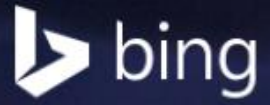




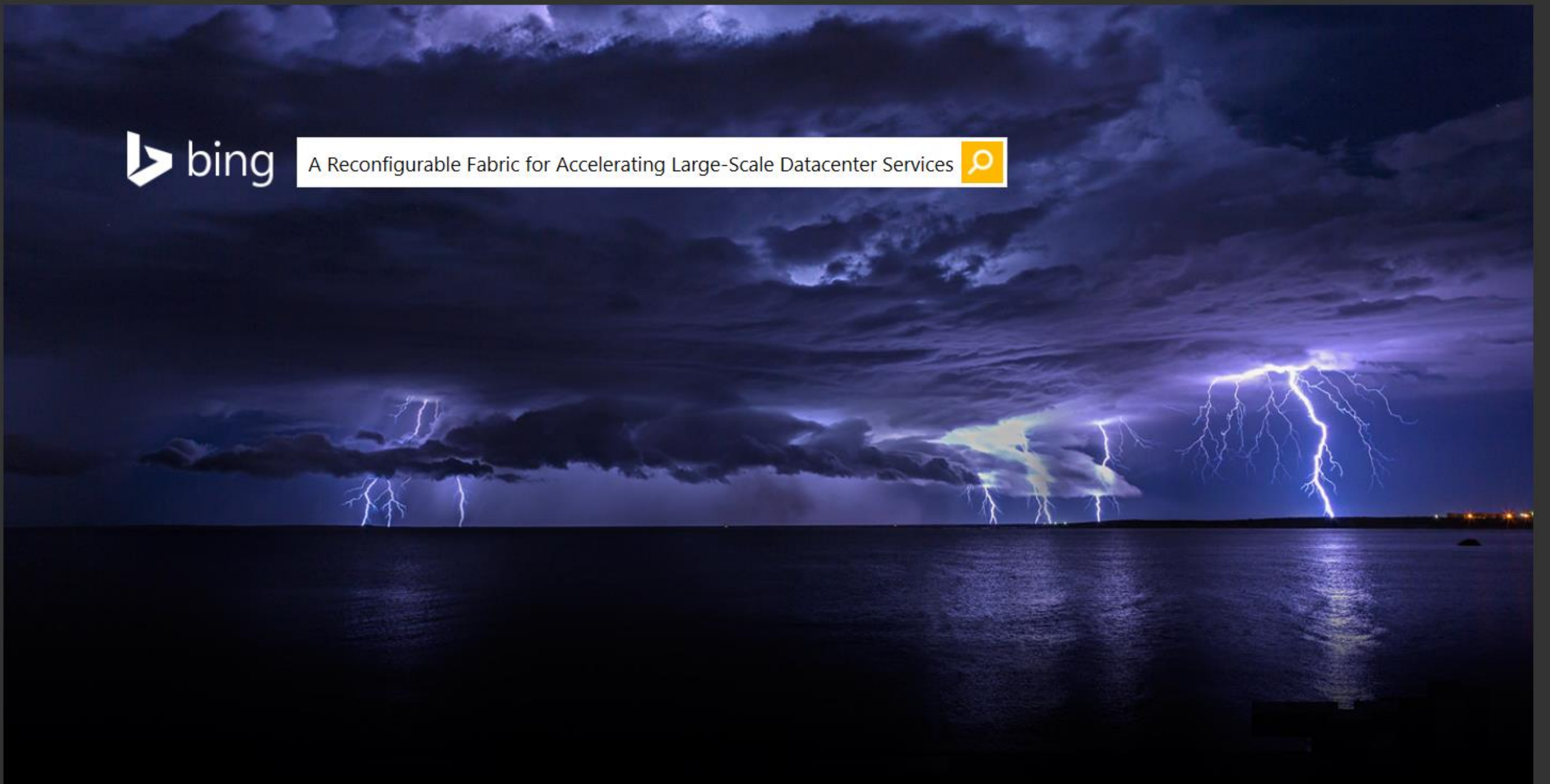
Microsoft Research

Faculty
Summit

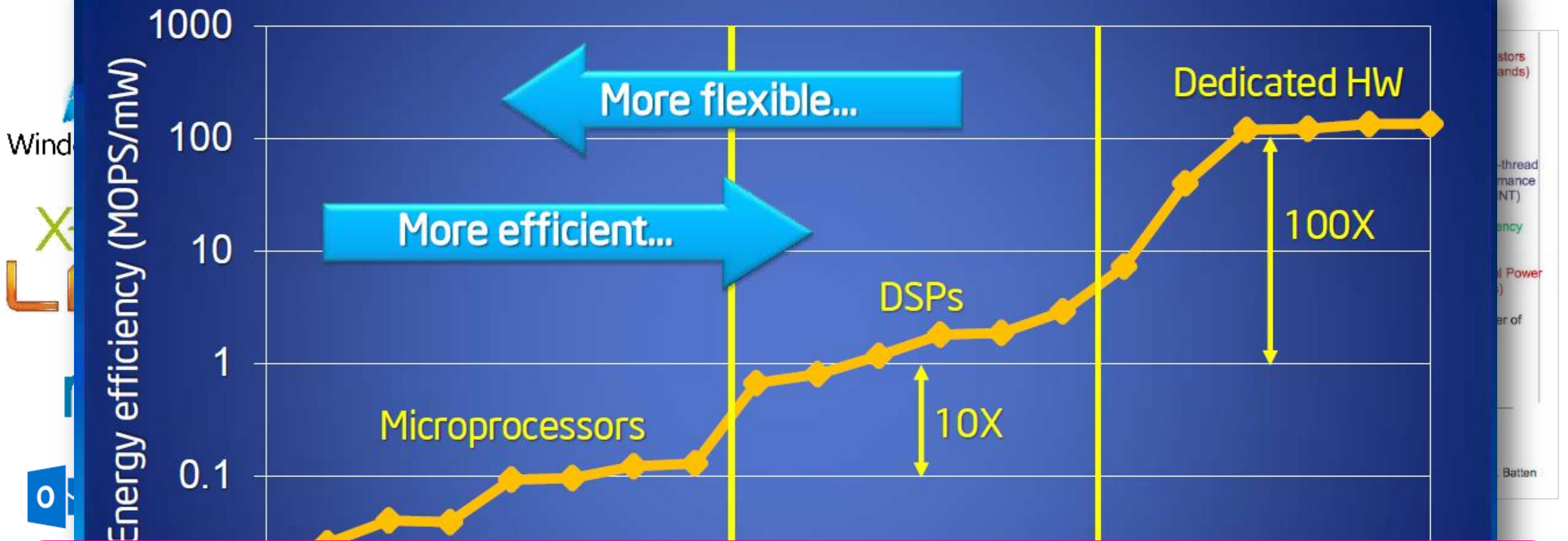
2014 15TH ANNUAL



A Reconfigurable Fabric for Accelerating Large-Scale Datacenter Services



Microsoft Cloud Services



Increase Efficiency with Hardware Specialization

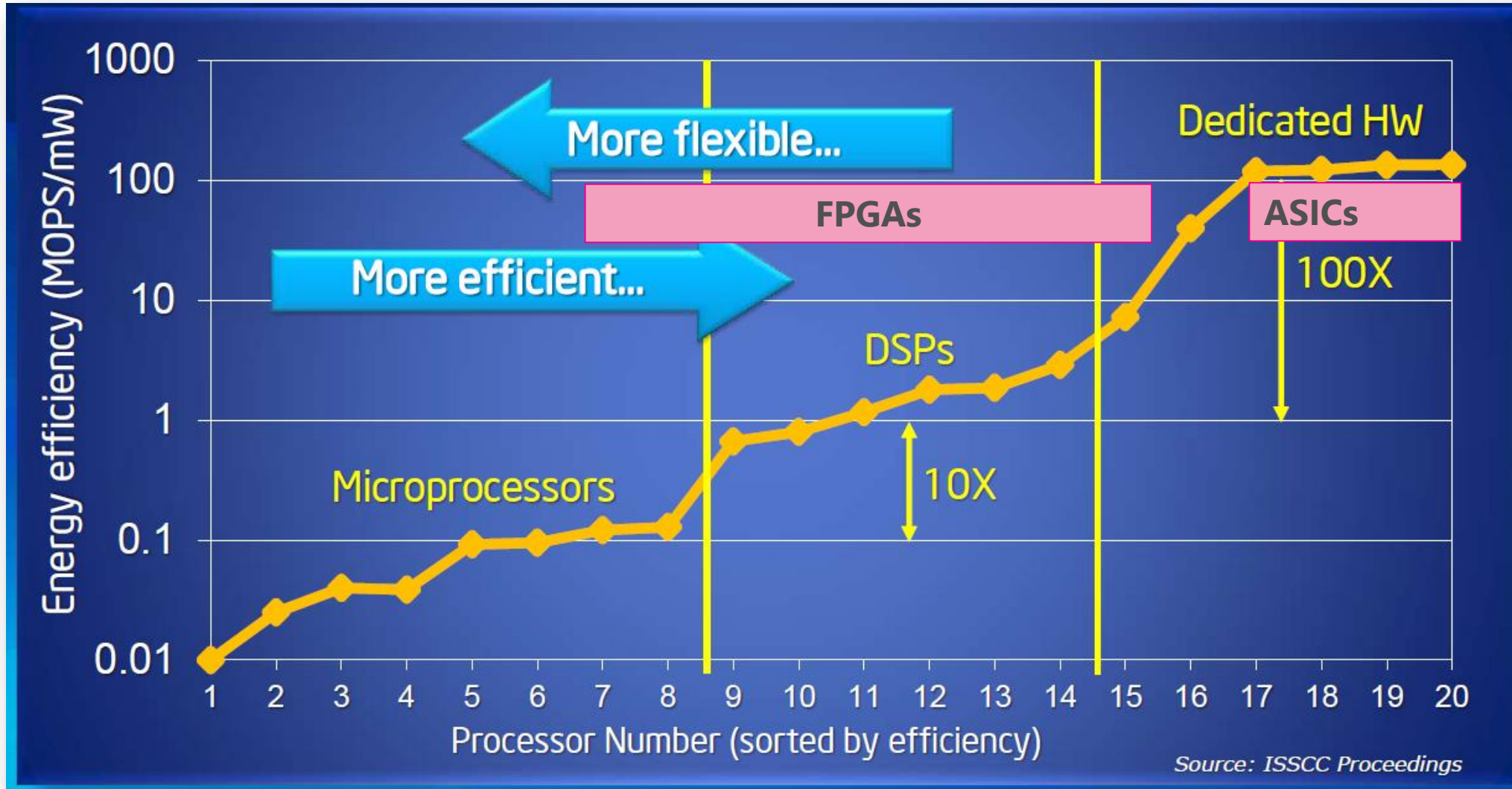
Datacenter Environment

- Software services change monthly
- Machines last 3 years, purchased on a rolling basis
- Machines repurposed $\sim 1/2$ 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



Flexibility

Efficiency



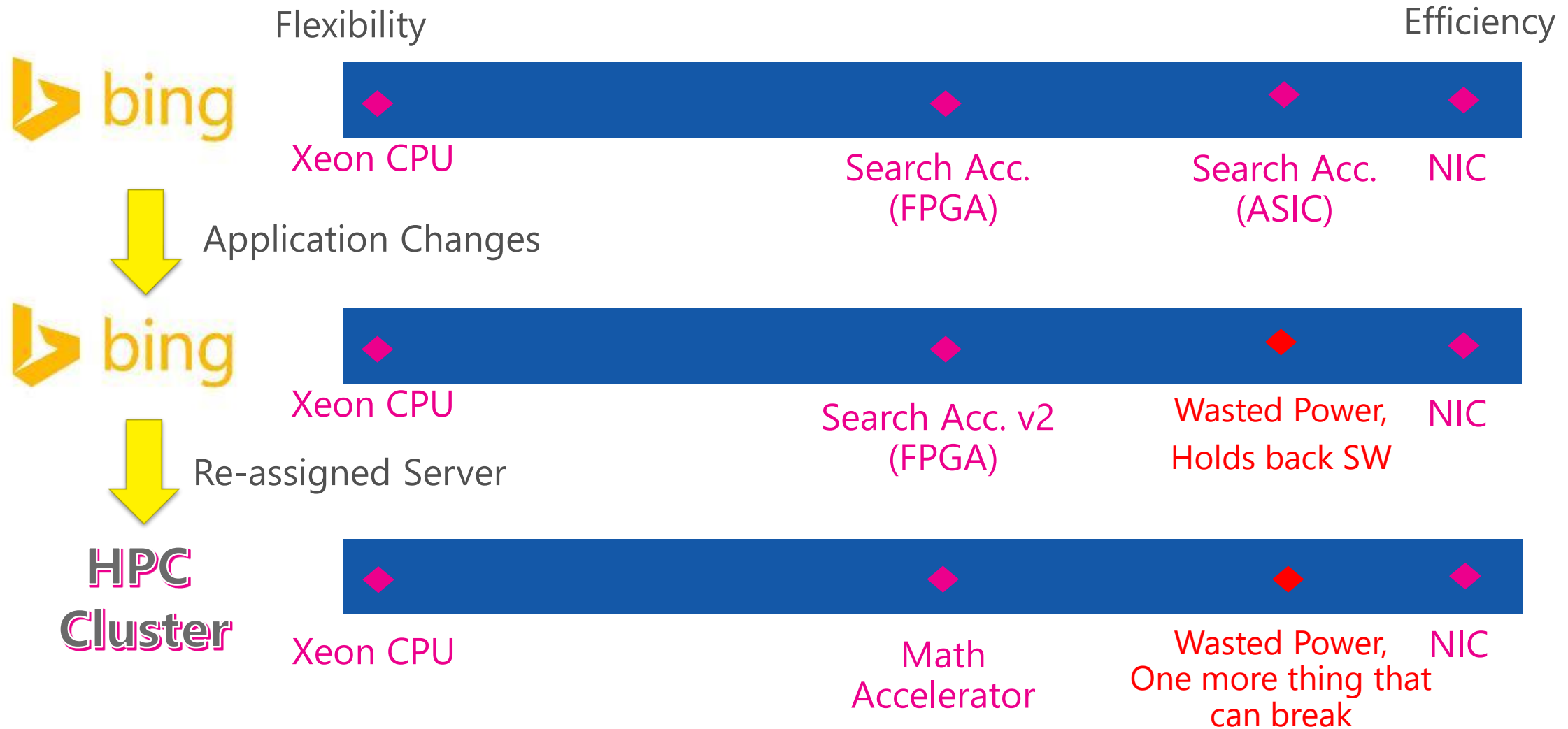
Xeon CPU

NIC



Accelerator Opportunities

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

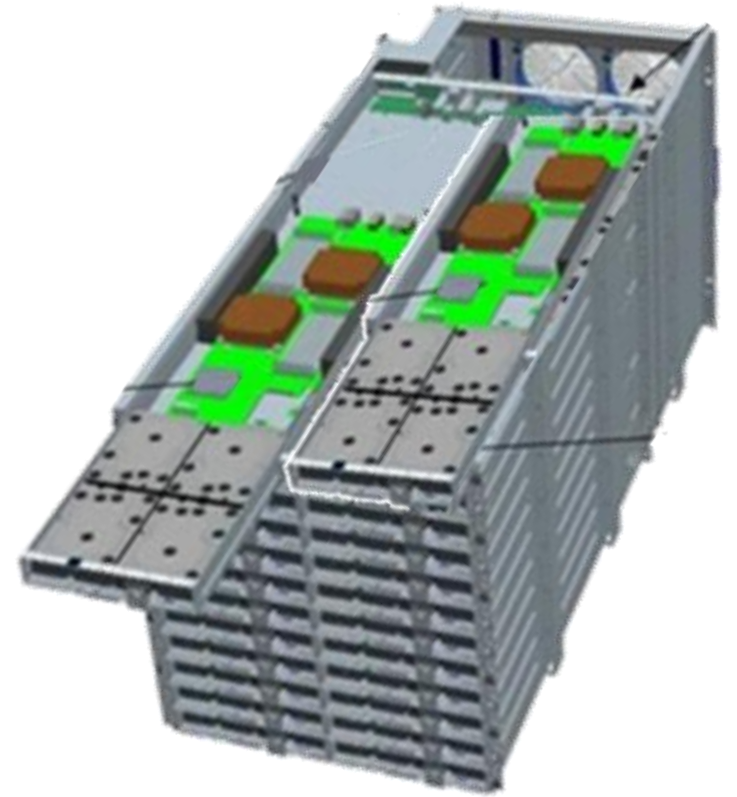
<10% Power Draw
(25W max, all from PCIe)

Don't Break Anything

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

Datacenter Servers

- Microsoft Open Compute Server
- 1U, 1/2 wide servers
- Enough space & power for 1/2 height, 1/2 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

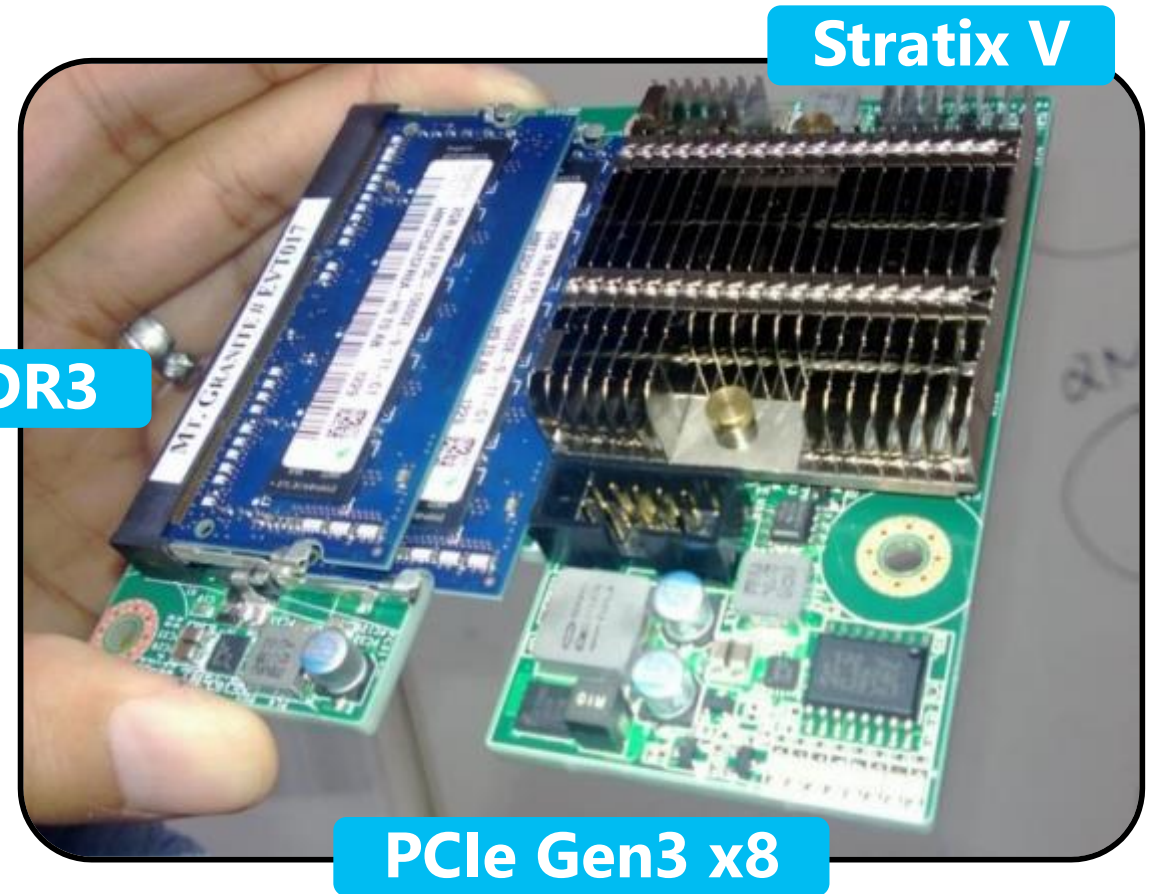


- Two 8-core Xeon 2.1 GHz CPUs
- 64 GB DRAM
- 4 HDDs, 2 SSDs
- No cable attachments to server

Air flow

Catapult FPGA Accelerator Card

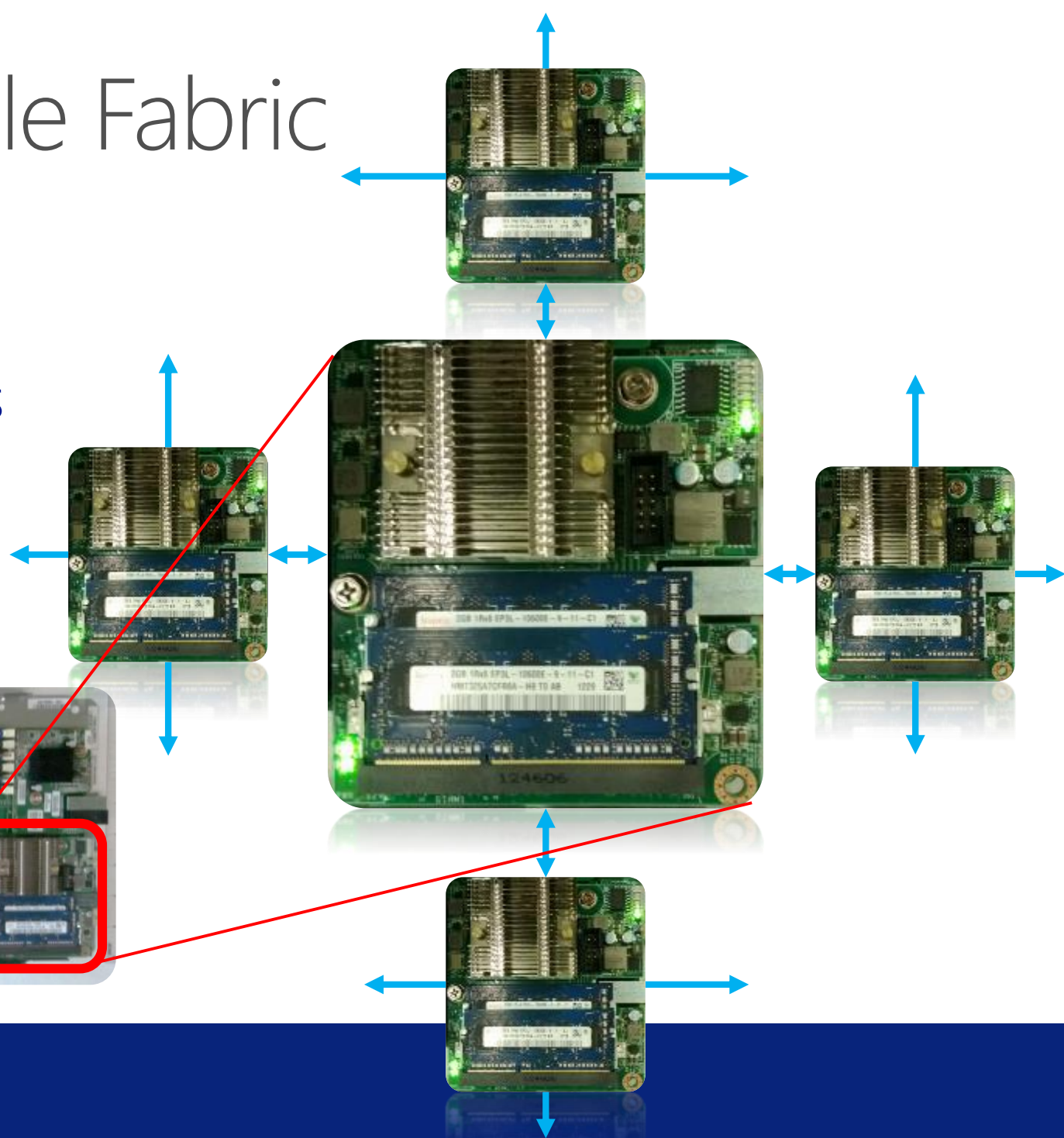
- Altera Stratix V D5
 - 172,600 ALMs
 - 2,014 M20Ks
 - 1,590 DSPs
- PCIe Gen 3 x8
- 8GB DDR3-1333
- Powered by PCIe slot
- Torus Network



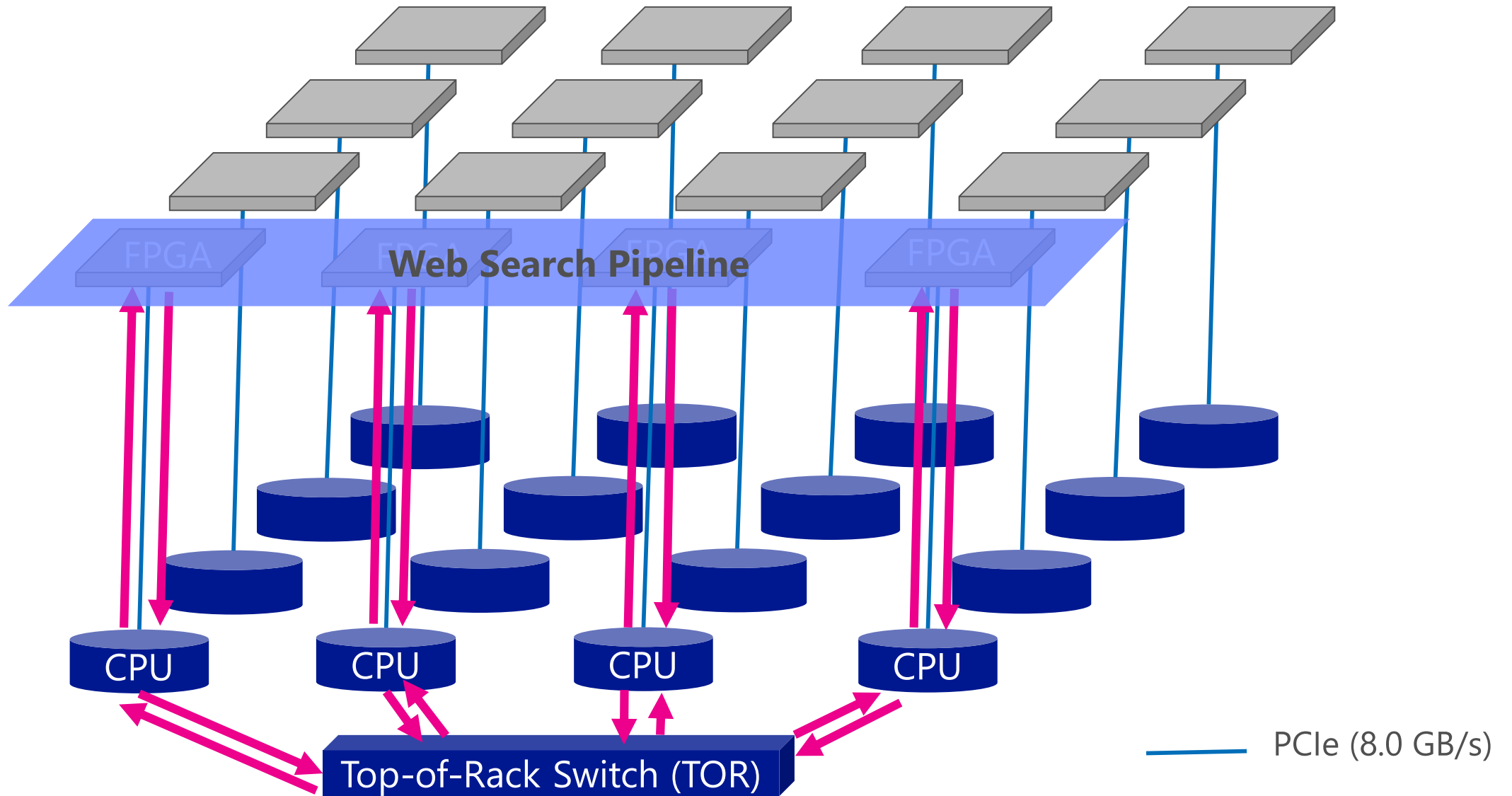
Scalable Reconfigurable Fabric

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

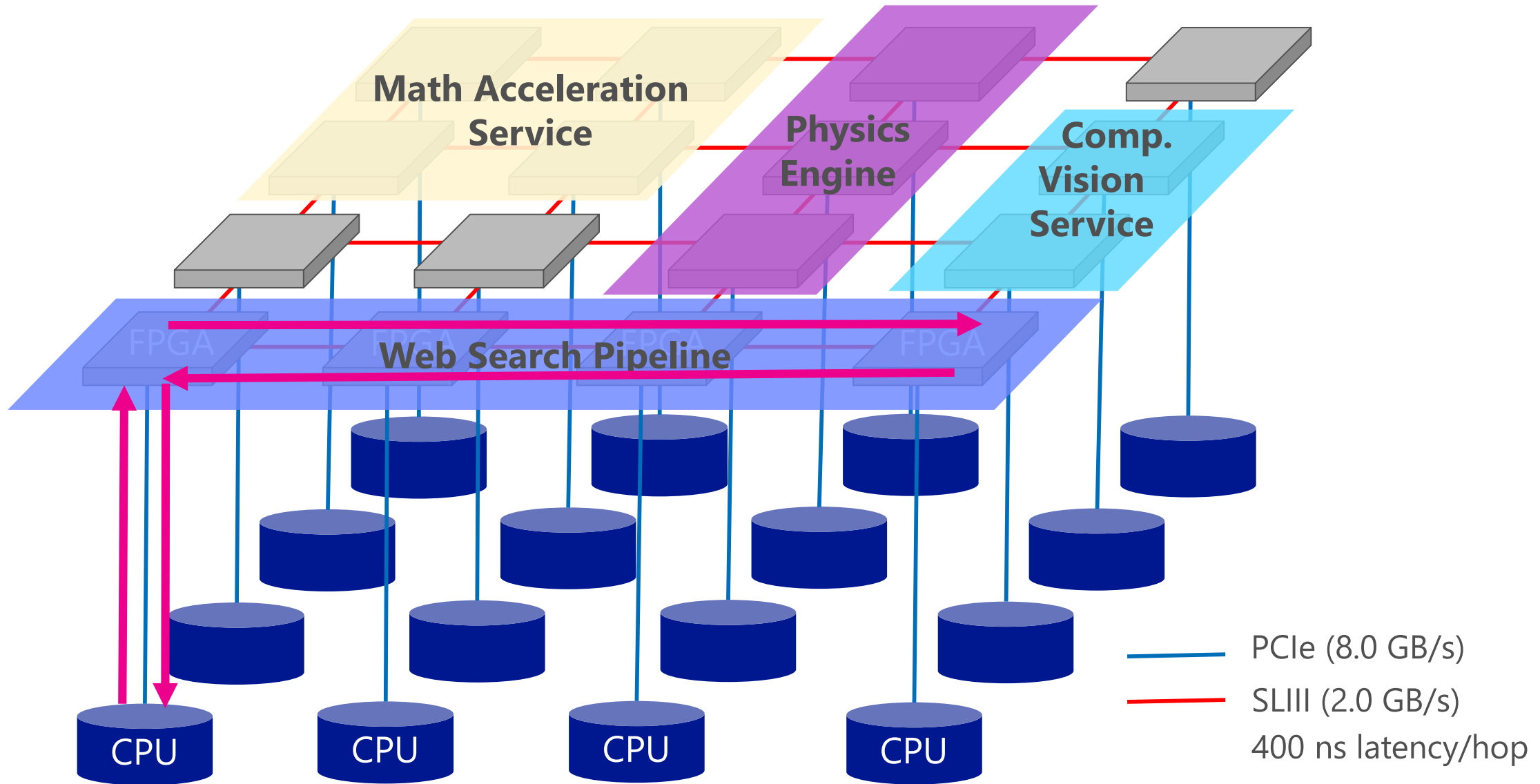
Data Center Server (1U, ½ width)



An Elastic Reconfigurable Fabric

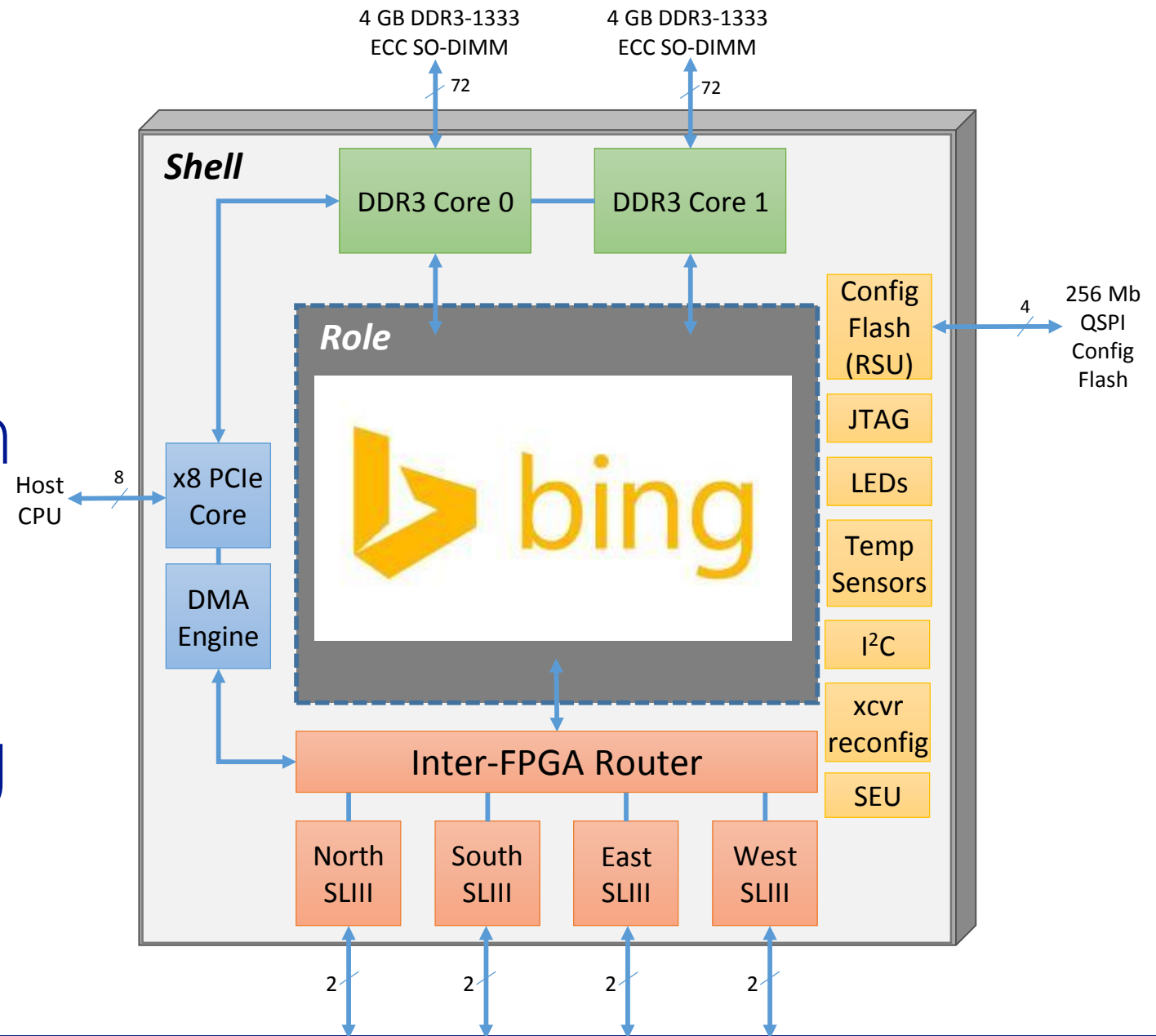


An Elastic Reconfigurable Fabric

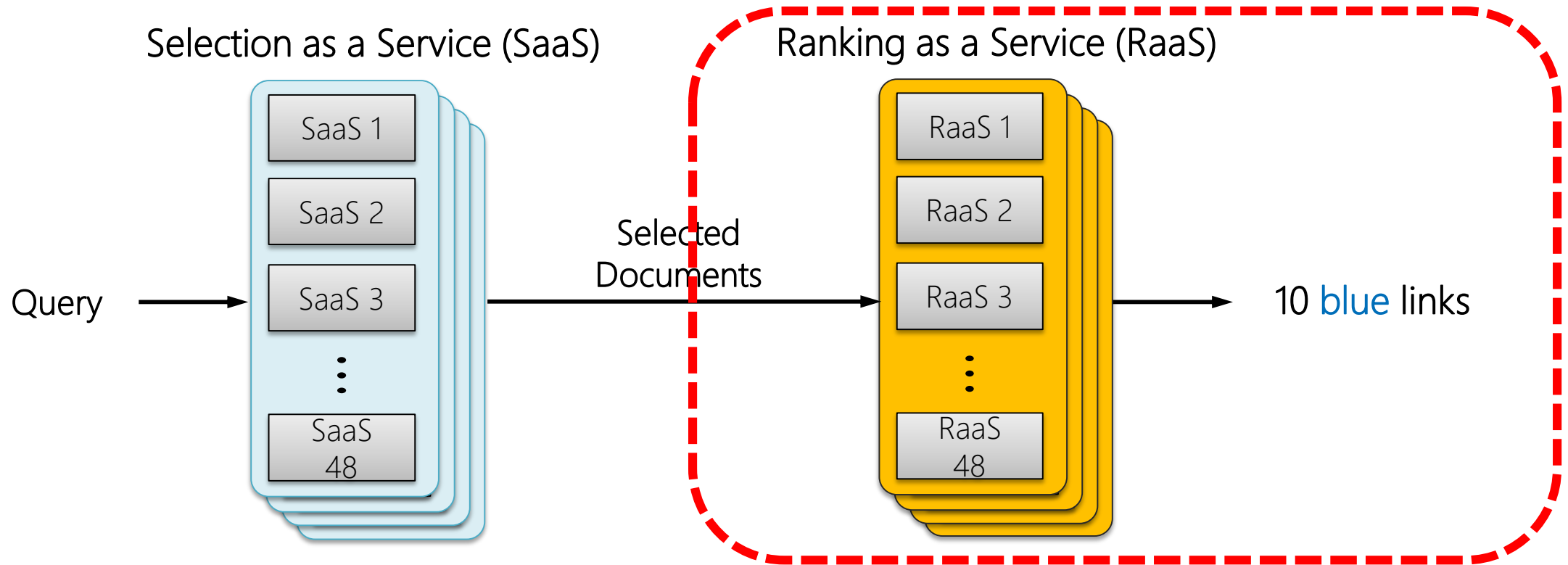


Shell & Role

- *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



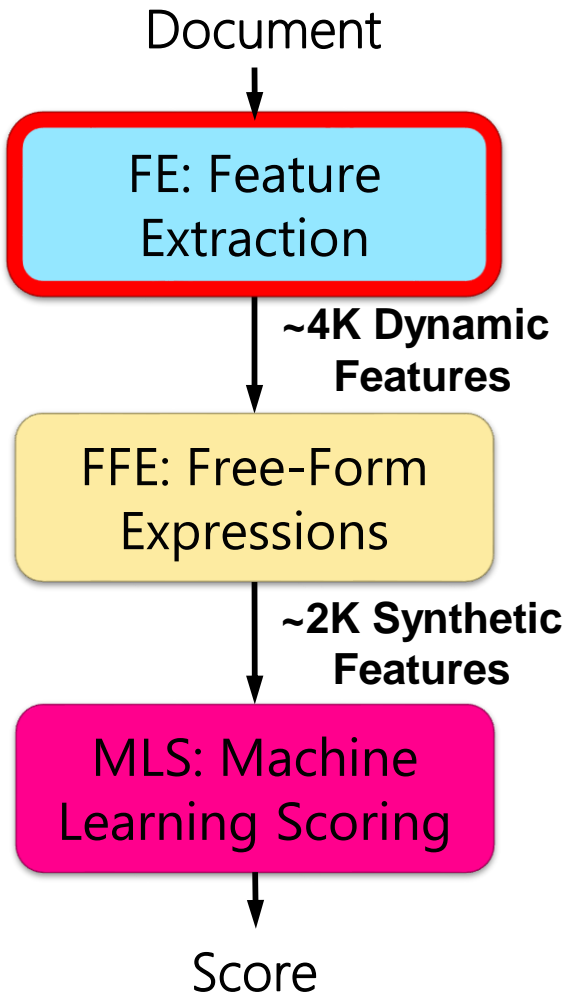
Selection-as-a-Service (SaaS)

- 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

Field-programmable gate array

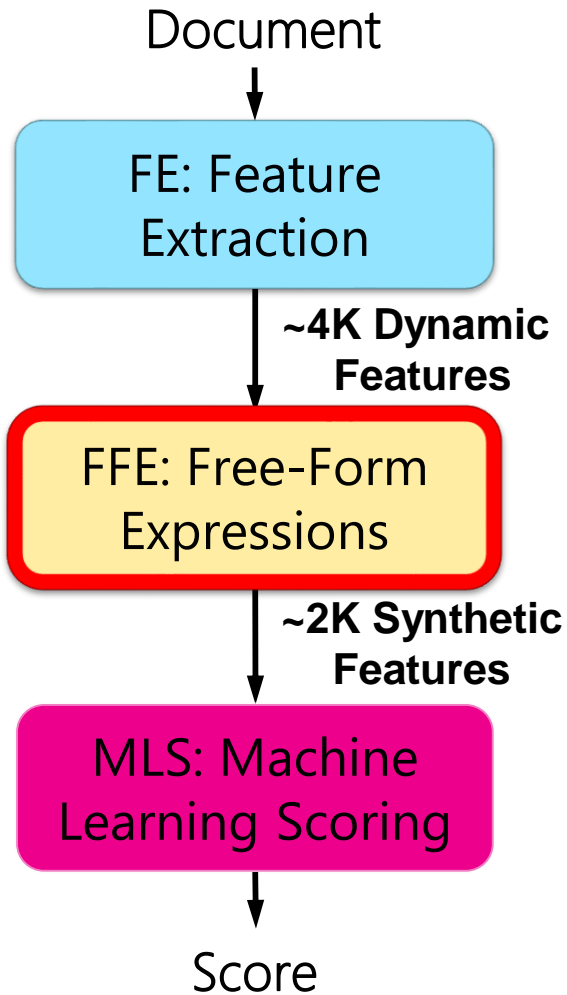
From Wikipedia, the free encyclopedia
(Redirected from **FPGAs**)

A **field-programmable gate array (FPGA)** is an **integrated circuit** designed to be configured by the customer or designer after manufacturing—hence "**field-programmable**". The **FPGA configuration** is generally specified using a **hardware description language (HDL)**, similar to that used for an **application-specific integrated circuit (ASIC)** (**circuit diagrams** were previously used to specify the **configuration** as they were for ASICs, but this is increasingly rare). **FPGAs** can be used to implement any logical function that an ASIC could perform. The ability to update the functionality after shipping, **partial re-configuration** of a portion of the design^[1] and the low non-recurring engineering costs relative to an ASIC design (notwithstanding the generally higher unit cost), offer advantages for many applications.^[2]

FPGAs contain **programmable logic** components called "logic blocks", and a hierarchy of reconfigurable interconnects that allow the blocks to be "wired together"—somewhat like many (changeable) logic gates that can be inter-wired in (many) different configurations. Logic blocks can be configured to perform complex **combinational functions**, or merely simple **logic gates** like **AND** and **XOR**. In most **FPGAs** the logic blocks also include memory elements, which may be simple **flip-flops** or more complete blocks of memory.^[2]

In addition to digital functions, some **FPGAs** have analog features. The most common analog feature is programmable **slew rate** and drive strength on each output pin, allowing the engineer to set slow rates on lightly loaded pins that would otherwise **ring** unacceptably, and to set stronger, faster rates on heavily loaded pins on high-

FFE: Free Form Expressions

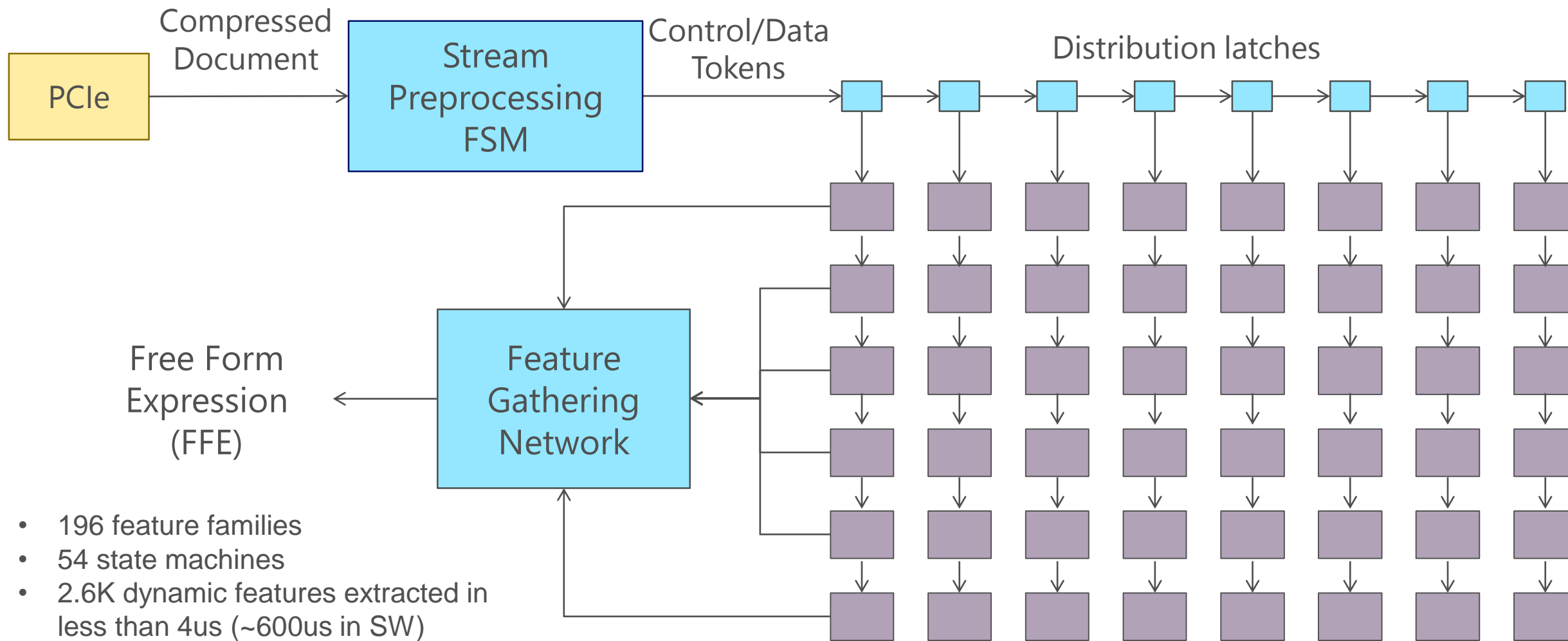


NumberOfOccurrences_0 = 7 NumberOfOccurrences_1 = 4 NumberOfTuples_0_1 = 1

$$\text{FFE \#1} = \frac{2 * \text{NumberOfOccurrences}_0 + \text{NumberOfOccurrences}_1}{2 * \text{NumberOfTuples}_{0_1}}$$

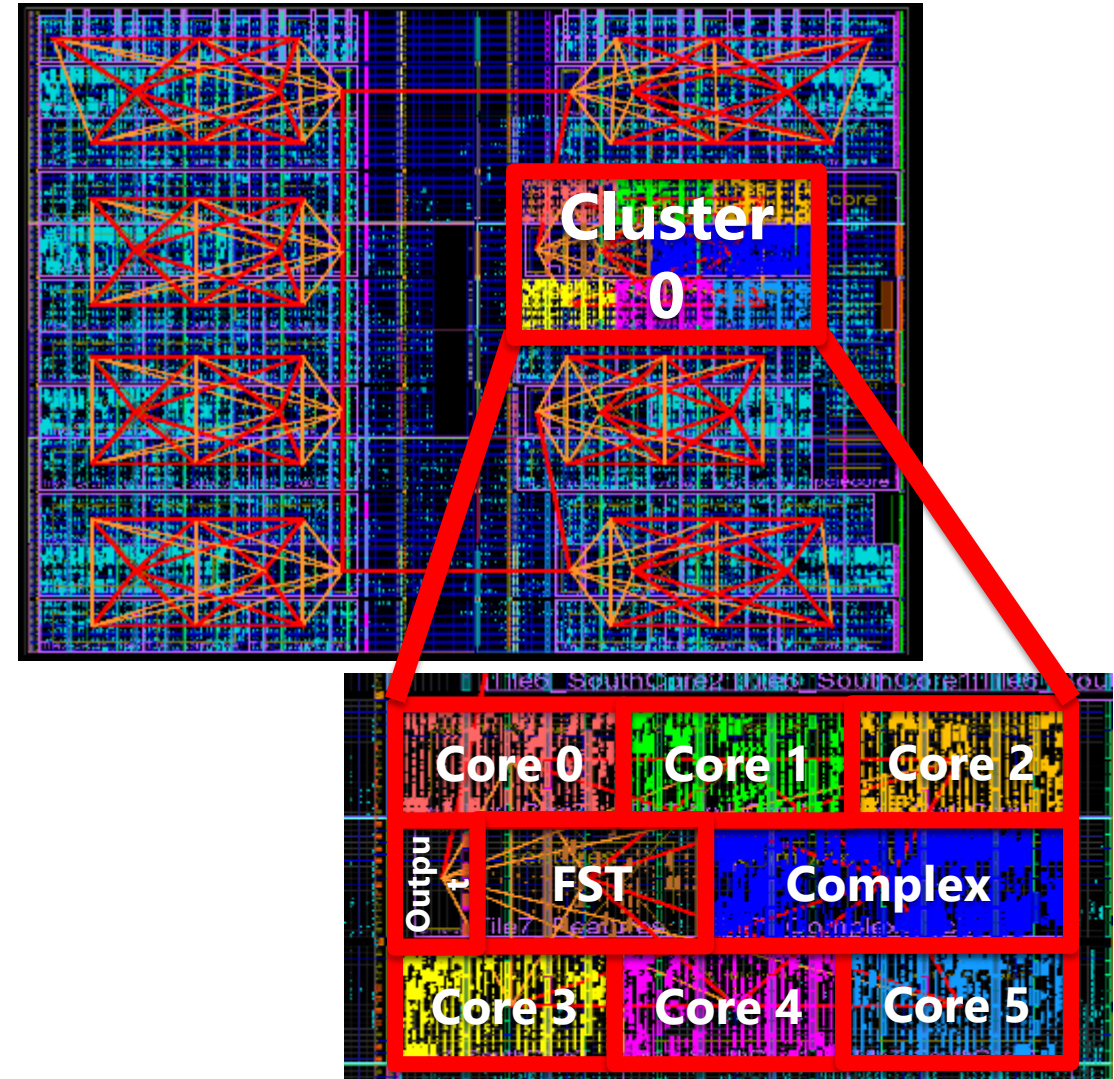
$$\text{FFE \#1} = 9$$

Feature Extraction Accelerator

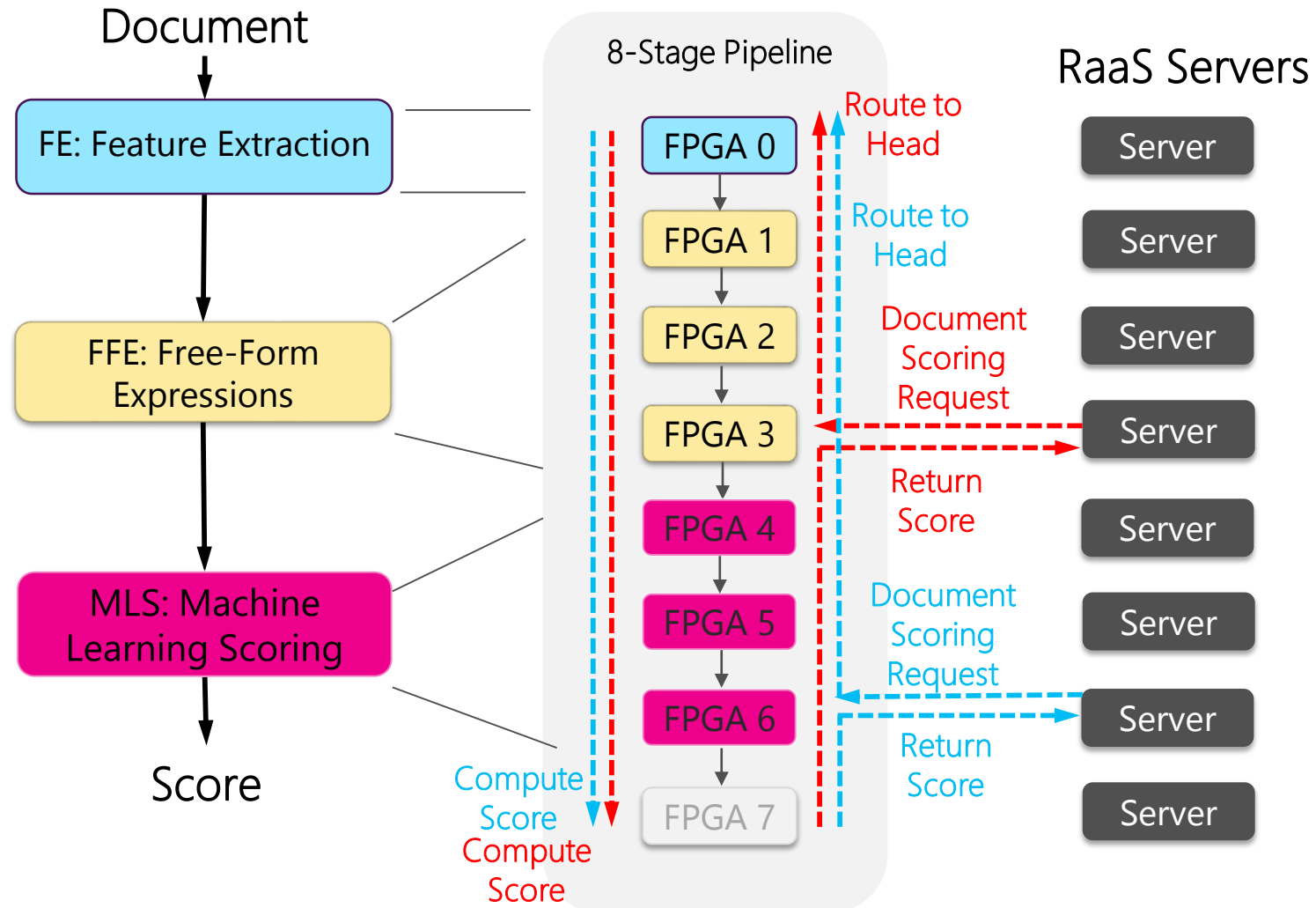


FFE Engines

- 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



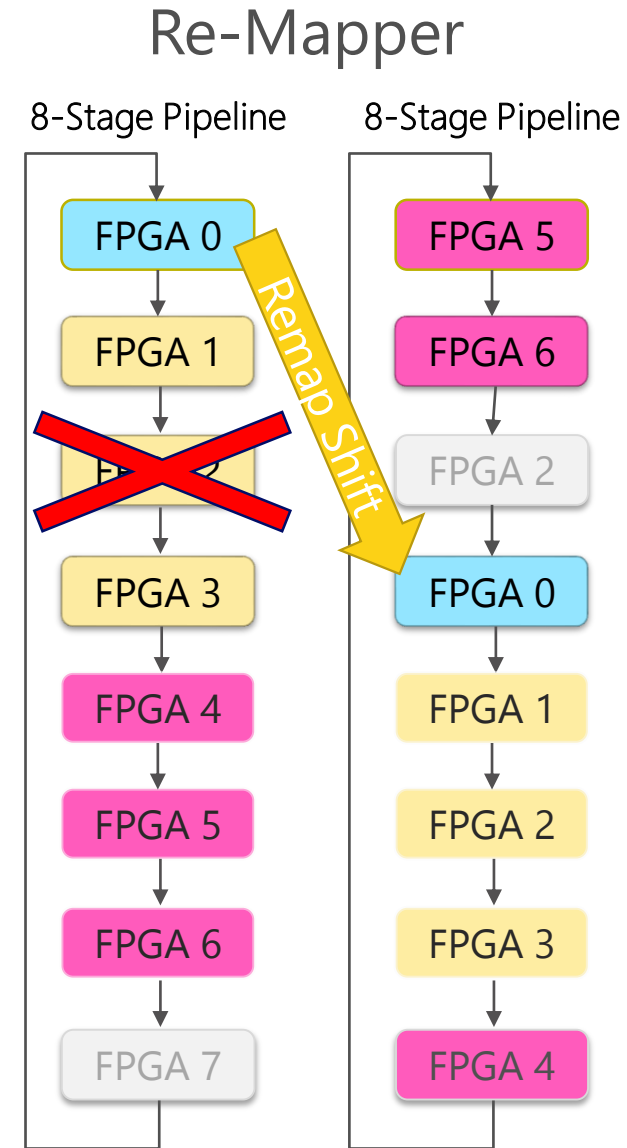
Scalable Deployment Challenges

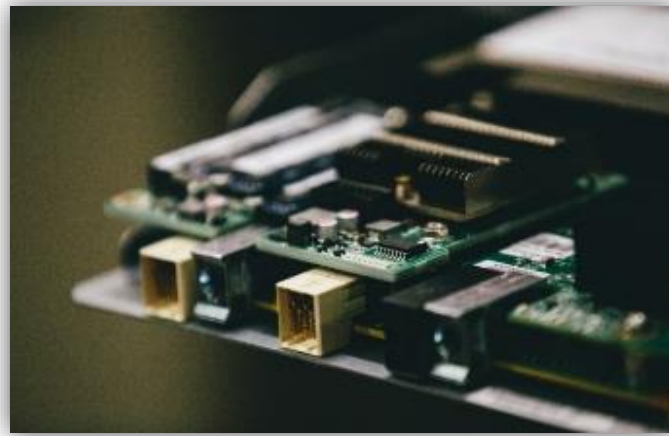
- Issues with Spanning Multiple FPGAs

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

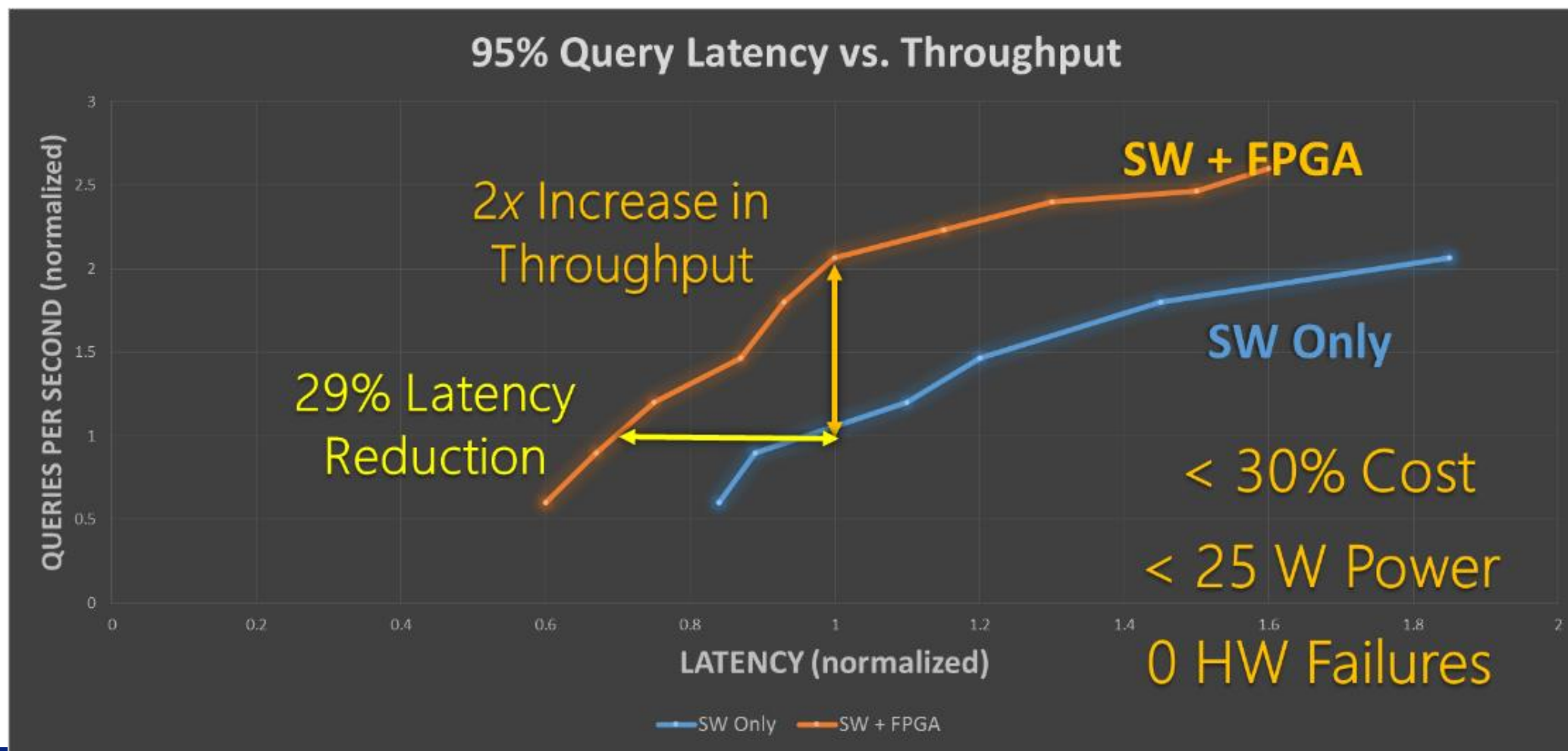




1,632 Server Pilot Deployed in a Production Datacenter

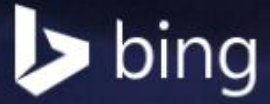
Accelerating Large-Scale Services – Bing Search

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: $\frac{1}{2}$ the number of ranking servers, lower latency, lower variance
- Proven scalability, proven resilience, and huge potential for future apps



But when will an FPGA handle my Bing Search?



"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



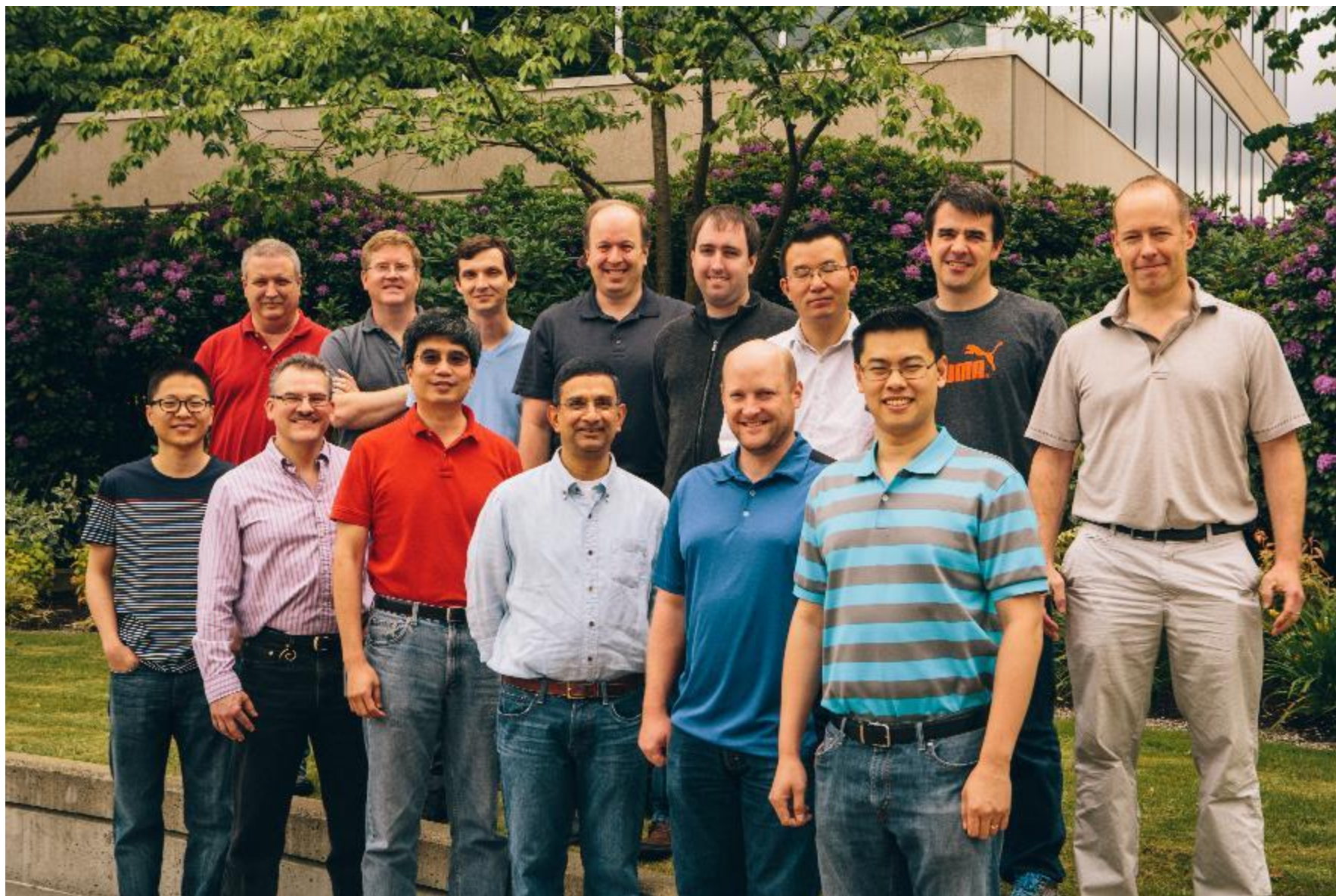


Top Row: Eric Peterson, Scott Hauck, Aaron Smith, Jan Gray, Adrian M. Caulfield, Phillip Yi Xiao, Michael Haselman, Doug Burger

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





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

