Microsoft

The Configurable Cloud -- *Accelerating Hyperscale Datacenter Services with FPGAs*

Andrew Putnam – Microsoft

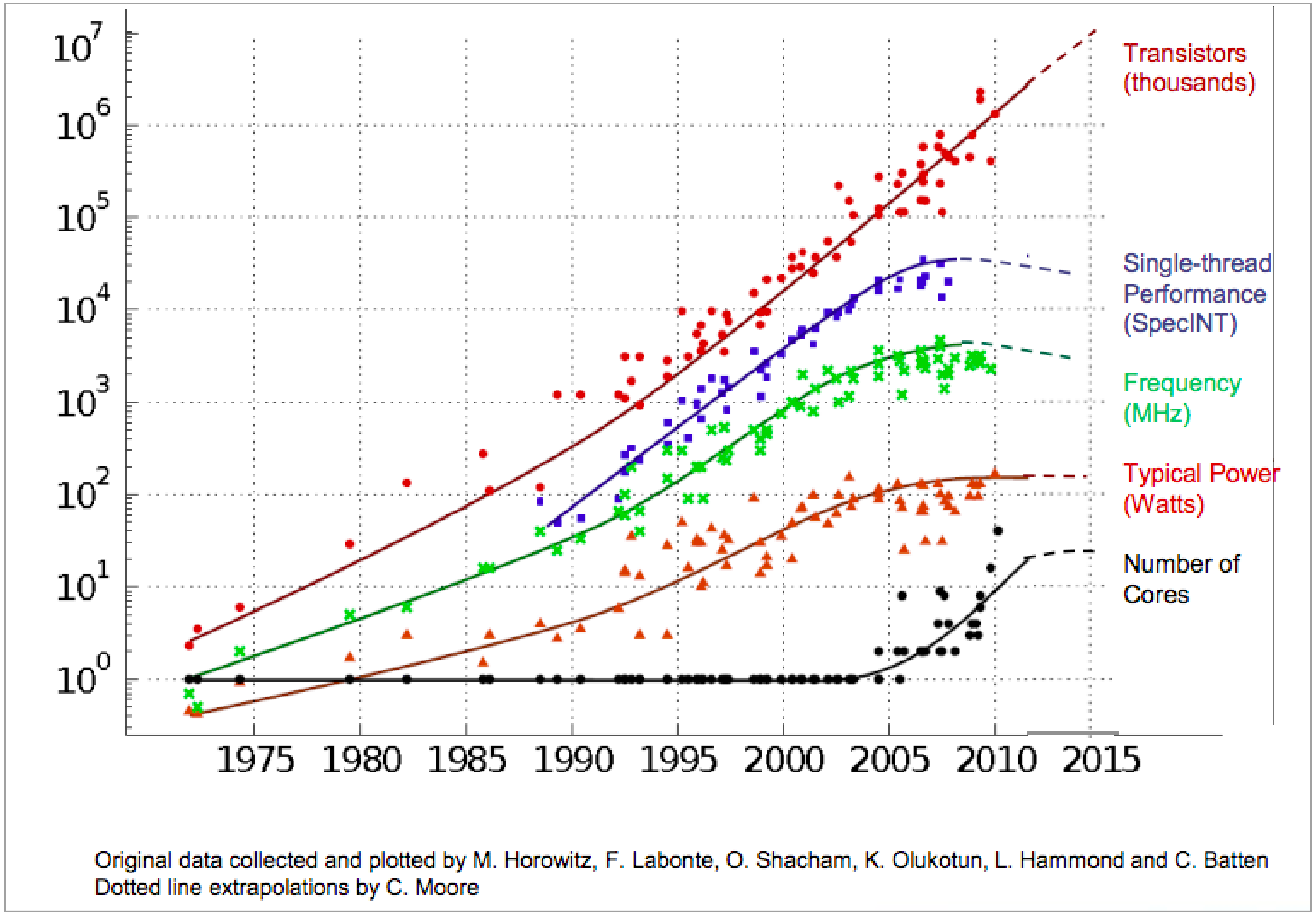
What is a Cloud / Big Data Application?

Microsoft Azure
bing
skype
Microsoft SQL Server
Visual Studio
COSMOS
LinkedIn
SharePoint
XBOX LIVE
Microsoft Dynamics
msn
Yammer
OneDrive
Microsoft Office 365
Outlook.com

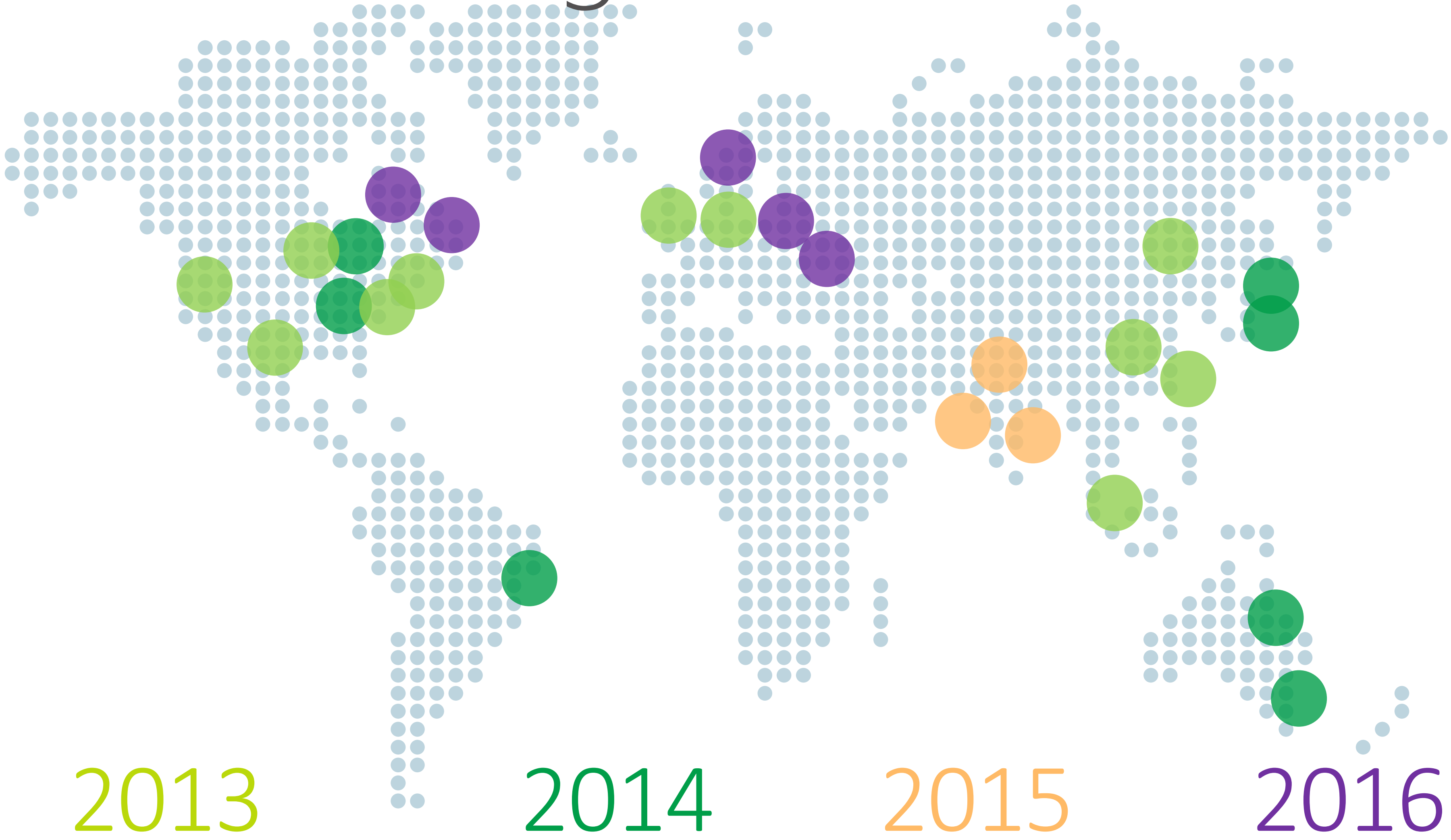
IaaS
PaaS
SaaS

Technology Scaling

- Moore's Law (transistors) is still alive
- Dennard Scaling (keeping energy under control) is dead
- 2x users requires 83% more servers
- Need increased *efficiency*



Datacenter Scaling



~100%+ Growth for the past 4 years

Modern HyperScale Datacenters

- Microsoft > 1,000,000 servers
- 100s of MegaWatts
- $100M+ power bill

TOP 10 Sites for November 2016

For more information about the sites and systems in the list, click on the links or view the complete list.

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	National Supercomputing Center in Wuxi China	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93,014.6	125,435.9	15,371
2	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
3	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
4	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890

Datacenter:
~100,000 dual-socket servers

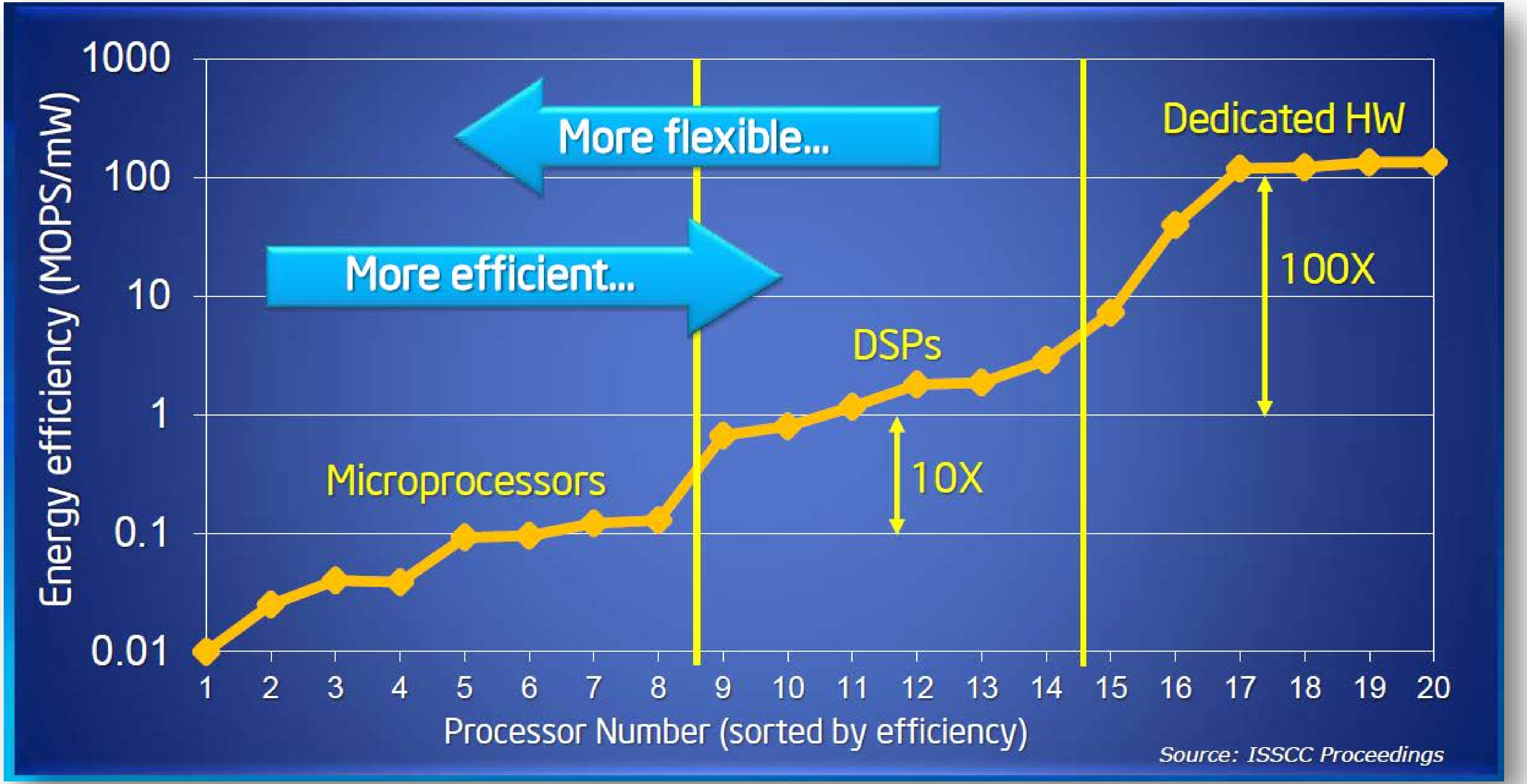
40,960 single-socket servers

16,000 dual-socket servers
3 Xeon Phi / server

18,688 single-socket servers
1 Tesla GPU / server

98,304 single-socket servers

Efficiency via Specialization



Source: Bob Broderson, Berkeley Wireless group

Silicon Technologies for Computing

More Flexible

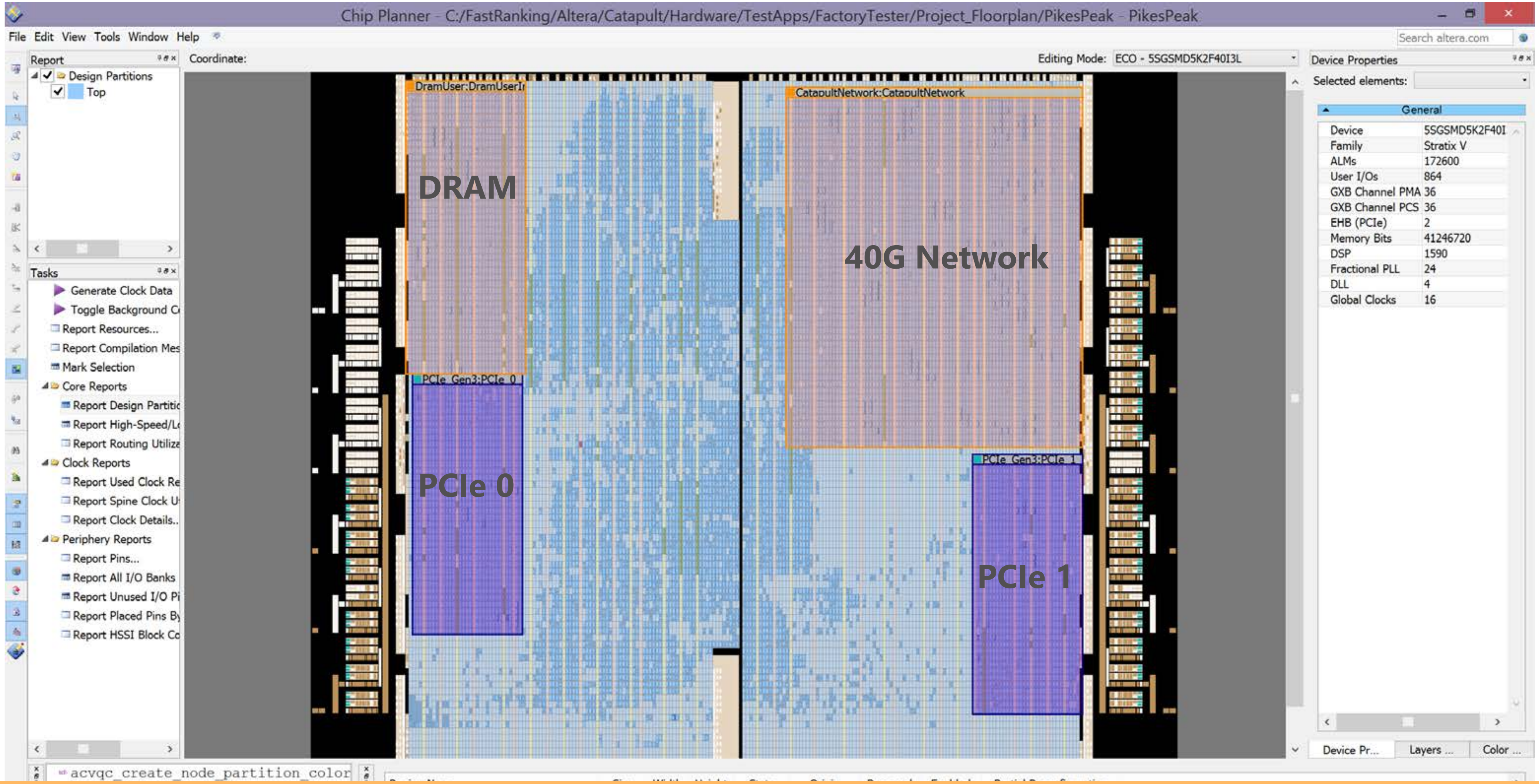
		Perf/W			
	CPUs	1X	Today's standard, most programmable, good for services changing rapidly	Conventional programming	C/C++
	Manycore CPUs	3X	Many simple cores (10s to 100s per chip), useful if software can be fine-grain parallel, difficult to maintain.		
	GPUs	5-30X	Good for data parallelism by merged threads (SIMD), High memory bandwidth, power hungry	Alternative programming	CUDA
	FPGAs	5-30X	Most radical fully programmable option. Good for streaming/irregular parallelism. Power efficient but currently need to program in H/W languages.		Verilog
	Structured ASICS	20-100X	Lower-NRE ASICs with lower performance/efficiency. Includes domain-specific (programmable) accelerators.	Can't change functionality	Verilog
	Custom ASICs	> 100X	Highest efficiency. Highest NRE costs. Requires high volume. Good for functions in very widespread use that are stable for many years.		

More Efficient

What are FPGAs?

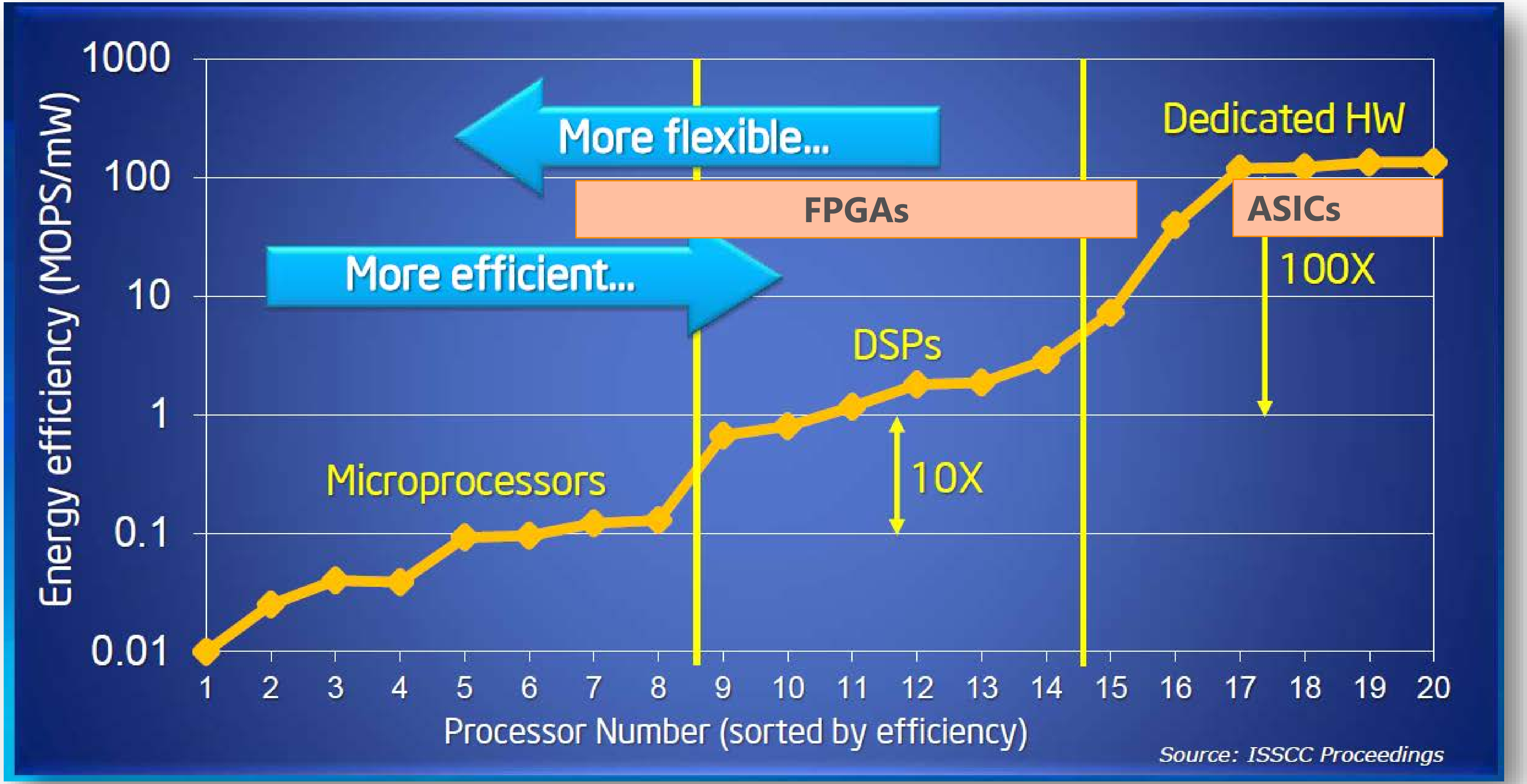
- <u>F</u>ield <u>P</u>rogrammable <u>G</u>ate <u>A</u>rray
- **FPGAs** are a sea of generic logic and interconnect
  - "Silicon Legos" – build them into exactly the right circuit for each task
- Special-purpose hardware (FPGAs) is faster and more efficient than general-purpose hardware (CPUs)
- **Change the hardware anytime!**
  - 100 ms to 1 second reconfiguration time

FPGA Physical Layout



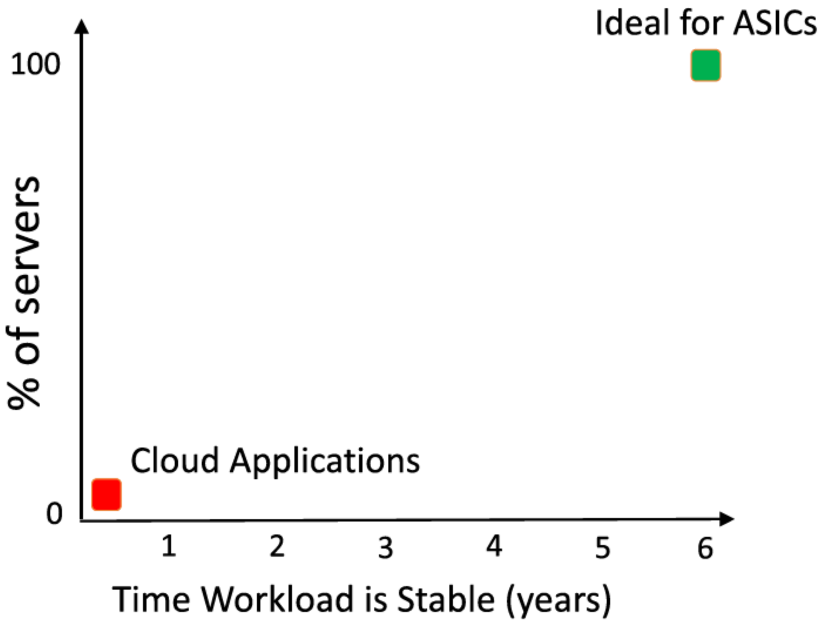
Customize both the processing logic *and* the I/O

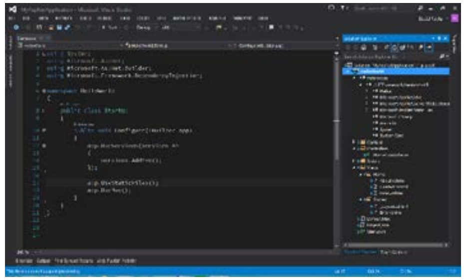
Efficiency via Specialization



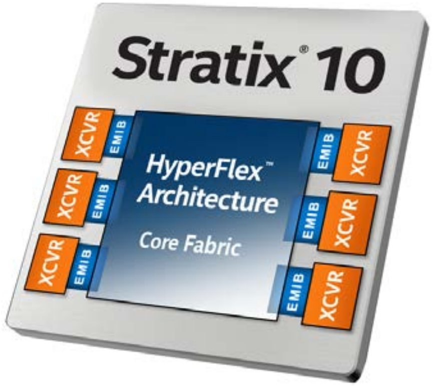
Source: Bob Broderson, Berkeley Wireless group

Why not use ASICs?

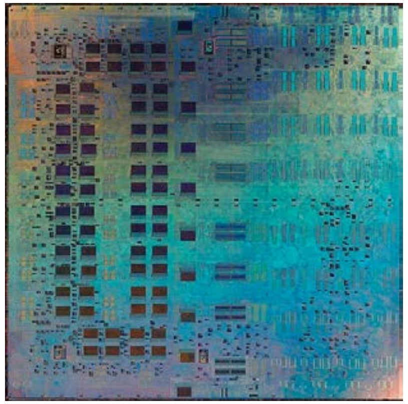




Software



FPGA



ASIC

Why not use GPUs?

- SIMD is best for batch jobs
- Customer-facing (interactive) workloads are small batches, need low latency
- Limited floating point for most workloads
    - Scientific computing and ML are exceptions
- *Optimize for the whole fleet, not for one application*

Also…. power

TOP 10 Sites for November 2016

For more information about the sites and systems in the list, click on the links or view the complete list.

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	National Supercomputing Center in Wuxi China	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93,014.6	125,435.9	15,371
2	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
3	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
4	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890

Datacenter:

~100,000 dual-socket servers

ALL WITH FPGAs!

40,960 single-socket servers

16,000 dual-socket servers
3 Xeon Phi / server

18,688 single-socket servers
1 Tesla GPU / server

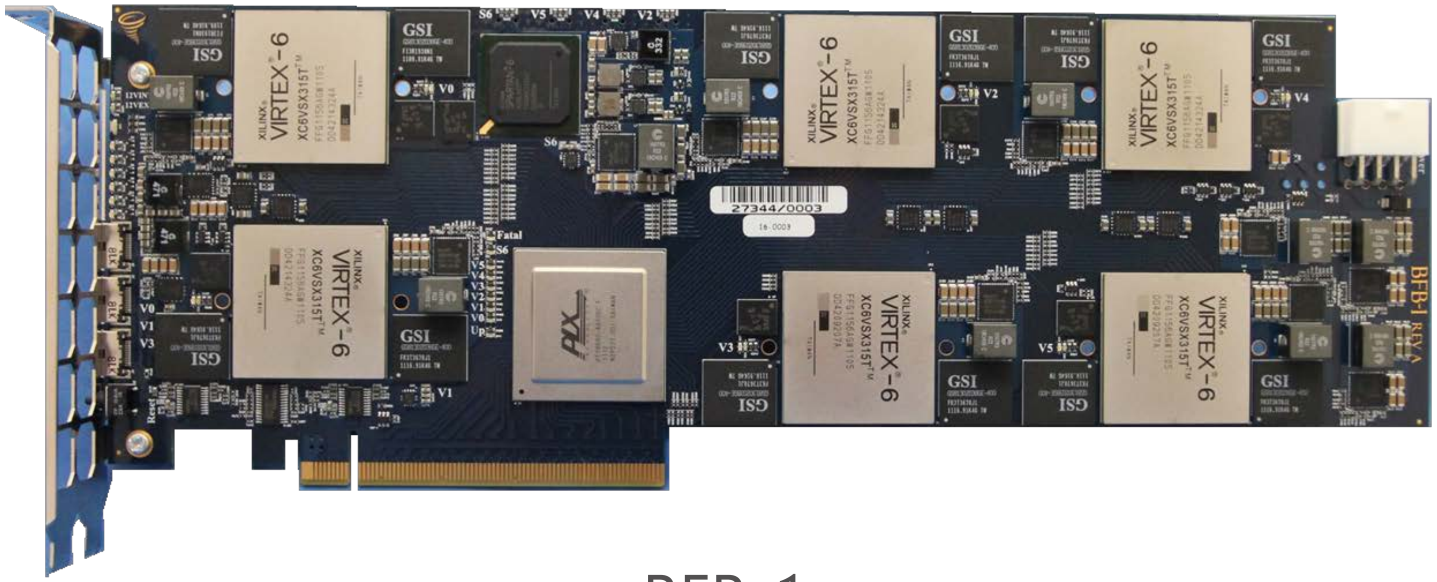
98,304 single-socket servers

Bing: How it all Began

- Launched June 1, 2009
  - MSN Search
  - Windows Live Search
  - Live Search
- October 2010: Approached Microsoft Research for help optimizing performance
- December 2010: Designed an FPGA accelerator for one critical stage
  - First designed at Starbucks – in PowerPoint

BFB-1

- Created a custom board with 6 big FPGAs
  - First designed at Starbucks (& Duke's Chowder House) – in PowerPoint

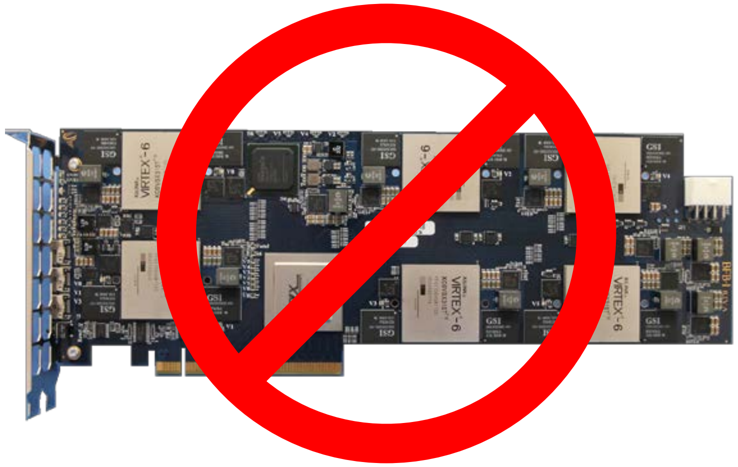


BFB-1

And Datacenter Operations says…

Centralized Model Complicates Deployment

- Single point of failure
- Server is different from the surrounding servers
- Complicates rack design, thermals, maintainability
- CPU network needed for FPGA communication
  - Definition of the Network In-cast problem
  - Precludes many latency-sensitive workloads
- Limited elasticity
  - What if you need more than six FPGAs?

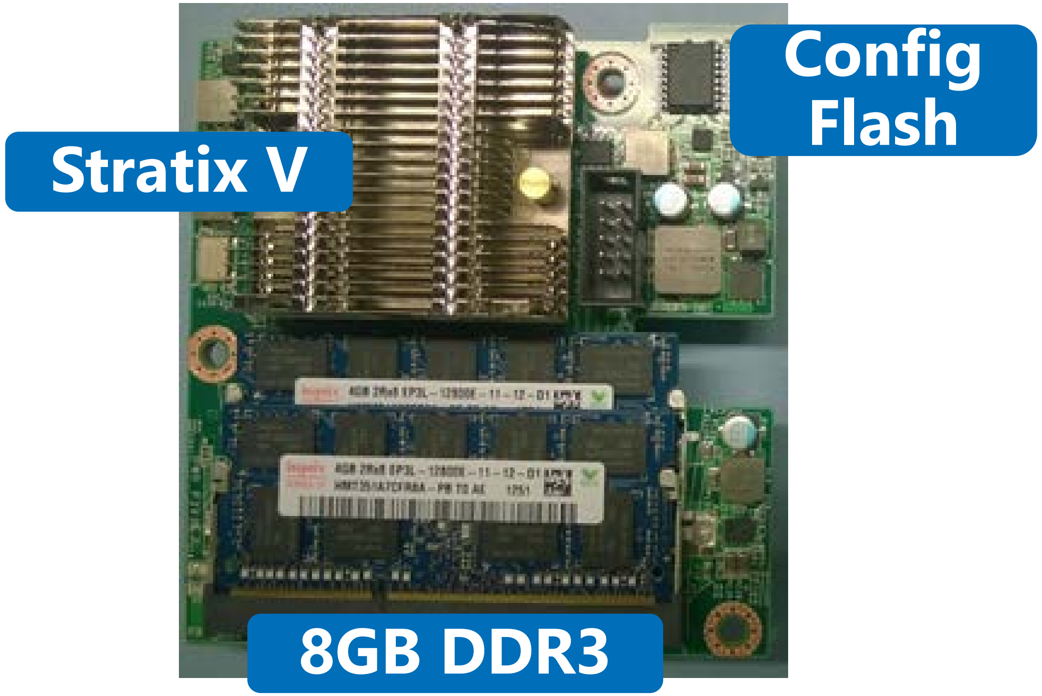


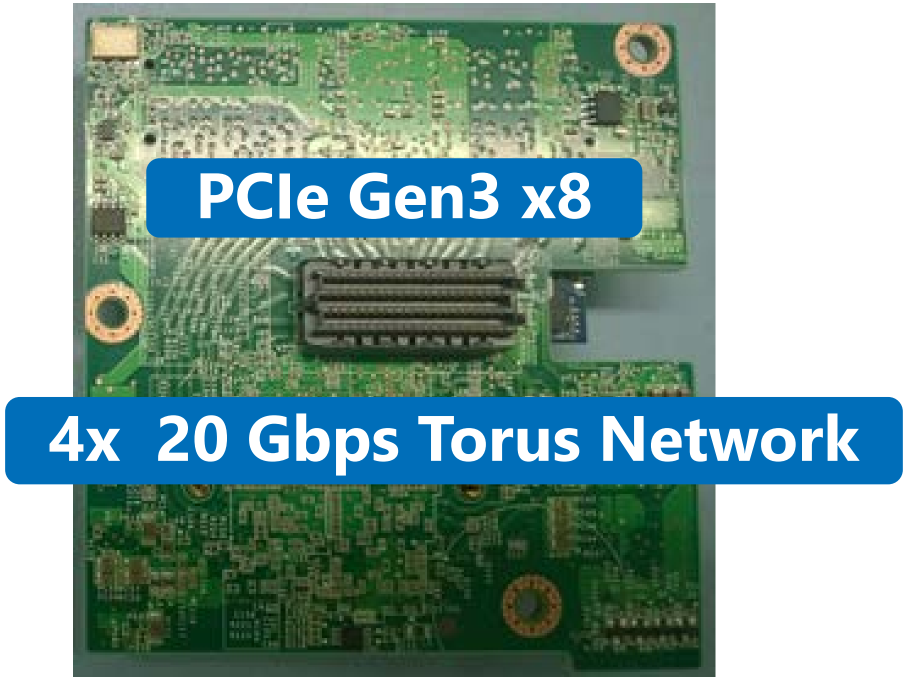
Fitting FPGAs in the Datacenter

- All servers should be the same
- How about just 1 FPGA in each server?
- Area must be small. Temperatures high. Power low.

Catapult v1 FPGA Accelerator Card

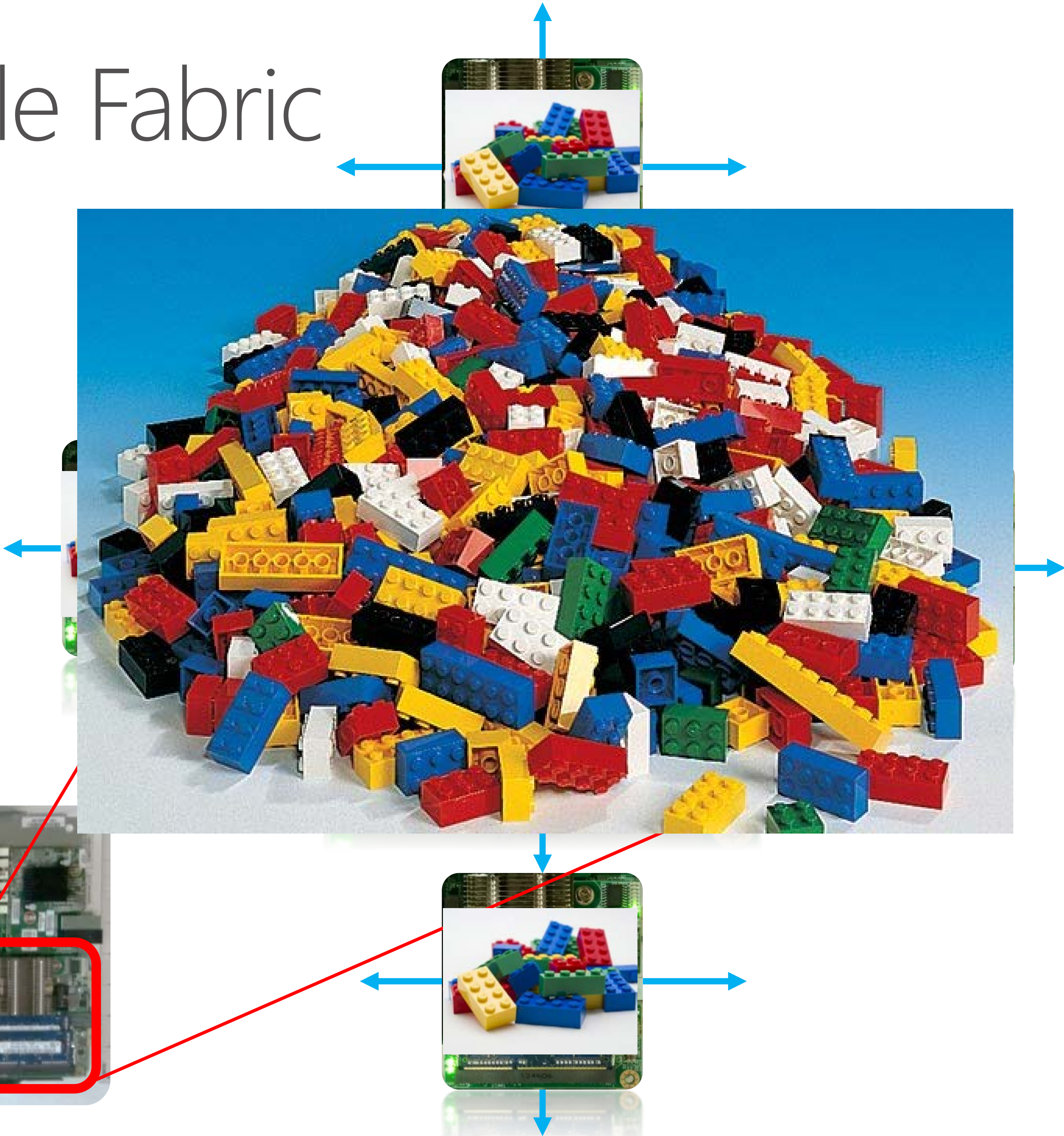
- Altera Stratix V GS D5
  - 172k ALMs, 2,014 M20Ks, 1,590 DSPs
- 8GB DDR3-1333
- 32 MB Configuration Flash
- PCIe Gen 3 x8
- 8 lanes to Mini-SAS SFF-8088 connectors
- Powered by PCIe slot



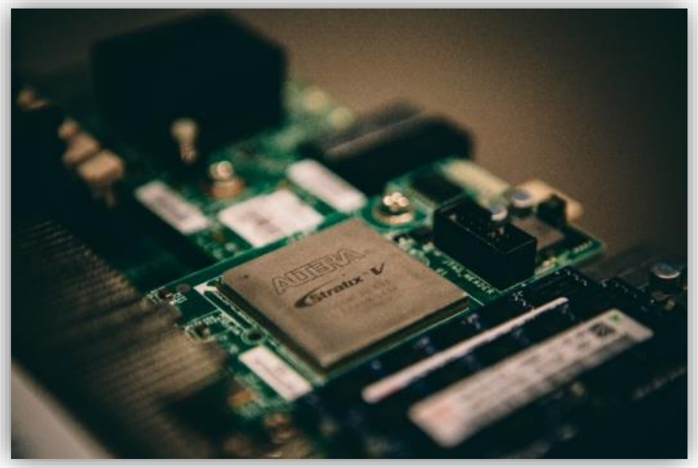


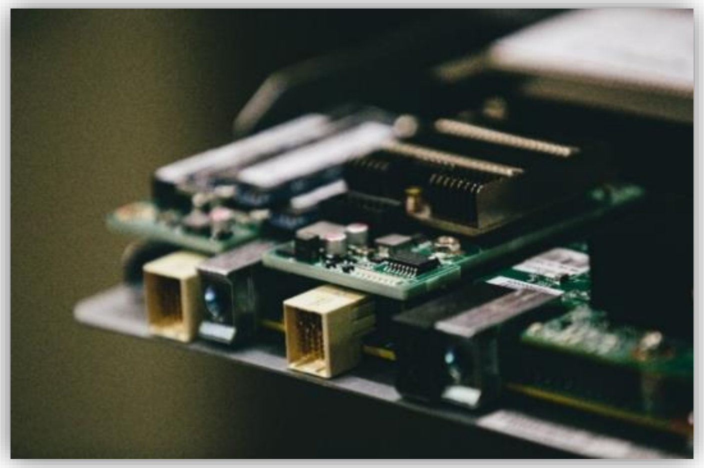
Scalable Reconfigurable Fabric

- 1 FPGA board per Server
- 48 Servers per ½ Rack
- Network among FPGAs:
  - 6x8 Torus at 20 Gb/link



Data Center Server (1U, ½ width)

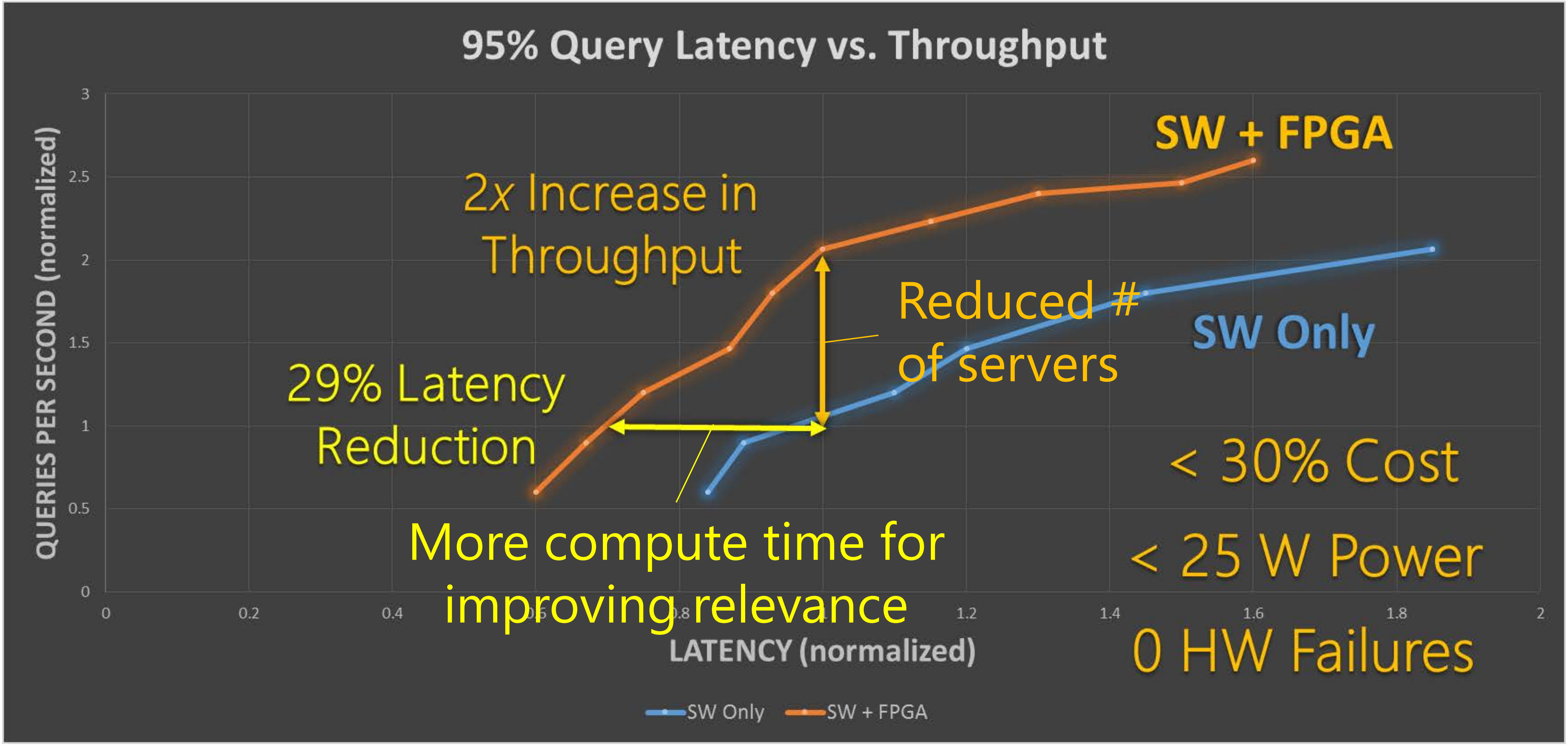




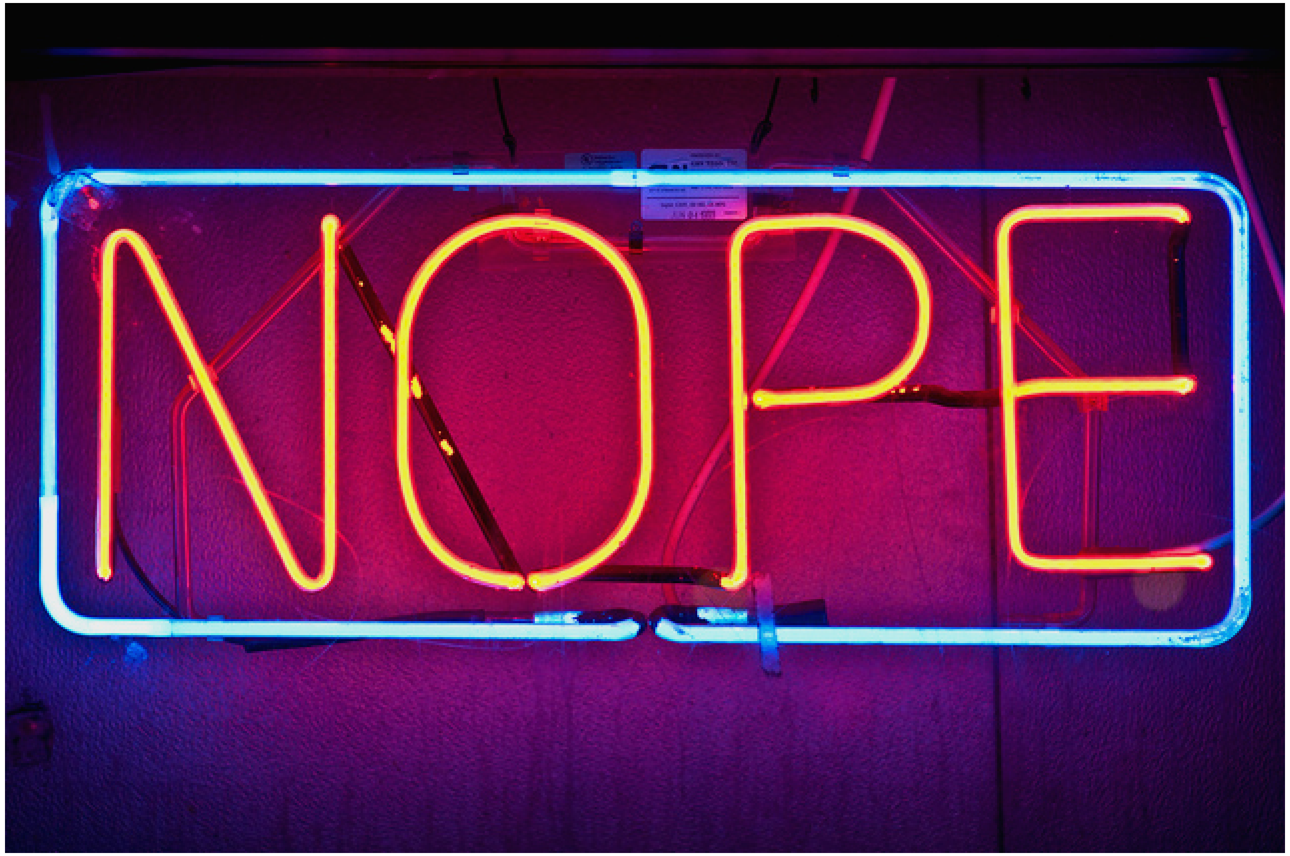

1,632 Server Pilot Deployed in a Production Datacenter

Accelerating Large-Scale Services – Bing Search

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

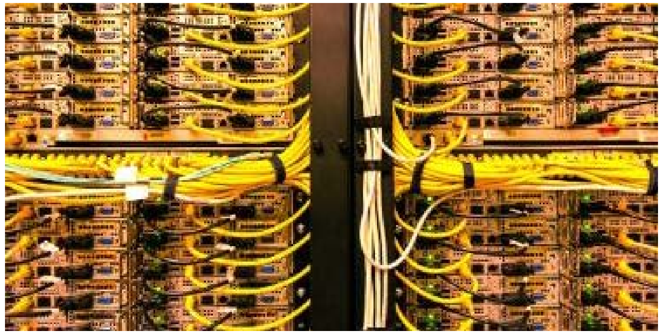


Was Catapult v1 deployed into production?



5.8+ billion
worldwide queries each month

bing



250+ million
active users

OneDrive

400+ million
active accounts

Outlook.com

2.4+ million
emails per day

Microsoft® Exchange Hosted Services

8.6+ trillion
objects in Microsoft Azure storage

Microsoft Azure

48+ million
users in 41 markets

XBOX LIVE

50+ million
active users

Microsoft® Office Web Apps

1 in 4
enterprise customers

Office 365

50+ billion
minutes of connections handled each month

skype™

200+ Cloud Services: **Diversity**

1+ billion customers · 20+ million businesses · 90+ markets worldwide

Workload Diversity

- Bing *was* the big dog, but Azure grew much faster
- Compute offload for Bing isn't enough to justify hyperscale deployment

- Could go Bing-specific
  - Misconfiguration is the leading cause of problems in the datacenter
  - Increased hardware diversity means increased chances of misconfigurations
- Could try to do offload for all the other services
  - But is there enough time to learn each new application?

Compute vs. Infrastructure

- *Compute* acceleration is application-specific
  - Bing Ranking
  - Machine Learning / DNNs

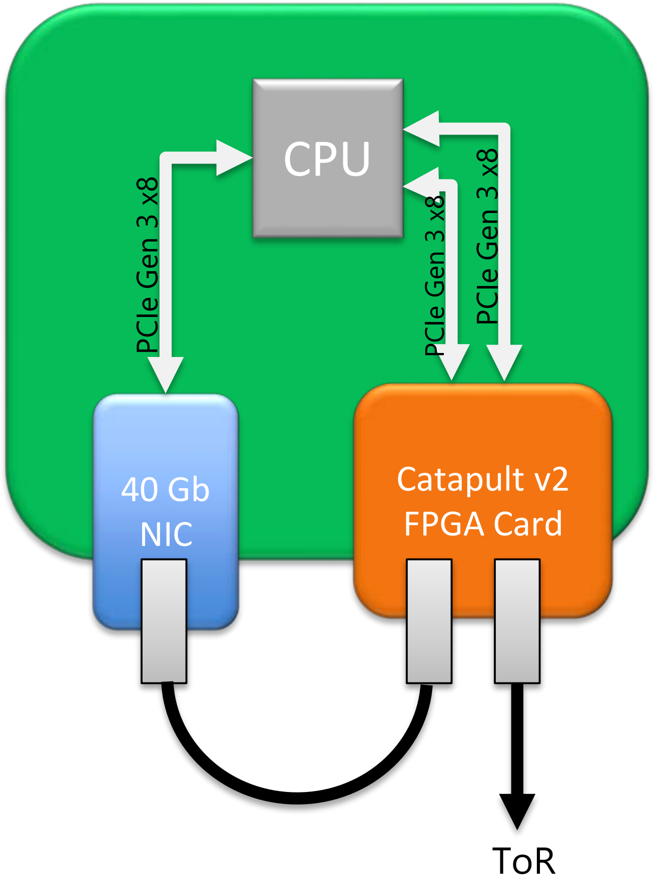
bing ≈ 1-2 years

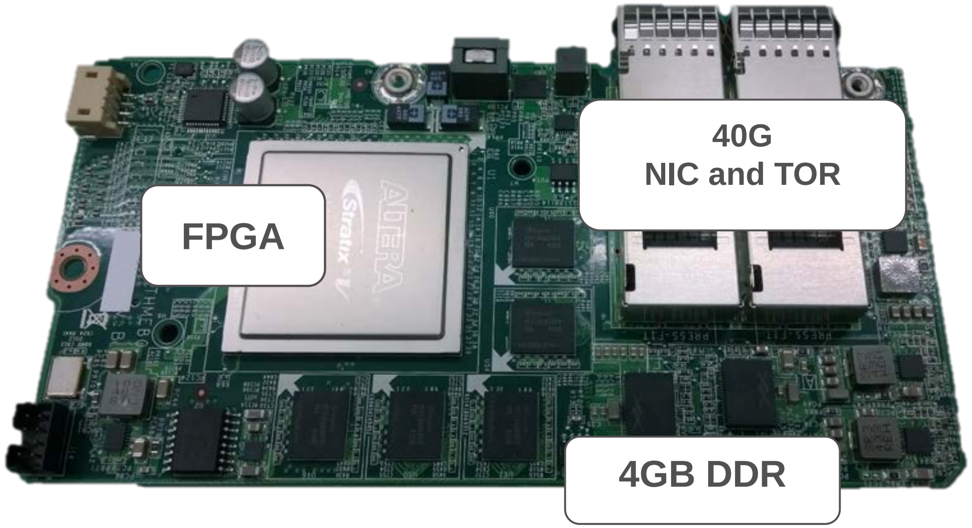
XBOX LIVE · Microsoft Azure · skype · SQL Server · Visual Studio · Microsoft Dynamics · msn · SharePoint · Outlook.com · OneDrive · LinkedIn · Office 365 · Yammer ≈ 12 years?

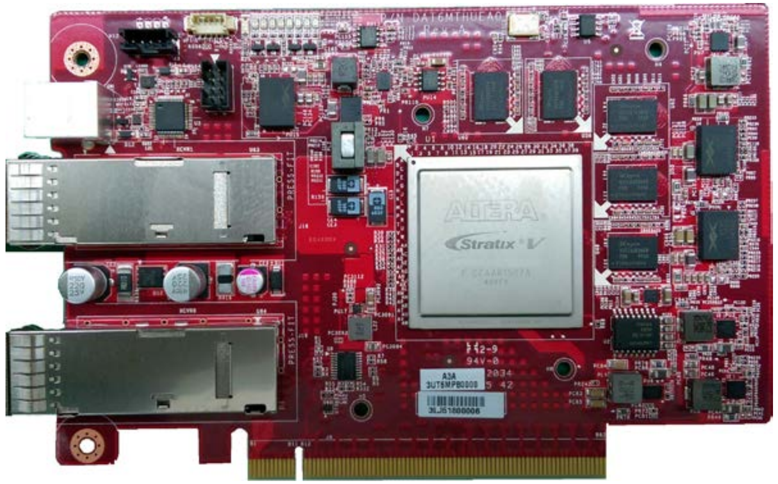
- *Infrastructure* acceleration benefits common software and services
  - Network offload and processing
  - Encryption / Compression
  - Security

- BOTH are critical when dealing with diverse cloud workloads

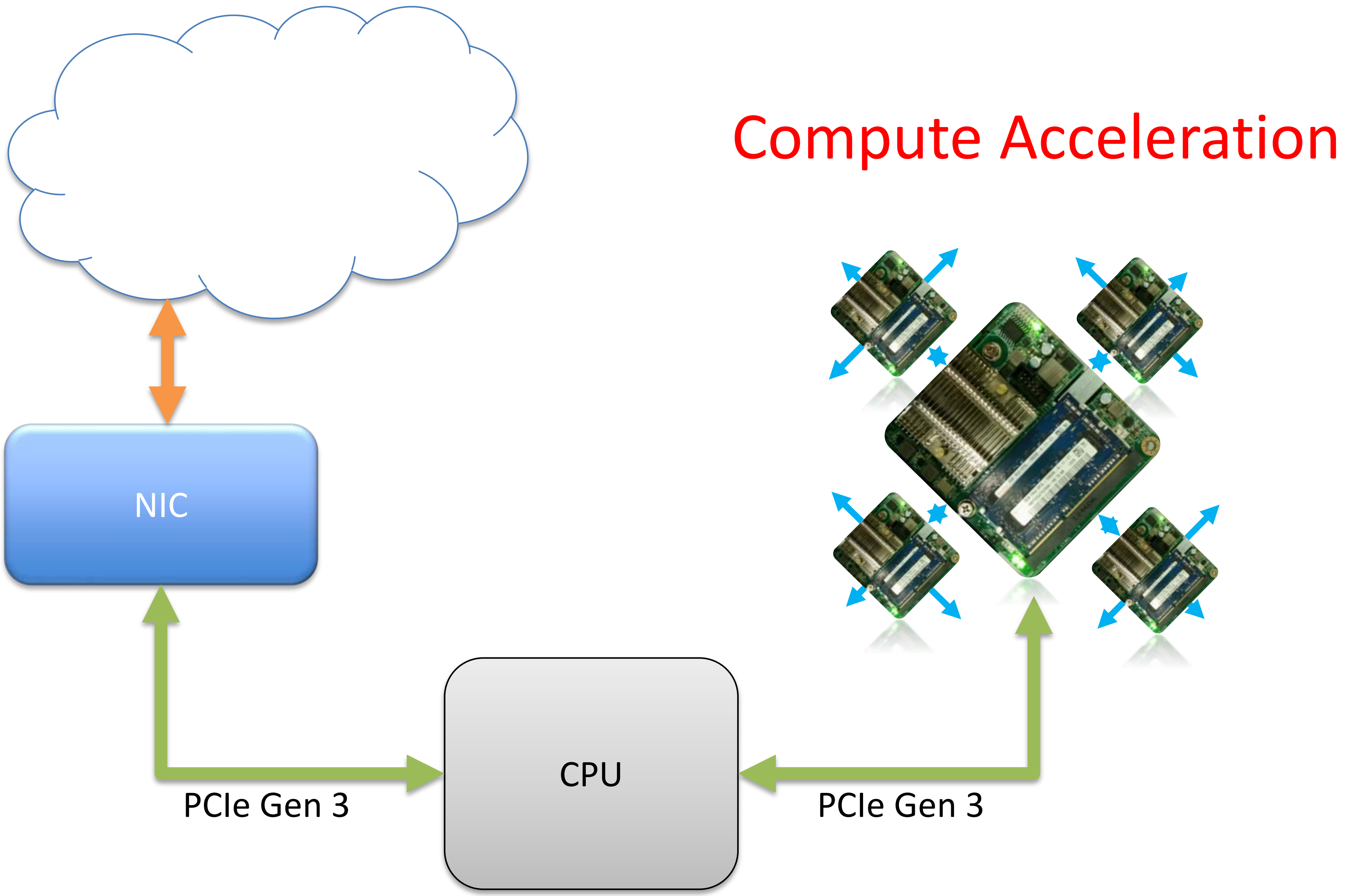
Catapult v2 – Bump in the Wire



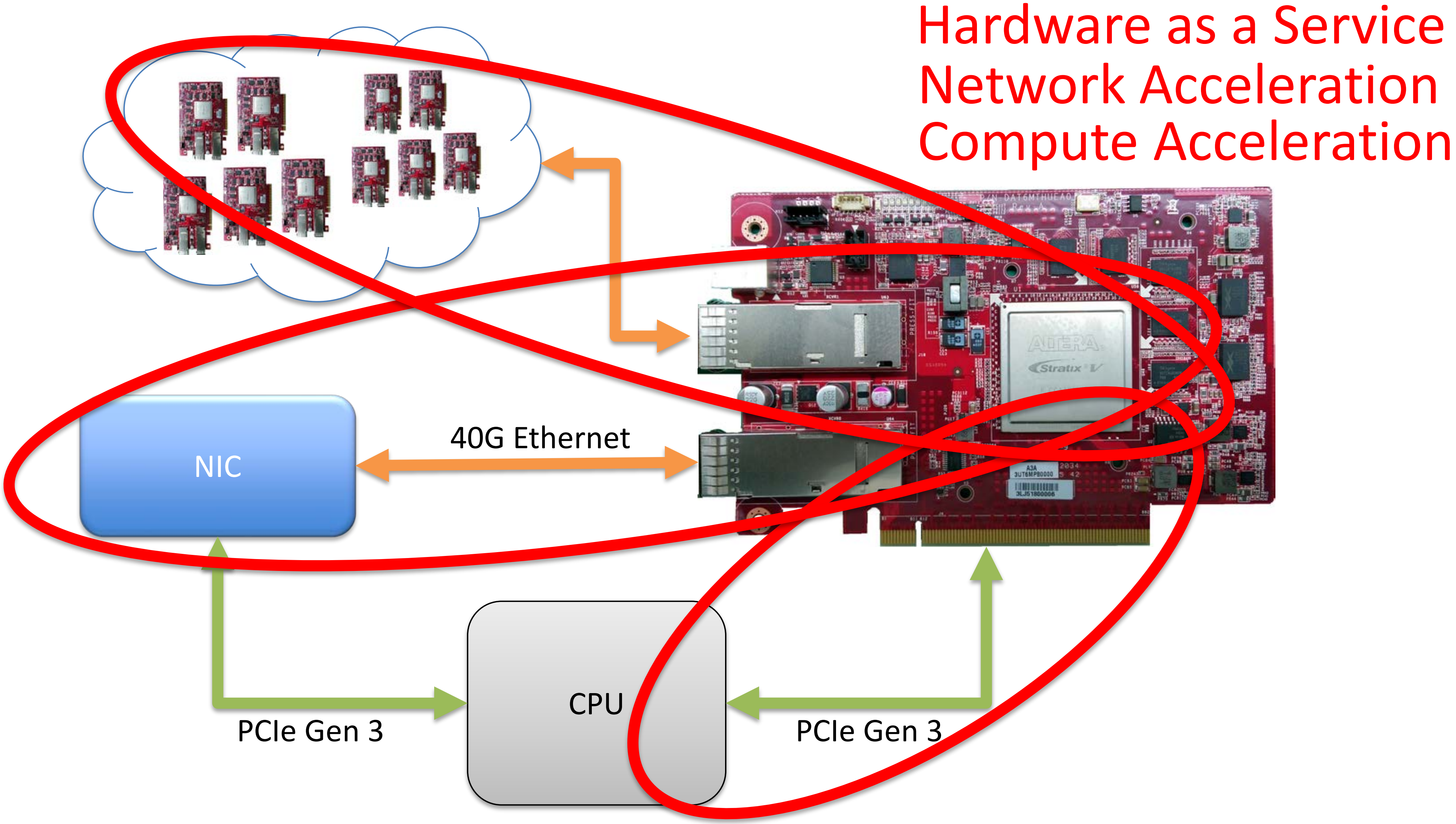




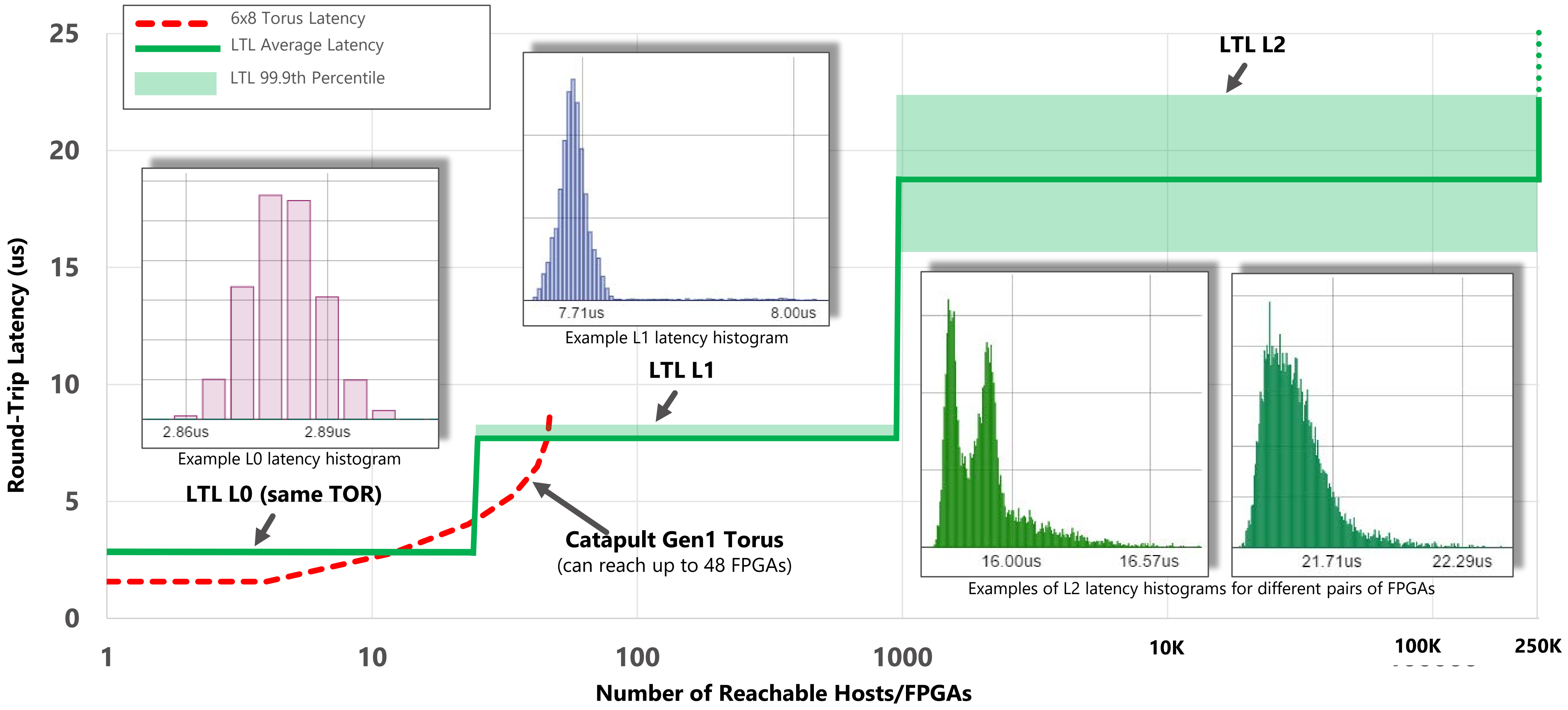
Catapult v1 Architecture



Bump-in-the-wire Architectue



Network Latencies

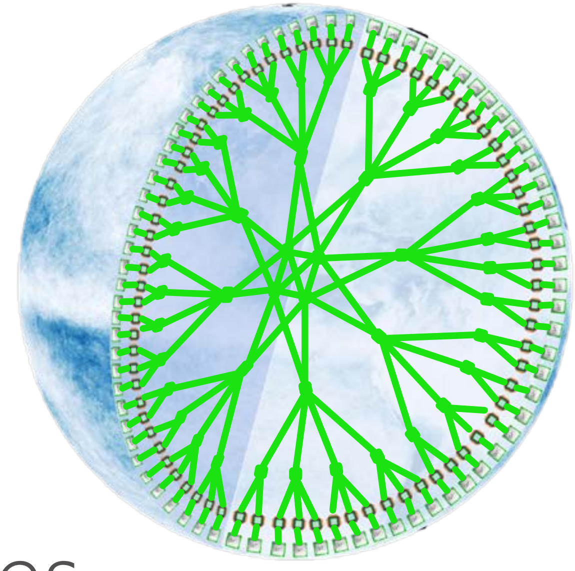


Inter-FPGA Communication

- LTL -- Lightweight Transport Layer
- A Low-latency, Reliable, Connection-based communication channel for FPGA-to-FPGA messaging over standard Ethernet network
- Send-side buffering and retransmit until recv. ACK
- Built in send-side queueing mechanism to handle serialization of messages during contention
- FPGAs can communicate without any CPU intervention

Benefits of Bump-in-the-Wire

- Compute & Infrastructure Acceleration with one board
- A global hyperscale FPGA fabric – 100k+ FPGAs
- Improved Robustness & Fault Tolerance
- Fewer hops between FPGAs
- Independent of physical location
- Customized network hardware & protocols
- Allows sharing of underutilized FPGA resources



All while retaining hardware homogeneity!

Was *this* Catapult deployed into production?

- **FPGAs Included in every new server for all major services**
- Deployed across 15 countries and 5 continents
- Already in large scale production in both Bing and Azure

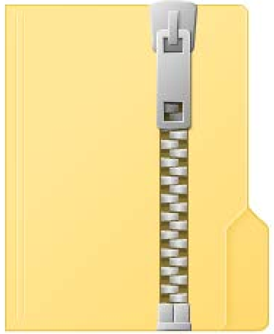
What workloads run well?

Microsoft Azure

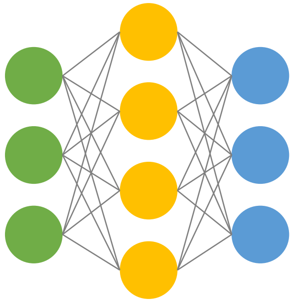
CNN

Encryption

bing



Compression

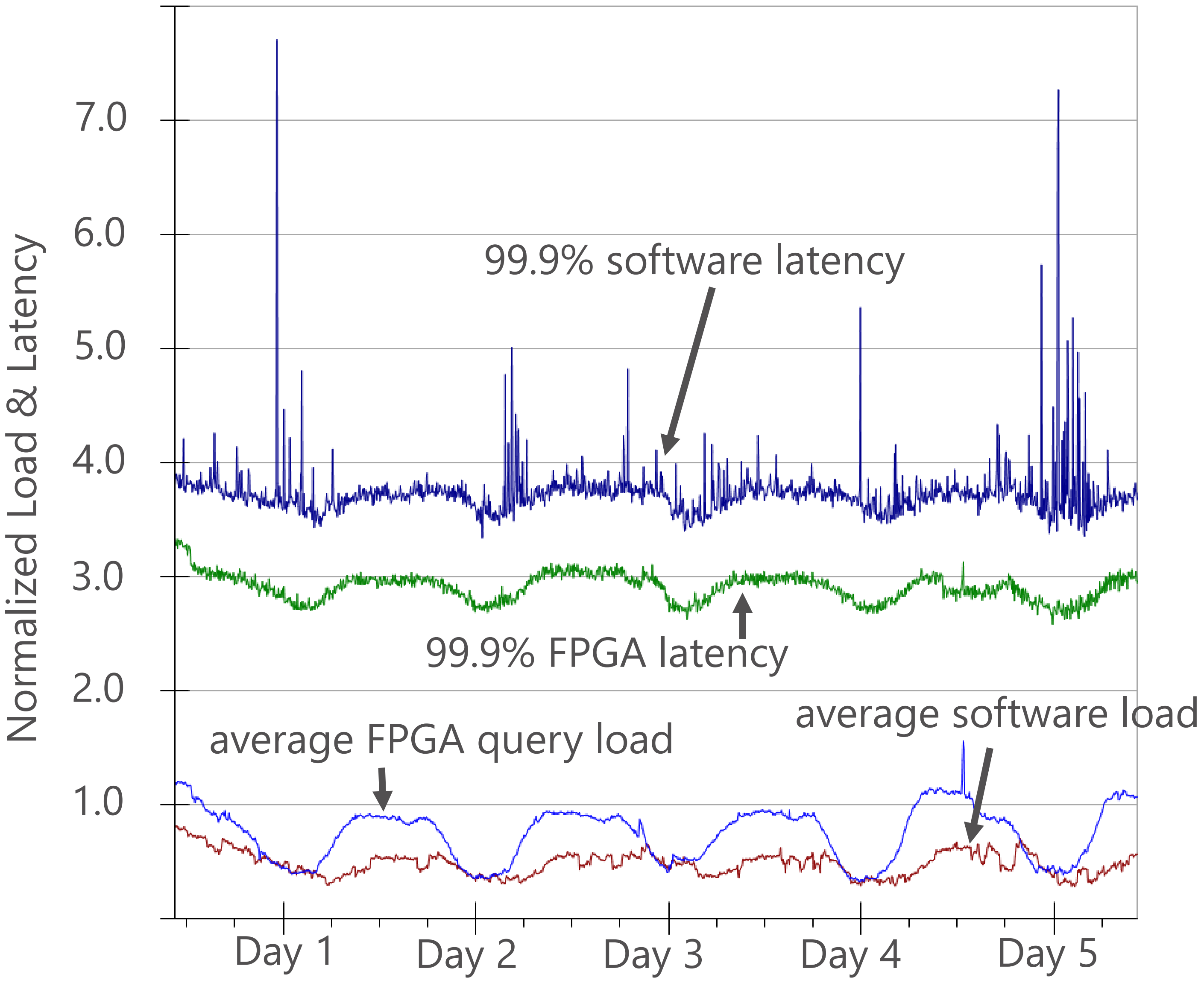


DNNs

Compute Acceleration -- Bing Ranking

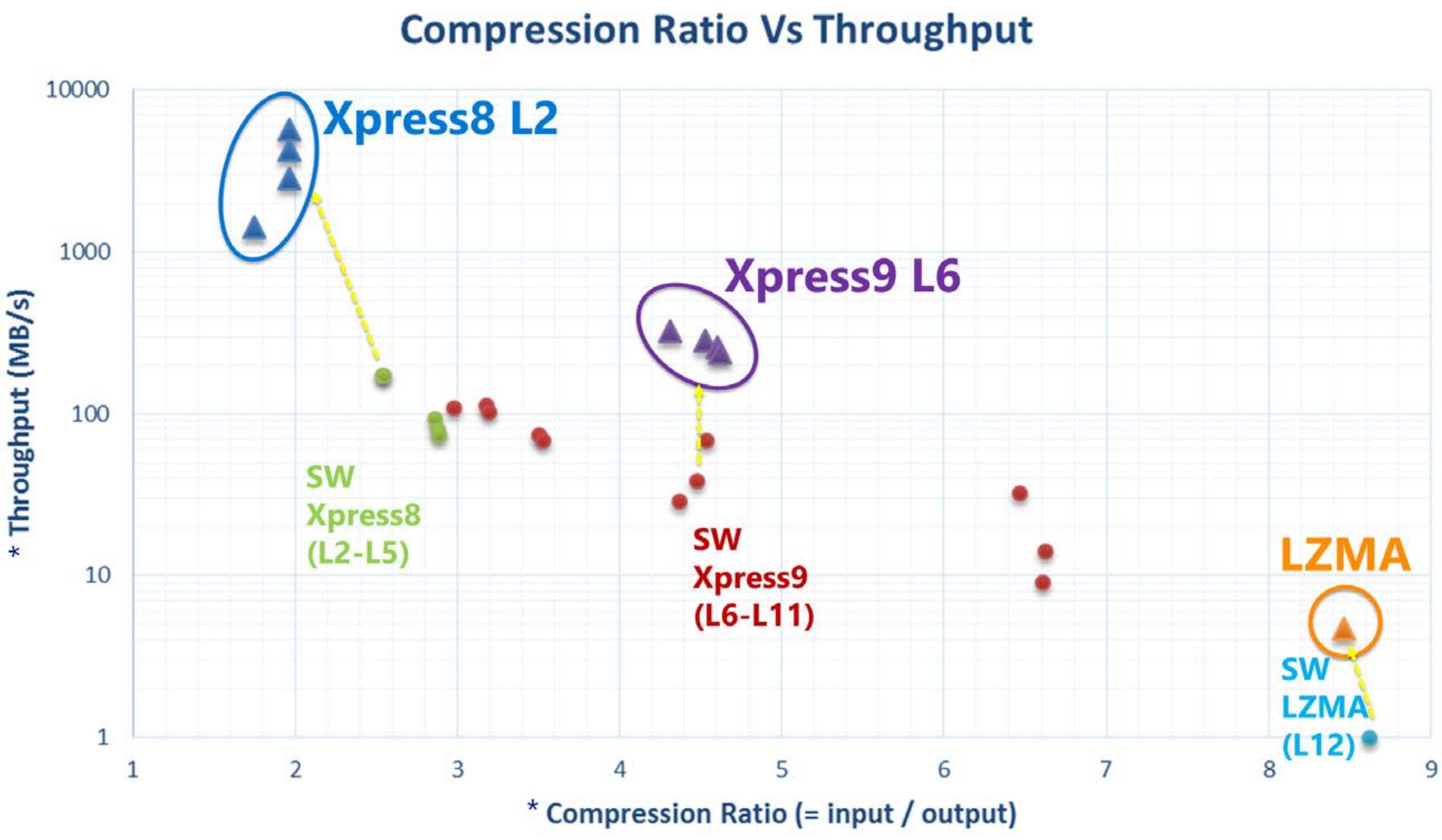
bing

- 2x Faster at 2x higher load
- Much lower variance



Data Compression

Microsoft Office 365 | Outlook.com



*- Measured on Canterbury dataset

Xpress8 L2 (5.6GB/s)
- 30x throughput
- 20% compression loss
- In-line compression

Xpress9 L6 (300MB/s)
- 4x throughput
- No compression loss
- Short/mid-term data

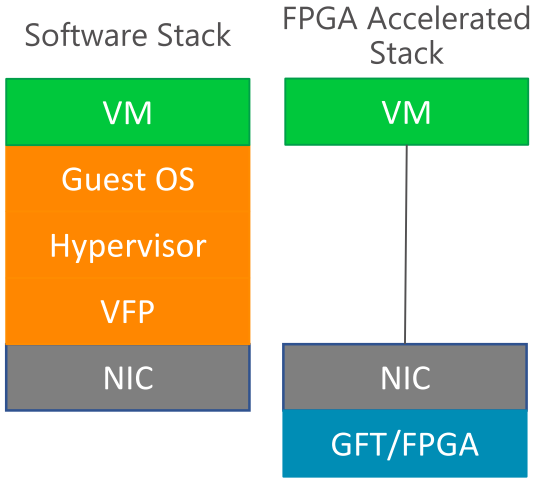
LZMA (5MB/s)
- 5x throughput
- 5% compression loss
- Long-term storage

Infrastructure Acceleration

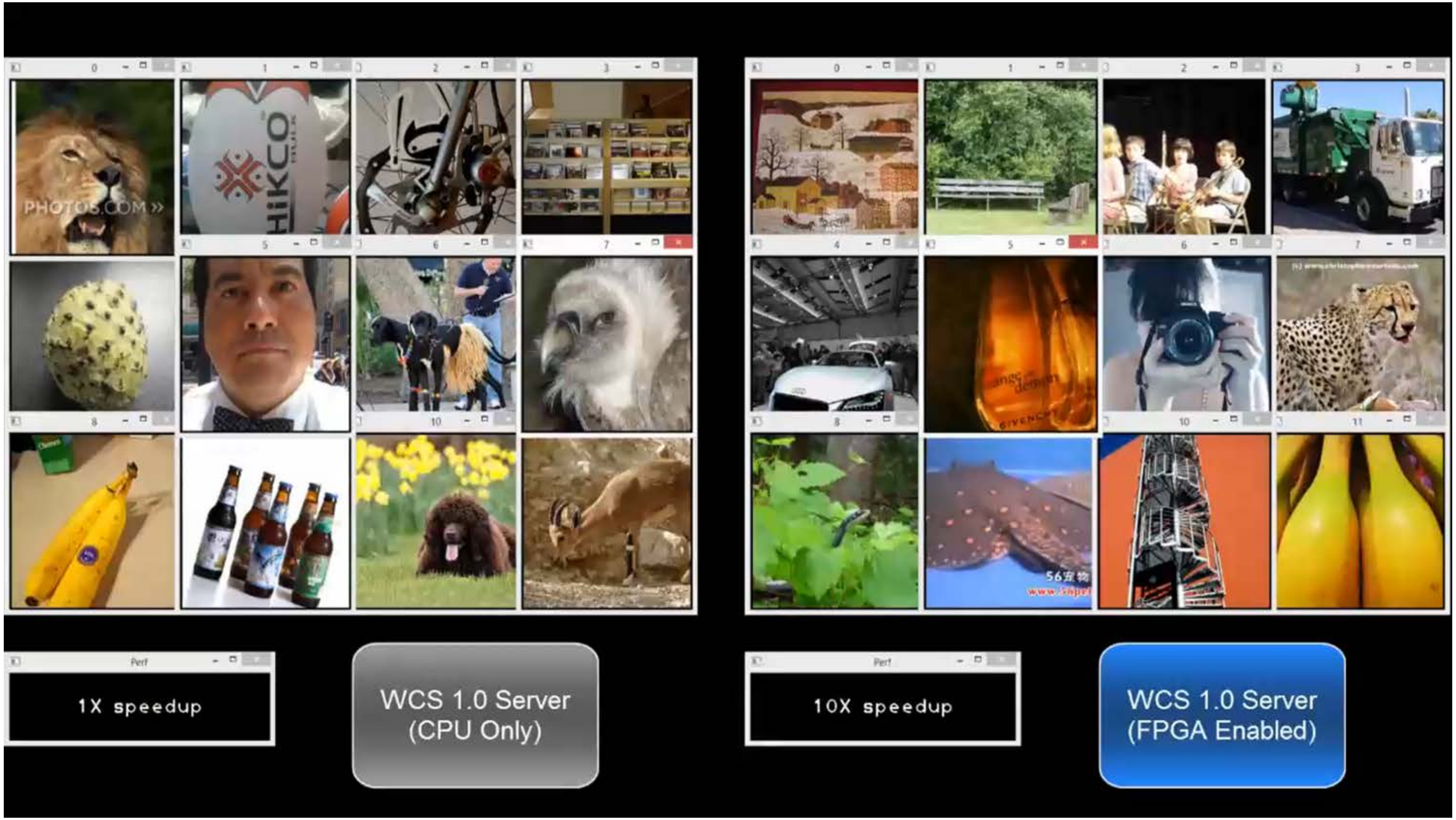
Microsoft Azure

SmartNIC: SDN and Crypto offload

- Generic Flow Table (GFT) rule based packet rewriting
- Enhanced network security
- 10x latency reduction vs software, CPU load now <1 core
- 25Gb/s throughput at 25µs latency – **the fastest cloud network**



Deep Learning -- Image Classification via CNN



2x 8-core 2.10 GHz Xeon (95W TDP)

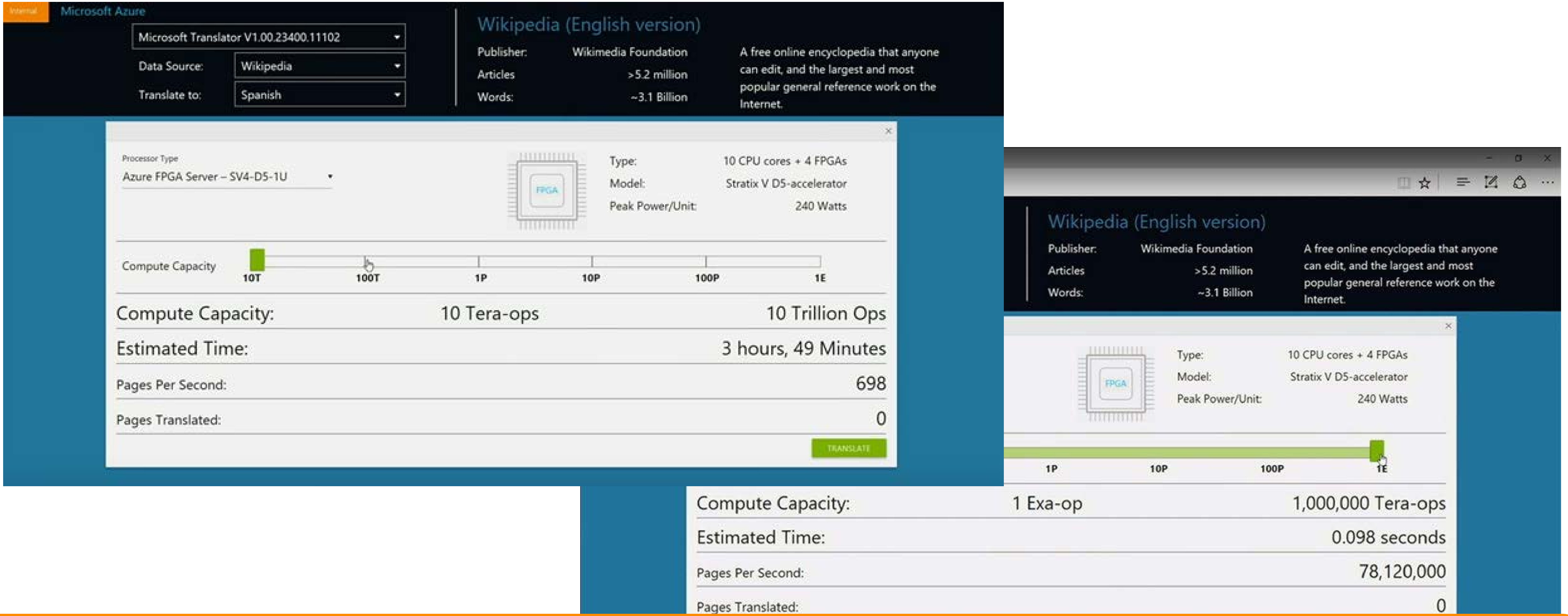
One Stratix V D5 FPGA (25 W)

Microsoft

Home | Agenda | Sessions | Expo | Pricing | MyIgnite | On-demand

Microsoft Ignite | September 26 – 30, 2016, Atlanta, GA

AI – Automatic Translation



- **>10x the AI capacity of the world's largest supercomputers with Catapult v2**
- **Next generation is more than 3x more powerful**

NEWS

Microsoft Claims Fastest Network in the Cloud

Microsoft CEO Satya Nadella demonstrated some of the AI supercomputing capabilities in its Azure public cloud that make it now among the fastest cloud services available. The secret sauce is the use of field-programmable gate arrays.

By Jeffrey Schwartz ■ 09/30/2016

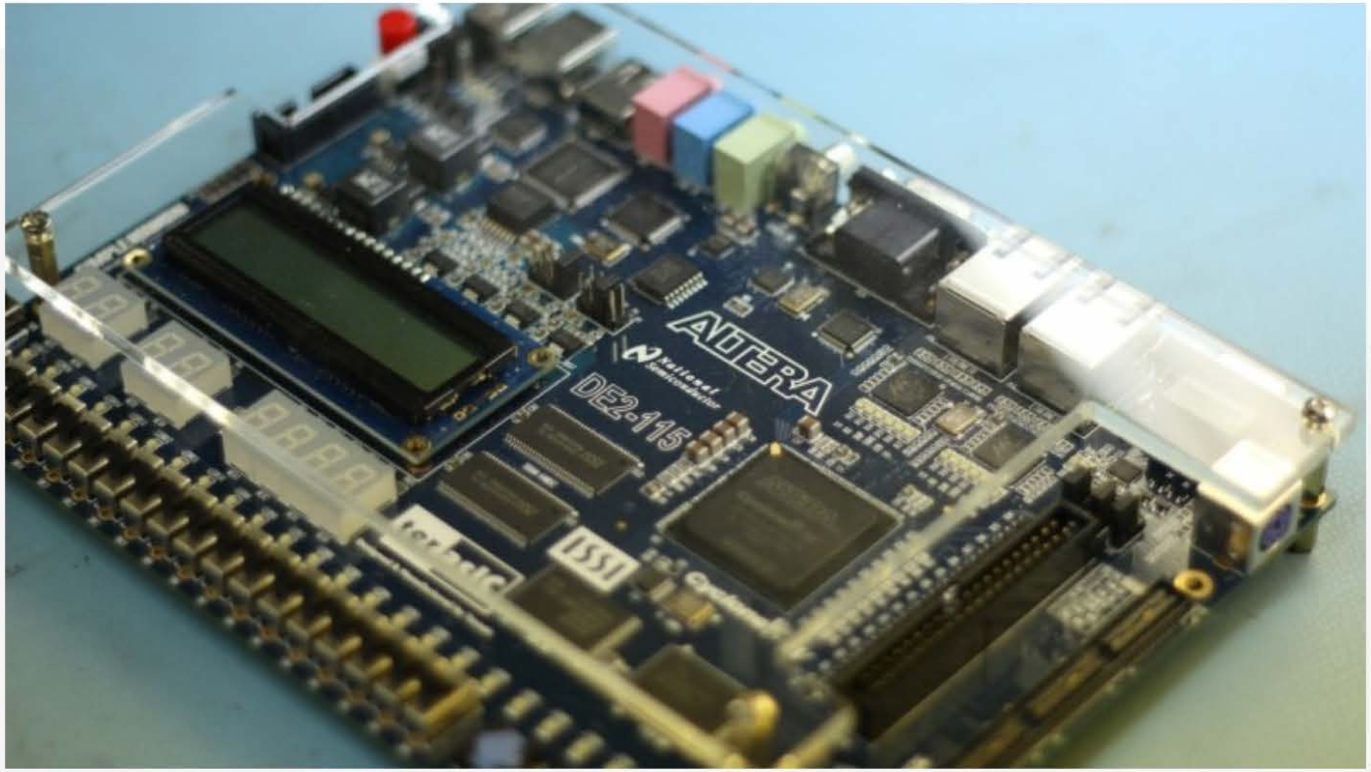
Microsoft CEO Satya Nadella demonstrated some of the AI supercomputing capabilities of the company's Azure platform. He said, in his keynote this week at Microsoft Ignite in Atlanta, that the company two years ago started upgrading every node in its Azure public cloud with software-defined network (SDN) infrastructure, developed using field-programmable gate arrays (FPGAs).

The result is that Microsoft's Azure public cloud fabric is now built on a 25 gigabit-per-second

INDUSTRY HARDWARE

Microsoft's Project Catapult is why Intel bought FPGA-maker Altera for $16.7 billion last year

By Shawn Knight on September 26, 2016, 4:45 PM



Intel last year acquired FPGA-maker Altera for $16.7 billion in cash, the chipmaker's largest purchase in history. As it turns out, Microsoft played a key role in Intel's decision to make the purchase.

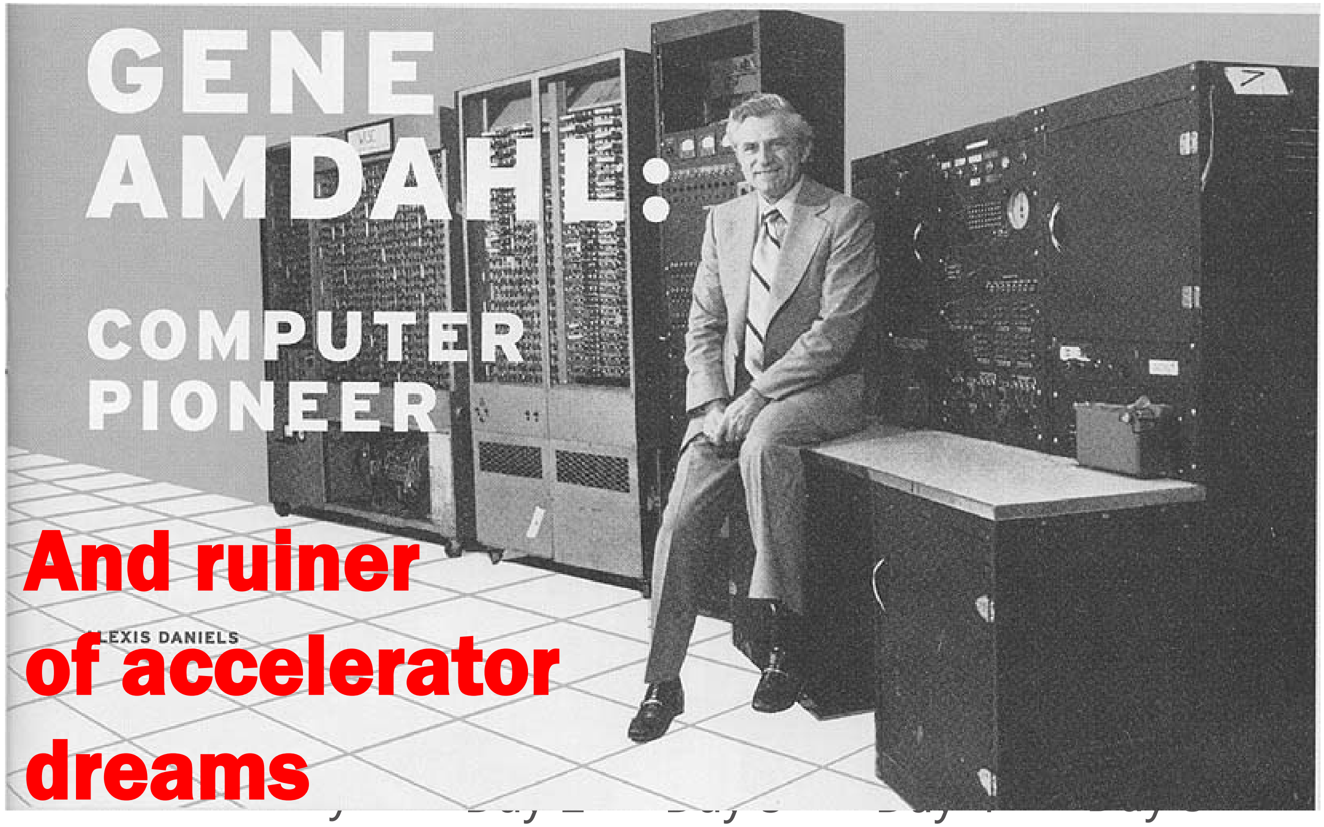
MOST POPULAR

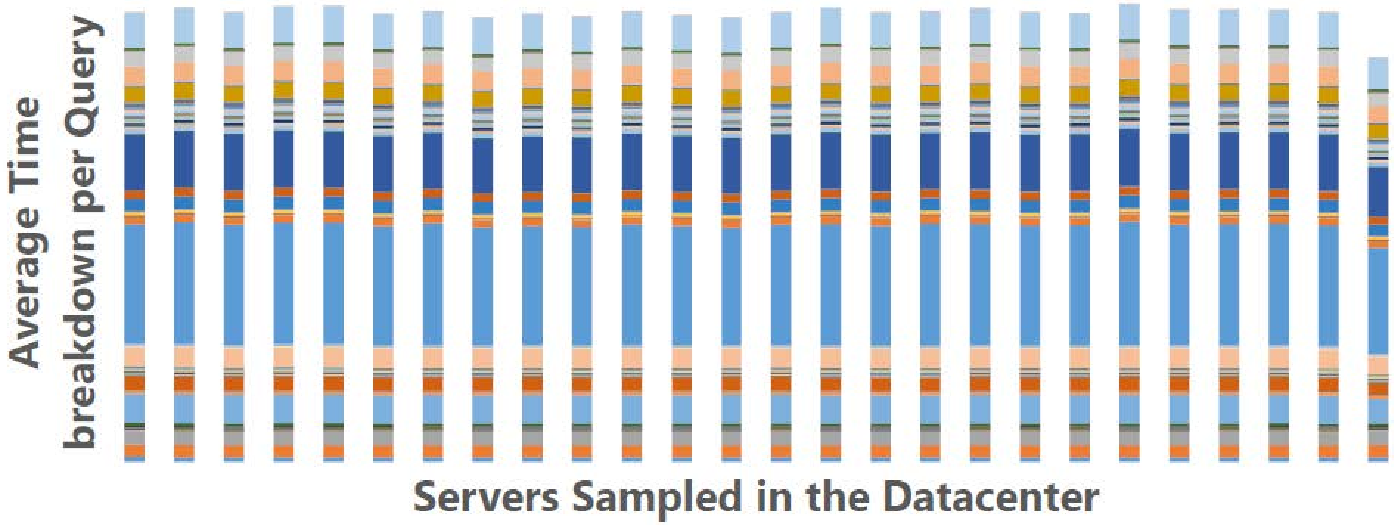
The Best PC Speakers


WHAT COULD POSSIBLY GO WRONG?

How do you program FPGAs?

- 10:1 or larger ratio of SW to HW programmers
- Verilog is the predominant programming language
- High-level synthesis (SDAccel, OpenCL, ROCCC, LegUp…) make it a little easier
  - Debugging still a challenge
  - Often requires platform-specific pragmas for good performance, which requires detailed knowledge of the FPGA architecture
- So far, most successful model is a custom contract between HW/SW, and still programming in Verilog


GENE
AMDAHL:
COMPUTER
PIONEER
LEXIS DANIELS
And ruiner
of accelerator
dreams


Average Time
breakdown per Query
Servers Sampled in the Datacenter

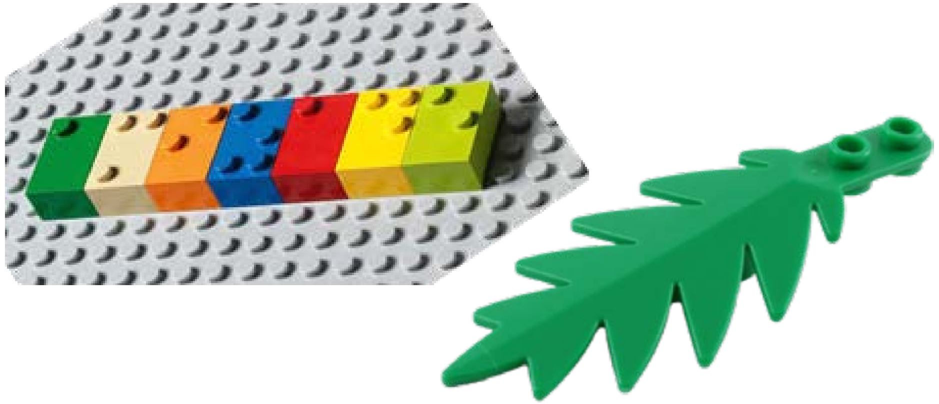
Usage Models

- Could allow users to have raw access to the FPGAs
  - Amazon EC2 F1
- High-Level Synthesis tools make this easier
- More common is a library of services built on top of FPGAs
  - DNN computation service, linear algebra, web search (Bing)…
- Infrastructure acceleration enables gradual migration to FPGAs

What am I worried about?

- I don't think the biggest problem is software engineers being able to program FPGAs
- I think our biggest problem is that we're going to make software engineers fight old battles
  - Libraries, linkers, backwards compatibility

OpenCL



Catapult Academic Program

- Donated 3 full racks of machines at TACC for research (96 machines per rack)
- Individual boards are available for in-house development use
- Very little effort to move from the academic cluster to the production machines

- See: http://research.microsoft.com/catapult for details

Conclusion

- Specialization with FPGAs is critical to the future Cloud
- FPGAs are harder to program, and improving that will greatly improve efficiency for Big Data / Cloud apps
- FPGAs allow optimization of both compute and I/O operations, so think beyond the core application
- What can you build with your global pile of Legos?



Microsoft