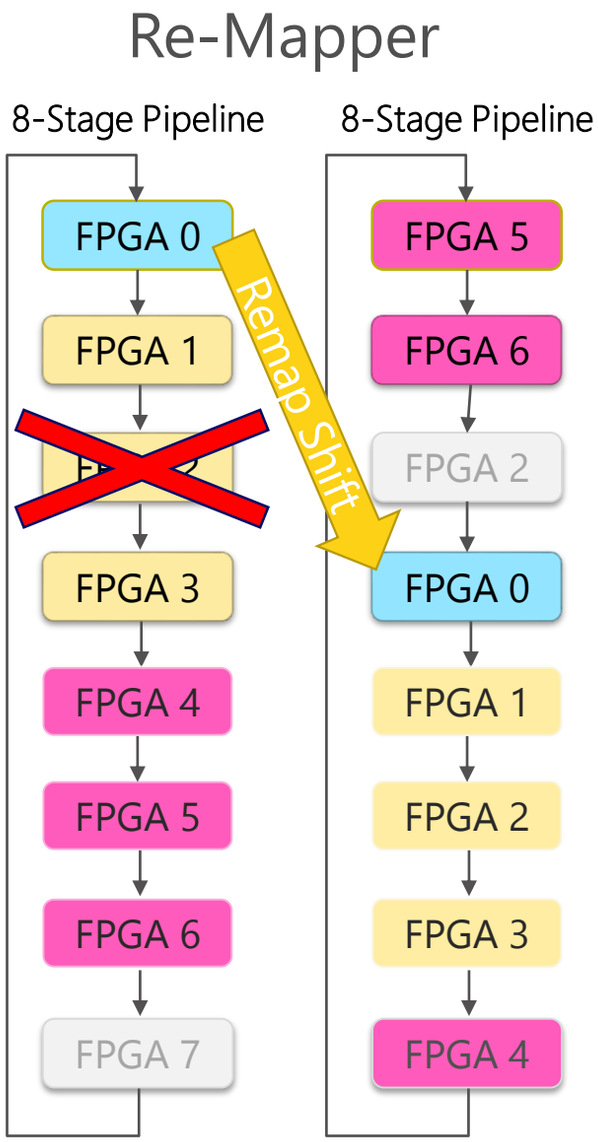
FPGA Accelerator for RaaS

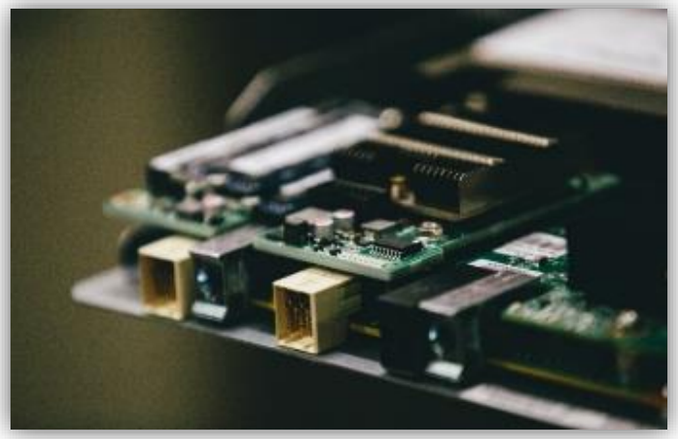


Scalable Deployment Challenges

- 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 for FPGA configuration
  - SEU scrubbing of the FPGA
  - Wiring and board check at integration

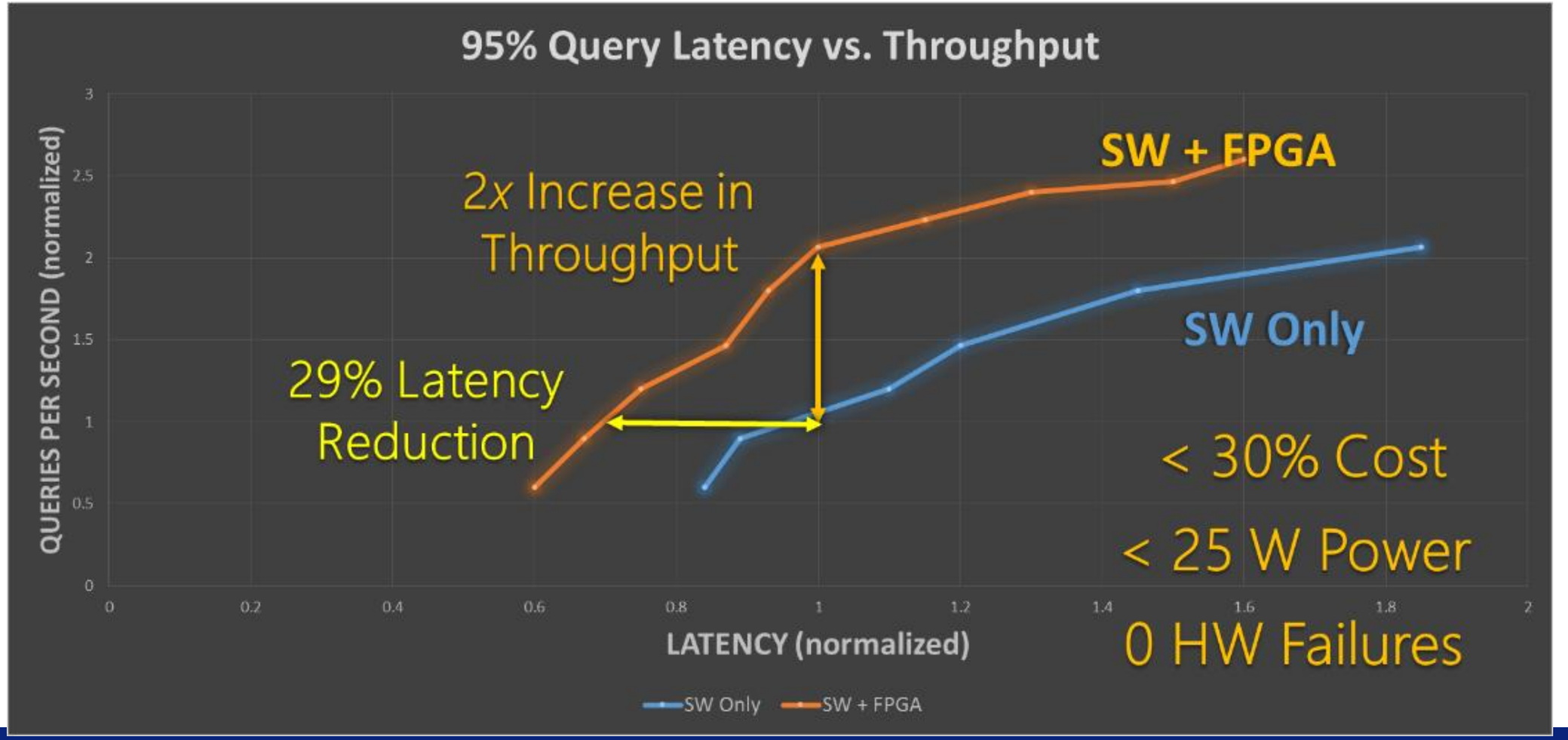




1,632 Server Pilot Deployed in a Production Datacenter

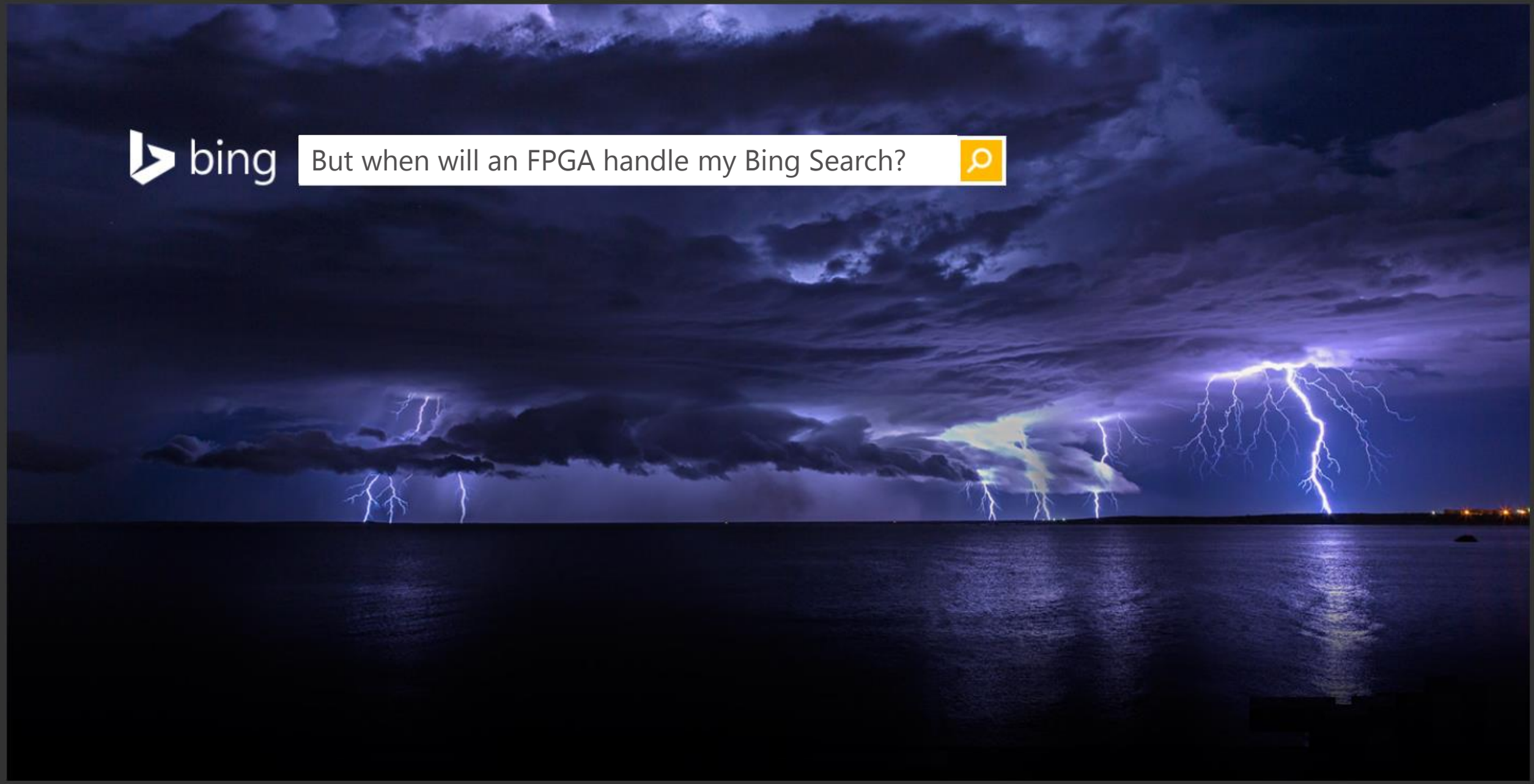
Accelerating Large-Scale Services – Bing Search

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



Conclusions

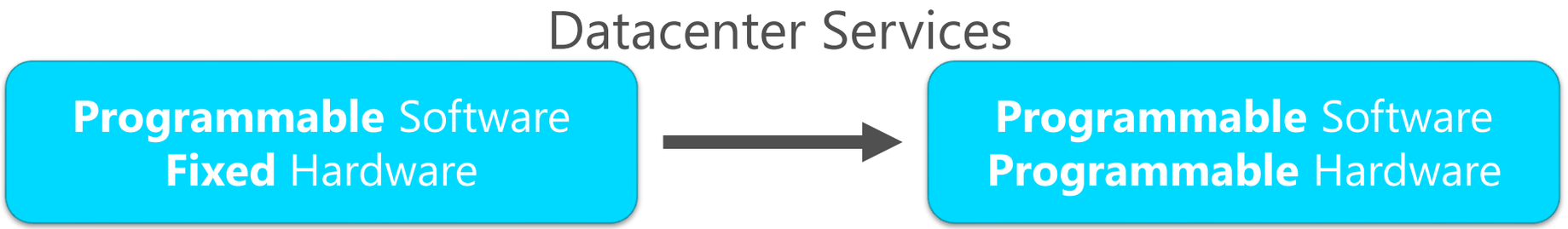
- Hardware specialization is a (the?) way to gain efficiency and performance
- An FPGA fabric offers a flexible, elastic pool of resources to accelerate services
- Results for one service: ½ the number of ranking servers, lower latency, lower variance
- Proven scalability, proven resilience, and huge potential for future apps


IMAGES VIDEOS MAPS NEWS SEARCH HISTORY MORE MSN OUTLOOK.COM
Make Bing my homepage
2
250
Sign in
bing
But when will an FPGA handle my Bing Search?
© Image Credits
© 2014 Microsoft | Privacy and Cookies | Legal | Advertise | About our ads | Help | Feedback

"This Isn't A Toy"

- *Bing is going into production with FPGAs*

- Early 2015 – Bing will begin serving searches based on computed by the FPGA fabric

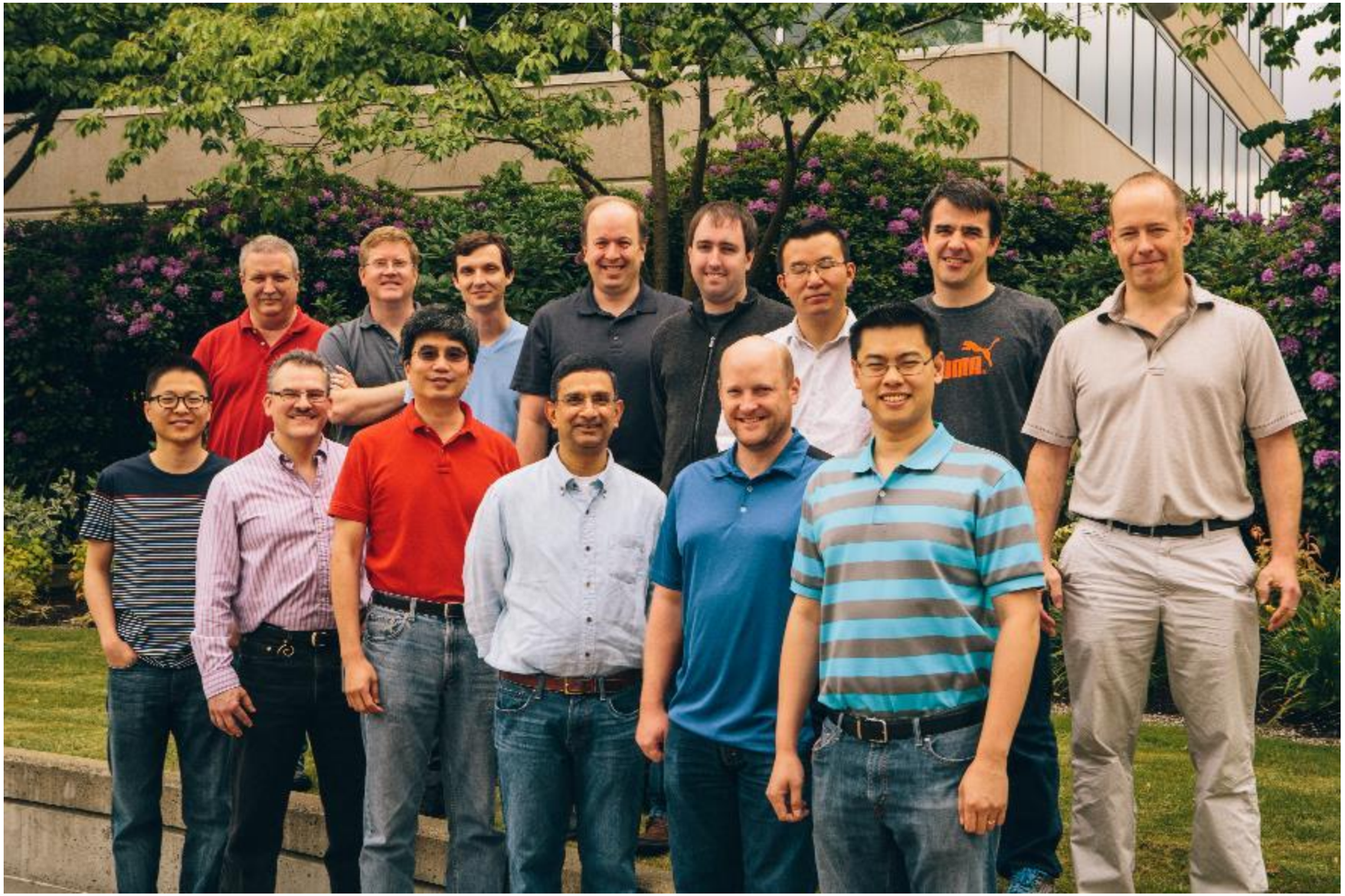


Microsoft

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

Save the planet and return
your name badge before you
leave (on Tuesday)