Microsoft

FPGA Datacenters for Hyperscale HPC

Andrew Putnam – Microsoft

Adrian Caulfield, Eric Chung, Hari Angepat, Daniel Firestone, Jeremy Fowers, Michael Haselman, Stephen Heil, Matt Humphrey, Puneet Kaur, Joo-Young Kim, Daniel Lo, Todd Massengill, Kalin Ovtcharov, Michael Papamichael, Lisa Woods, Sitaram Lanka, Derek Chiou, Doug Burger

HPC with the Cloud?

- The idea *sounds* great
- Pay for compute only when you use it
- When it breaks, it's someone else's problem
- No need to call the realtor / utility company when you want a bigger machine
- New hardware just shows up. No retrofits needed.

Questions from HPC to the Cloud

- Is the cloud really big enough for Exascale HPC?
- Are there examples of really big applications scaling on the cloud?
- Won’t virtual machines kill performance?
- Doesn’t HPC need fast specialized networks?
- Can clouds support specialized hardware?

Different Motivations

- HPC is focused on performance
  - Or at least performance within a certain energy budget
- Cloud is focused on supporting as many different customers as possible

ARM Clusters

Cloud Servers
with Accelerators

HPC Machines

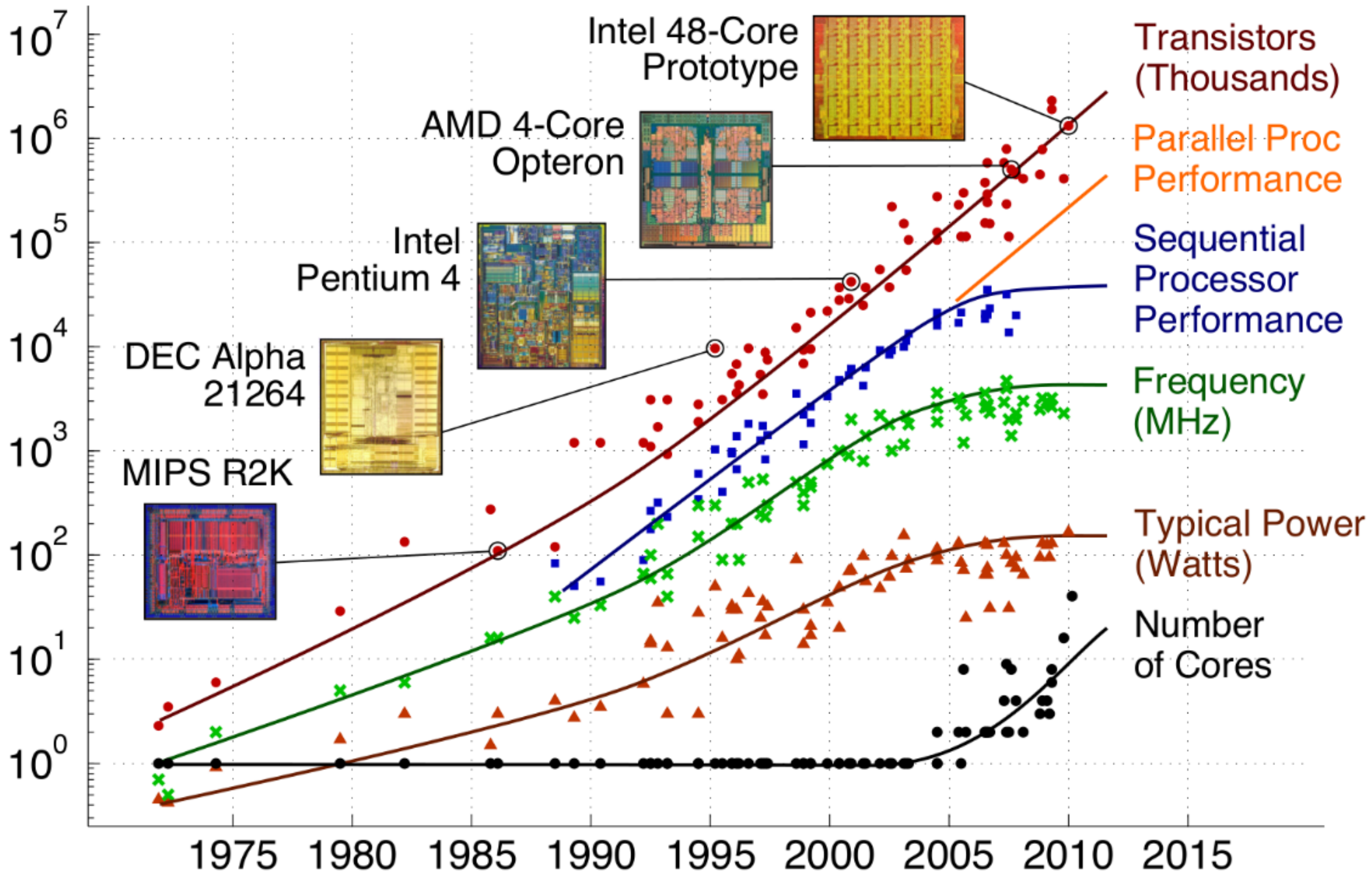
Questions from the Cloud to HPC*

- Is it *really* no expense spared?
- Isn't part of the need for maximum performance that you have to share the machine with others?
  - Run fast and get off
- How much use do older HPC machines really get?
- How much time is spent developing code specific to a single HPC machine?

*- I can't answer these. We need your input.

Scaling – What got us in this mess

- Moore's Law (transistors) is still alive
- Dennard Scaling (keeping energy under control) is dead
- Need improved performance, lower power … *Efficiency*



HPC changes since 2012

- Only 4 new top machines
- Most now have accelerators

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Titan** - Cray XK7, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x , Cray Inc. DOE/SC/Oak Ridge National Laboratory United States	560,640	17,590.0	27,112.5	8,209
2	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom , IBM DOE/NNSA/LLNL United States	1,572,864	16,324.8	20,132.7	7,890
3	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect , Fujitsu RIKEN Advanced Institute for Computational Science (AICS) Japan	705,024	10,510.0	11,280.4	12,660
4	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom , IBM DOE/SC/Argonne National Laboratory United States	786,432	8,162.4	10,066.3	3,945
5	**JUQUEEN** - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect , IBM Forschungszentrum Juelich (FZJ) Germany	393,216	4,141.2	5,033.2	1,970

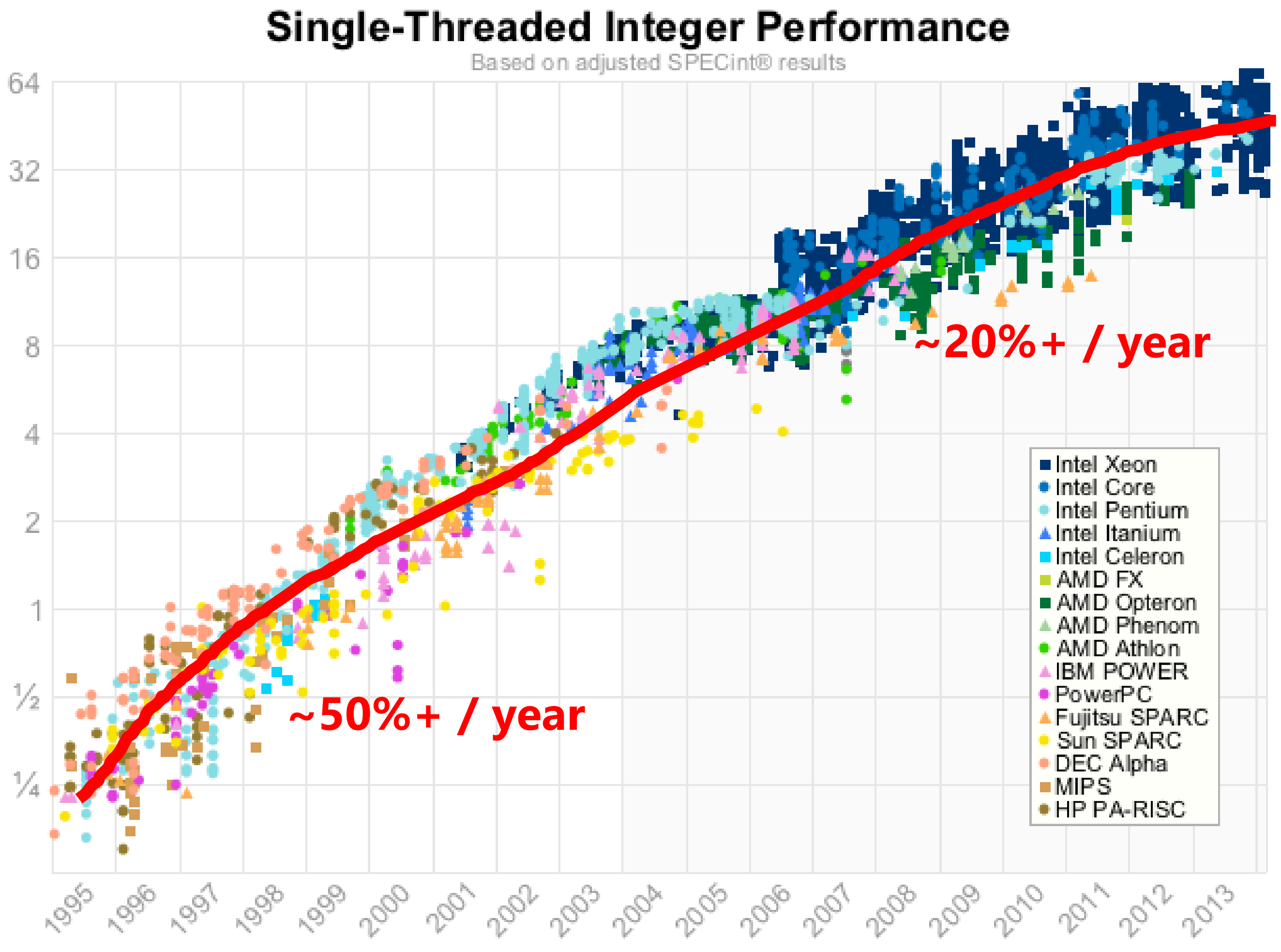
Nov. 2012

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway , NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
2	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P , NUDT National Super Computer Center in Guangzhou China	3,120,000	33,862.7	54,902.4	17,808
3	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 , Cray Inc. Swiss National Supercomputing Centre (CSCS) Switzerland	361,760	19,590.0	25,326.3	2,272
4	**Gyoukou** - ZettaScaler-2.2 HPC system, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 700Mhz , ExaScaler Japan Agency for Marine-Earth Science and Technology Japan	19,860,000	19,135.8	28,192.0	1,350
5	**Titan** - Cray XK7, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x , Cray Inc. DOE/SC/Oak Ridge National Laboratory United States	560,640	17,590.0	27,112.5	8,209

Nov. 2017

Technology Scaling for the Cloud

- Interactive Cloud apps rely on single threaded performance
- Performance depends on slowest 0.1% of machines (99.9%)
- 2x users requires 83% more cores



Jeff Preshing, Henk Poley, http://preshing.com/20120208/a-look-back-at-single-threaded-cpu-performance/

Cloud Server Changes

	2012	2017	Ratio
CPU Cores	16	36	2.25x
Storage	4 TB HDD	4 TB SDD 32 TB HDD	9x
Network	1Gb	50Gb	**50x**

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

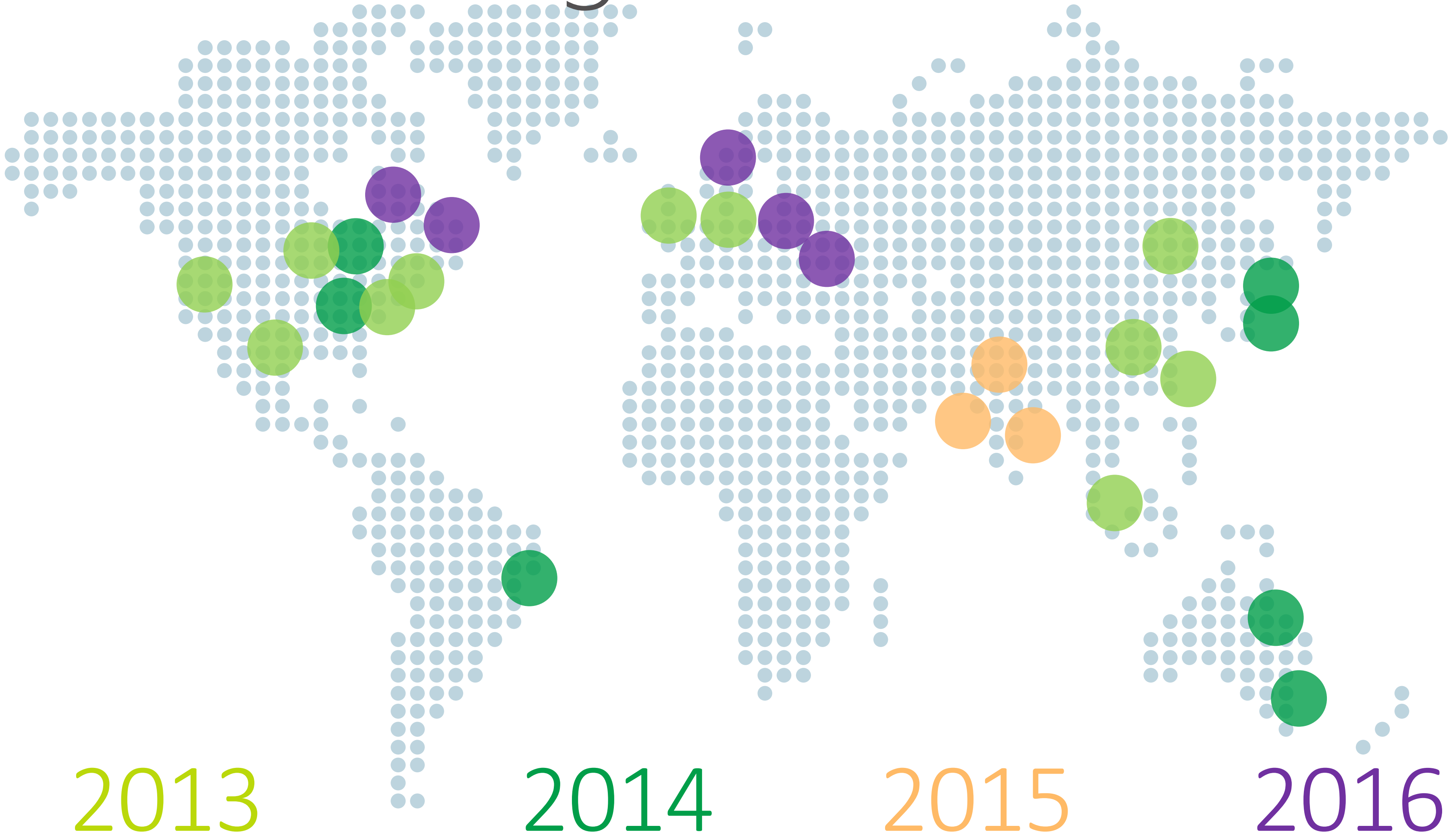
40,960 single-socket servers

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

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

98,304 single-socket servers

Datacenter Scaling



~100%+ Growth for the past 5 years

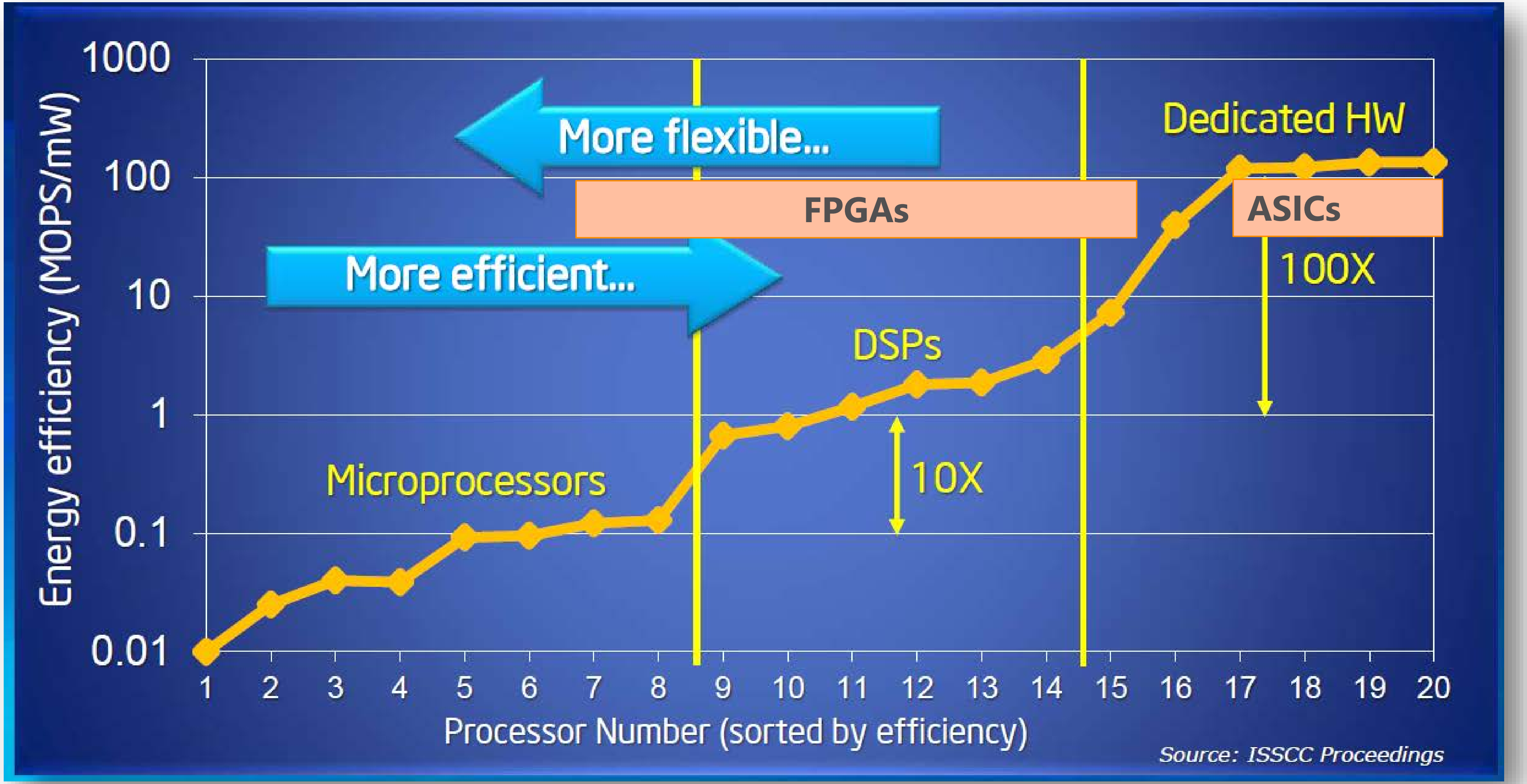
	2012	2017
Compute Instances	100K	Millions
Azure Storage	10’s of PB	Exabytes
Datacenter Network	10’s of Tbps	Pbps

Questions from HPC to the Cloud

- ✓Is the cloud really big enough to handle Exascale HPC?
- Are there examples of really big applications scaling on the cloud?
- Won't virtual machines kill performance?
- Doesn't HPC need fast specialized networks?
- Can clouds support specialized hardware?

But scaling with *efficient* computing is much cheaper than simply buying more hardware!

Efficiency via Specialization



Source: Bob Broderson, Berkeley Wireless group

Silicon Technologies for Computing

More Flexible

Homogeneous

	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.		

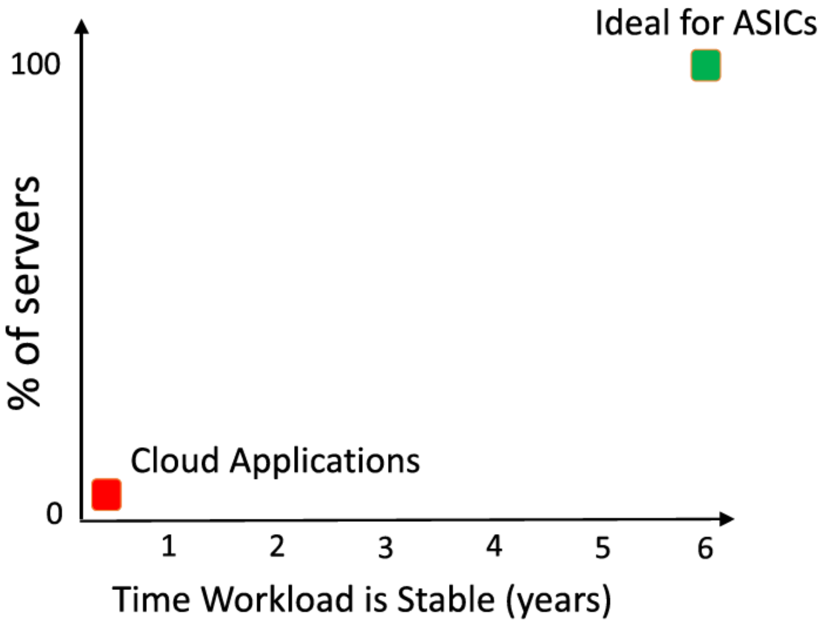
Specialized

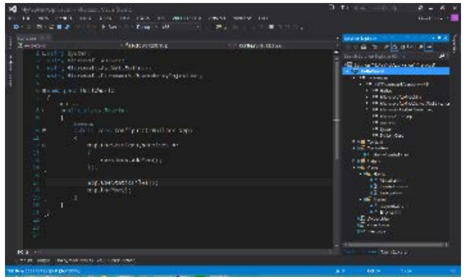
More Efficient

Why not use GPUs?

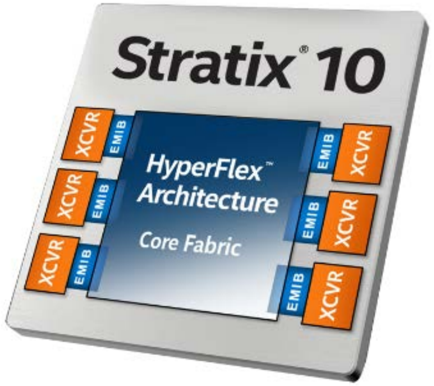
- Power
- Customer-facing (interactive) workloads are small batches, need **low latency**
- Power & Cost mean limited deployments (HPC only)
- *Optimize for the whole fleet, not for one application*

Why not use ASICs?

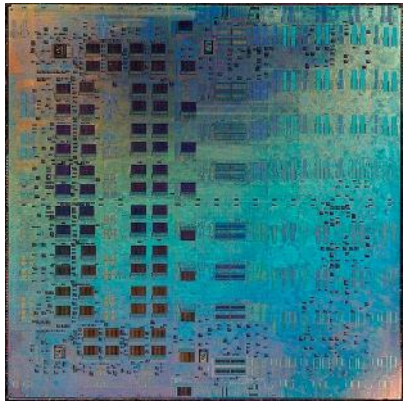




Software



FPGA

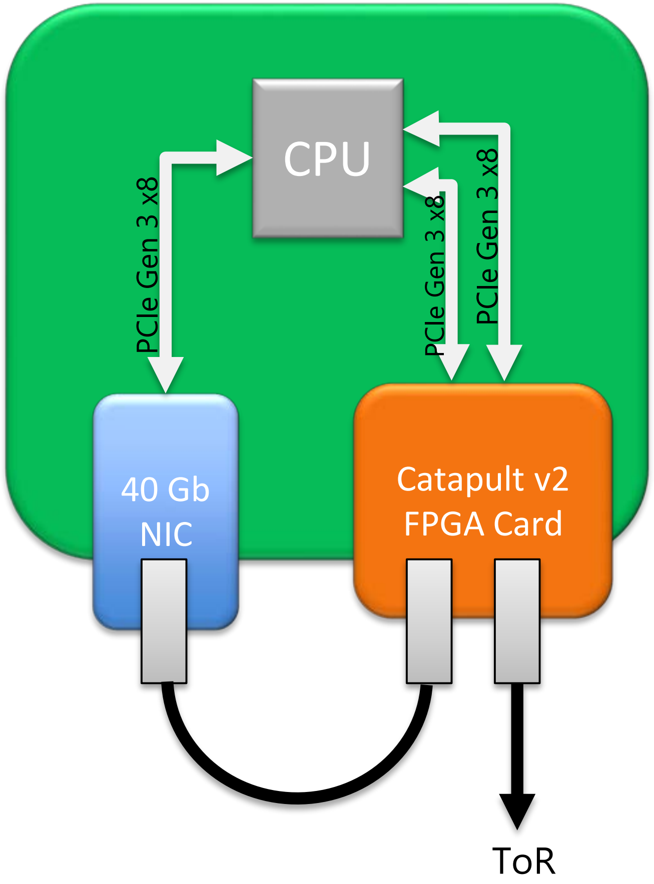


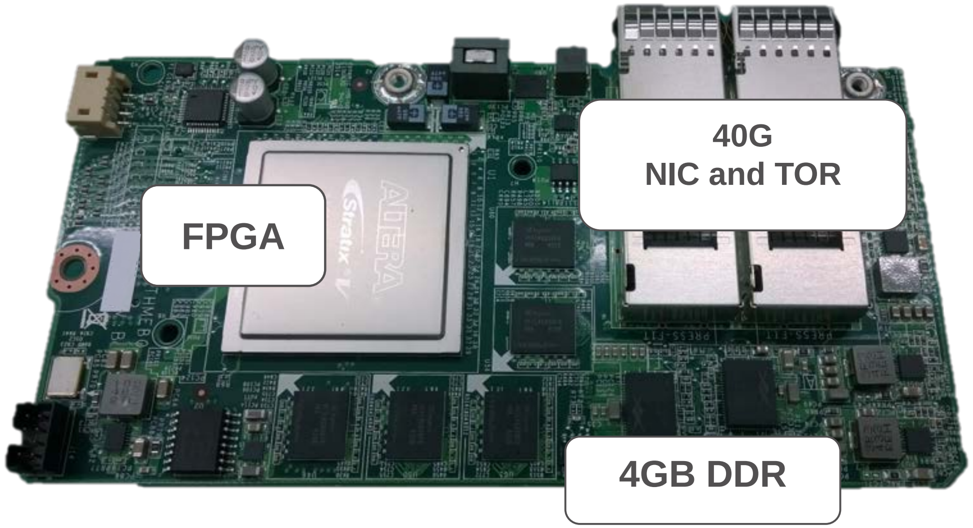
ASIC

Fitting FPGAs in the Datacenter

- All servers should be the same
- One FPGA per server keeps servers homogeneous
- Area must be small. Temperatures high. Power low.

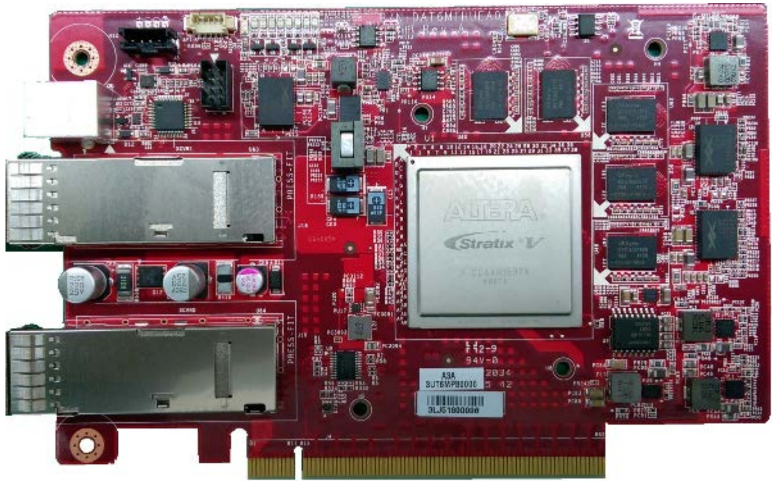
Catapult v2 – Bump in the Wire



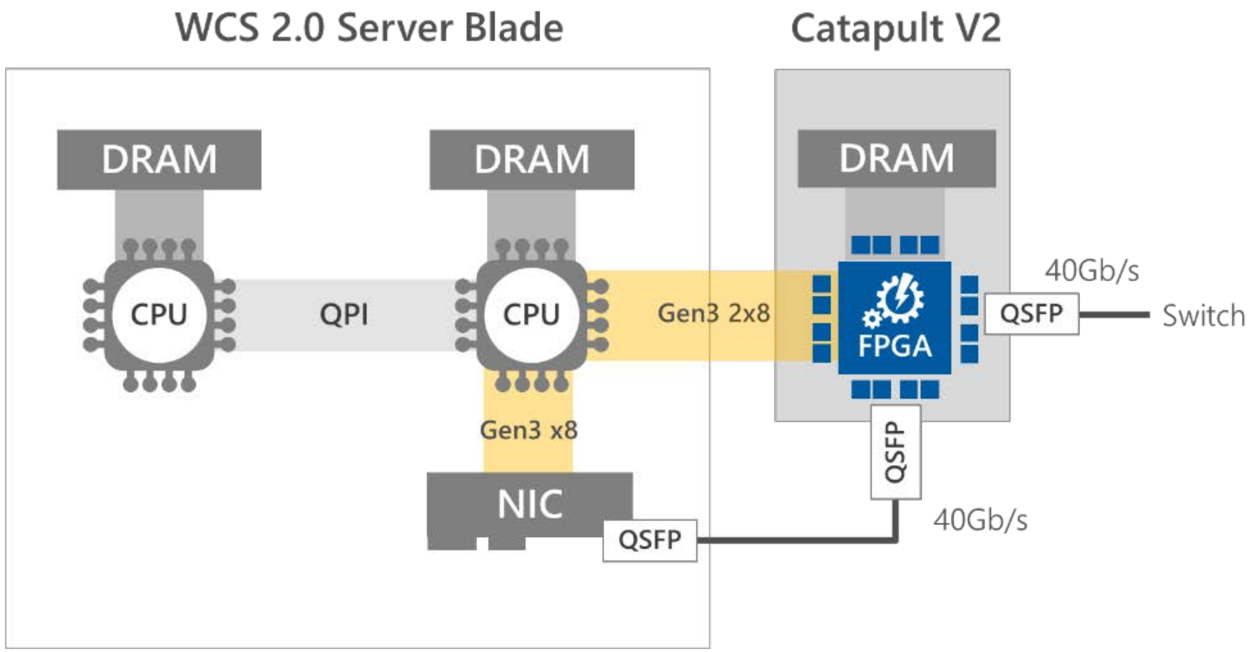


WCS Mezz

General PCIe

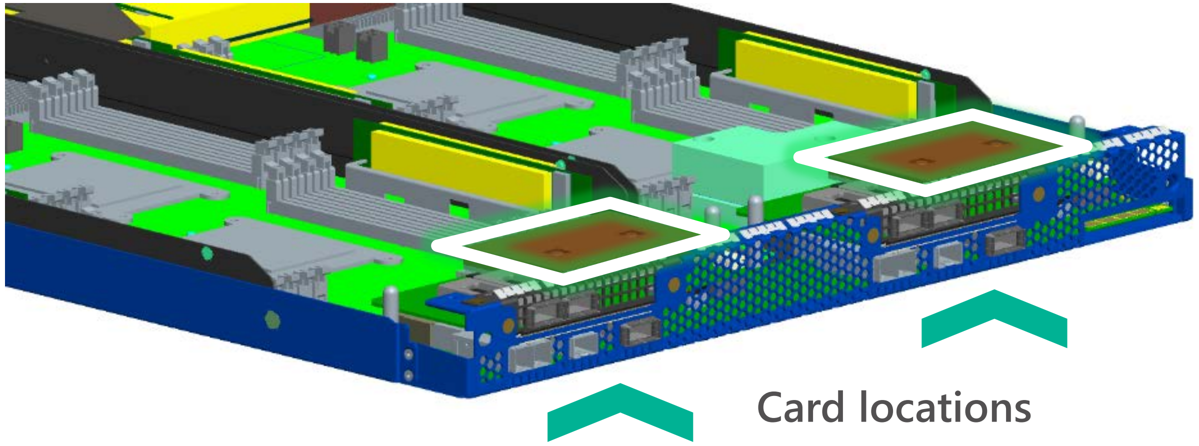


FPGAs Are Deployed in MSFT Servers Worldwide



Catapult v2 Mezzanine card

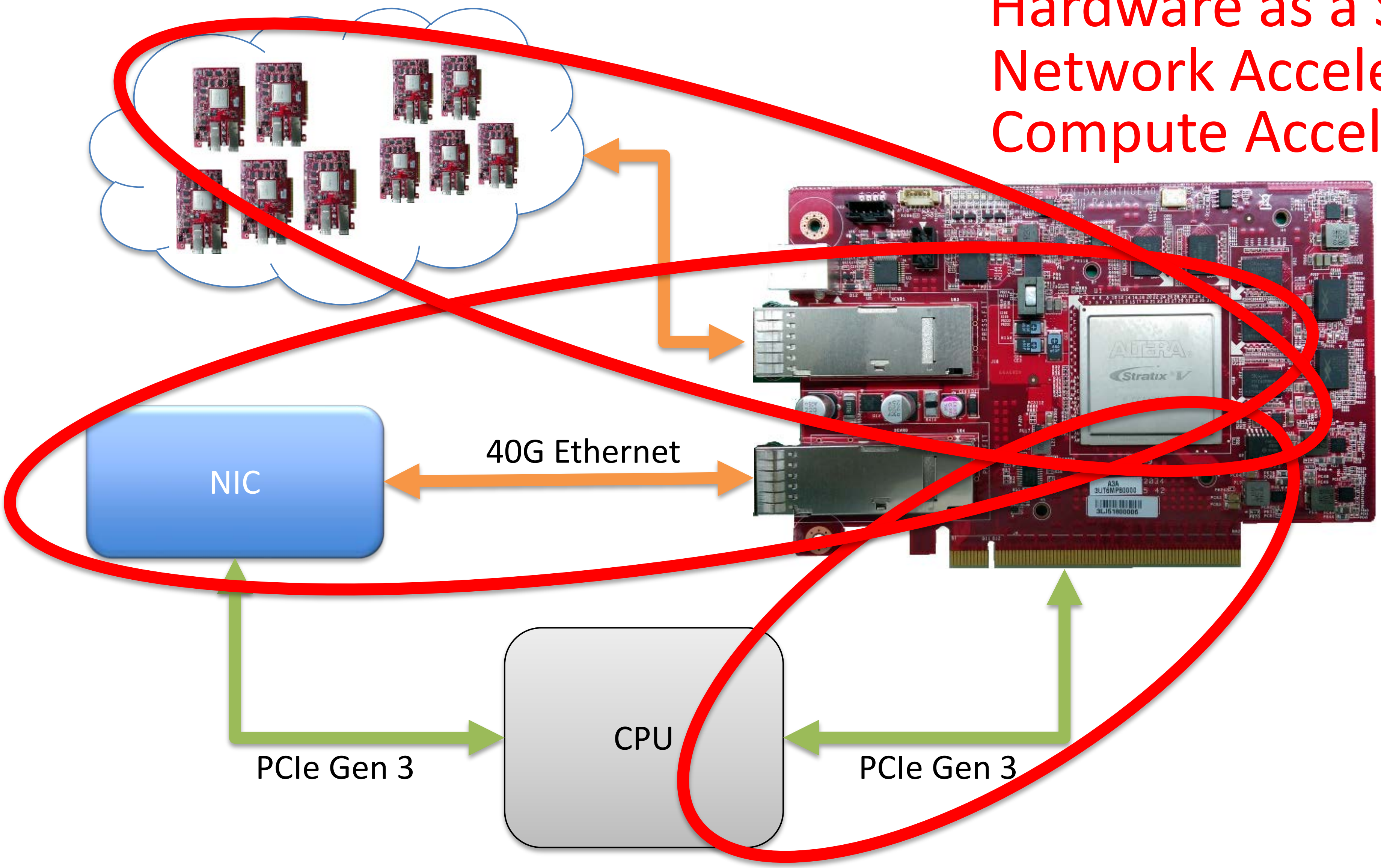
WCS Gen4.1 Blade with NIC and Catapult FPGA



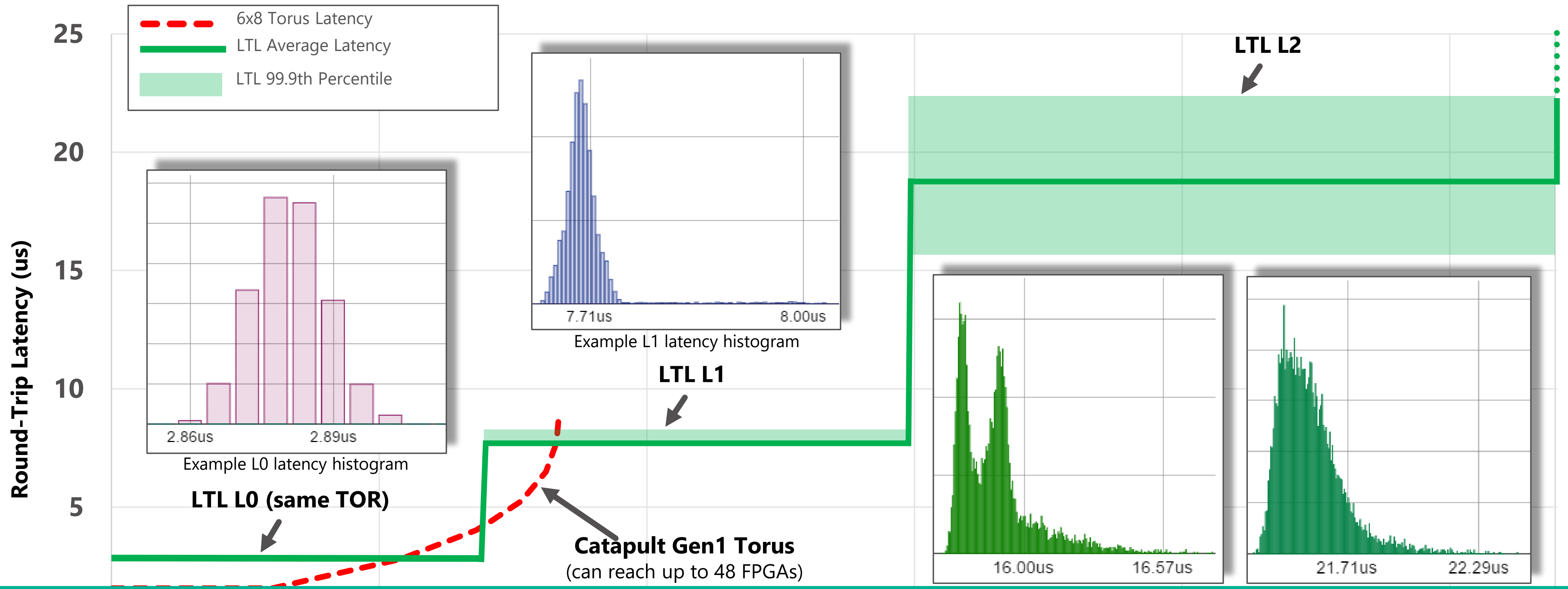
[ISCA'14, HotChips'14, MICRO'16]

Bump-in-the-wire Architectue

Hardware as a Service
Network Acceleration
Compute Acceleration



Network Latencies



- Extremely low latency (Similar to Infiniband)
- Add compute into the network

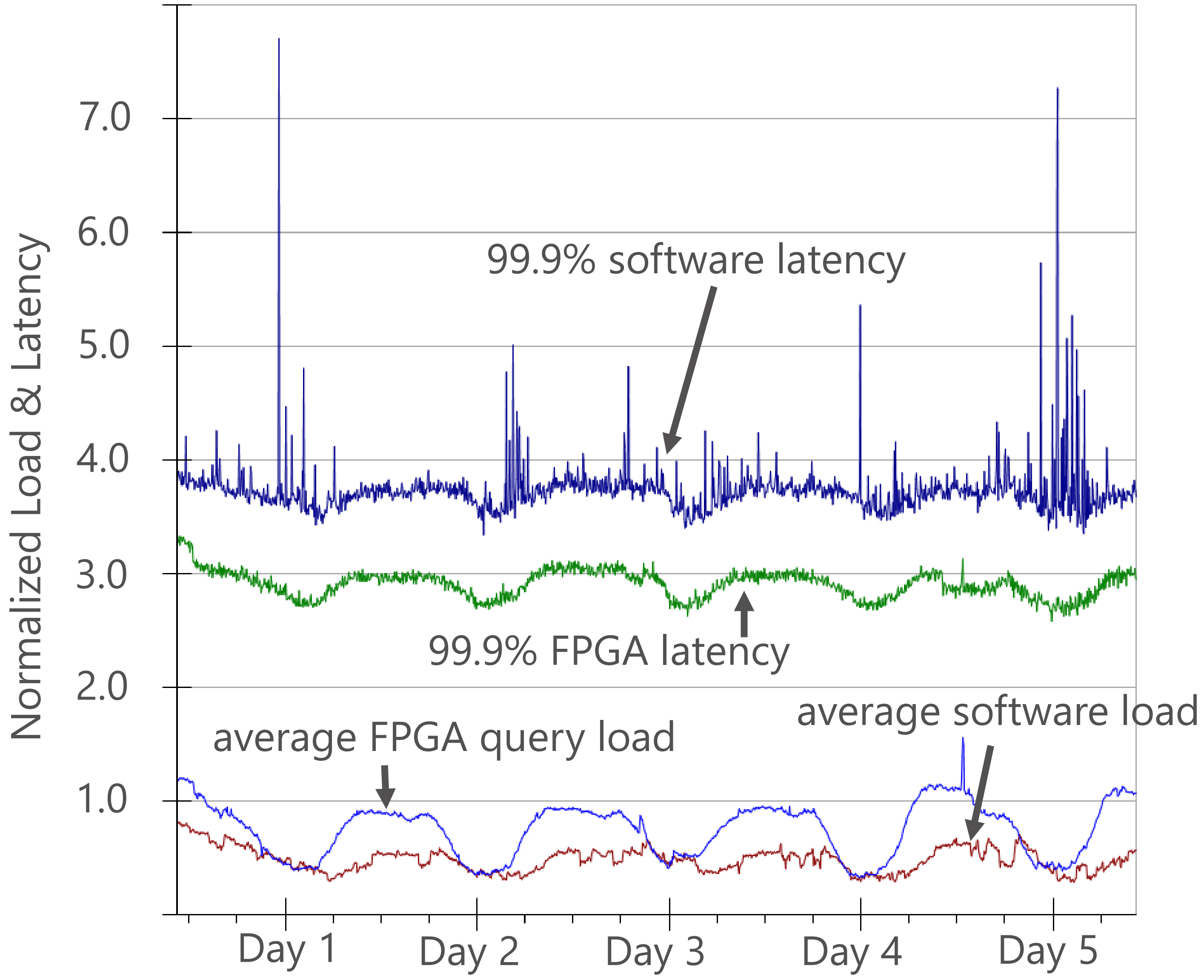
- **FPGAs Included in every new server for all major services**
- Deployed across 16 countries and 6 continents
- Already in large scale production in both Bing and Azure

Questions from HPC to the Cloud

- ✓Is the cloud really big enough to handle Exascale HPC?
- Are there examples of really big applications scaling on the cloud?
- Won’t virtual machines kill performance?
- Doesn’t HPC need fast specialized networks?
- Can clouds support specialized hardware?

Compute Acceleration -- Bing Ranking

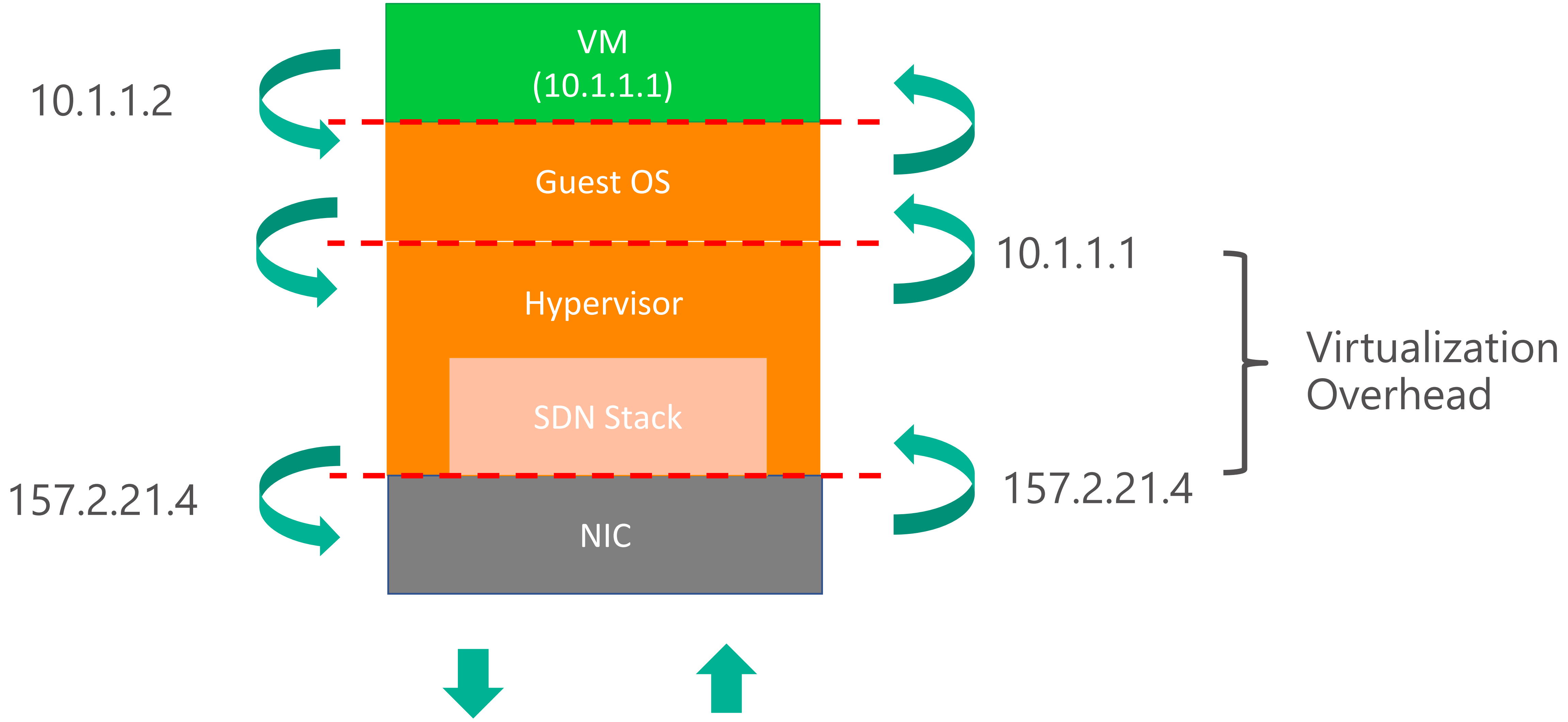
- 2x Faster at 2x higher load
- Much lower variance
- 1,632+ machines per instance



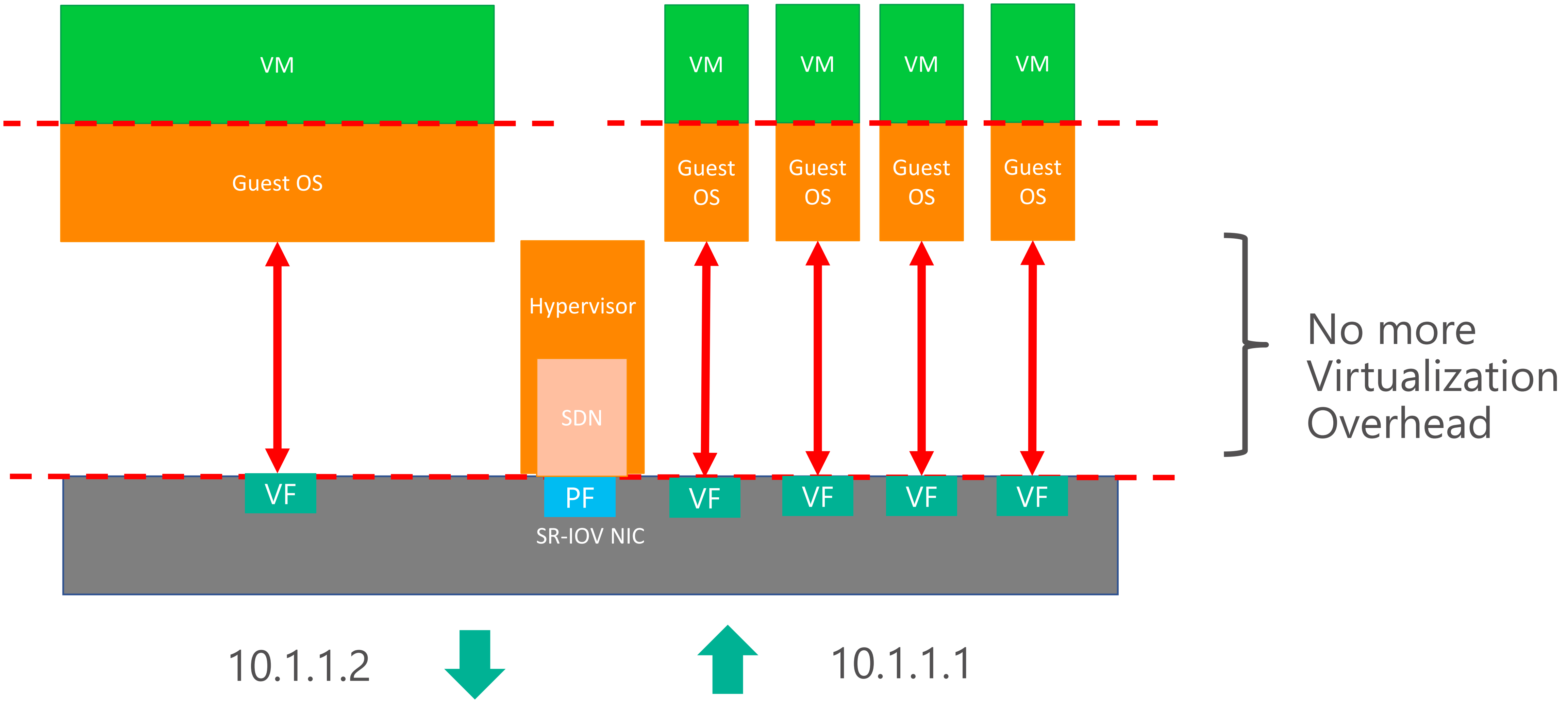
Questions from HPC to the Cloud

- ✓ Is the cloud really big enough to handle Exascale HPC?
- ✓ Are there examples of really big applications scaling on the cloud?
- Won't virtual machines kill performance?
- Doesn't HPC need fast specialized networks?
- Can clouds support specialized hardware?

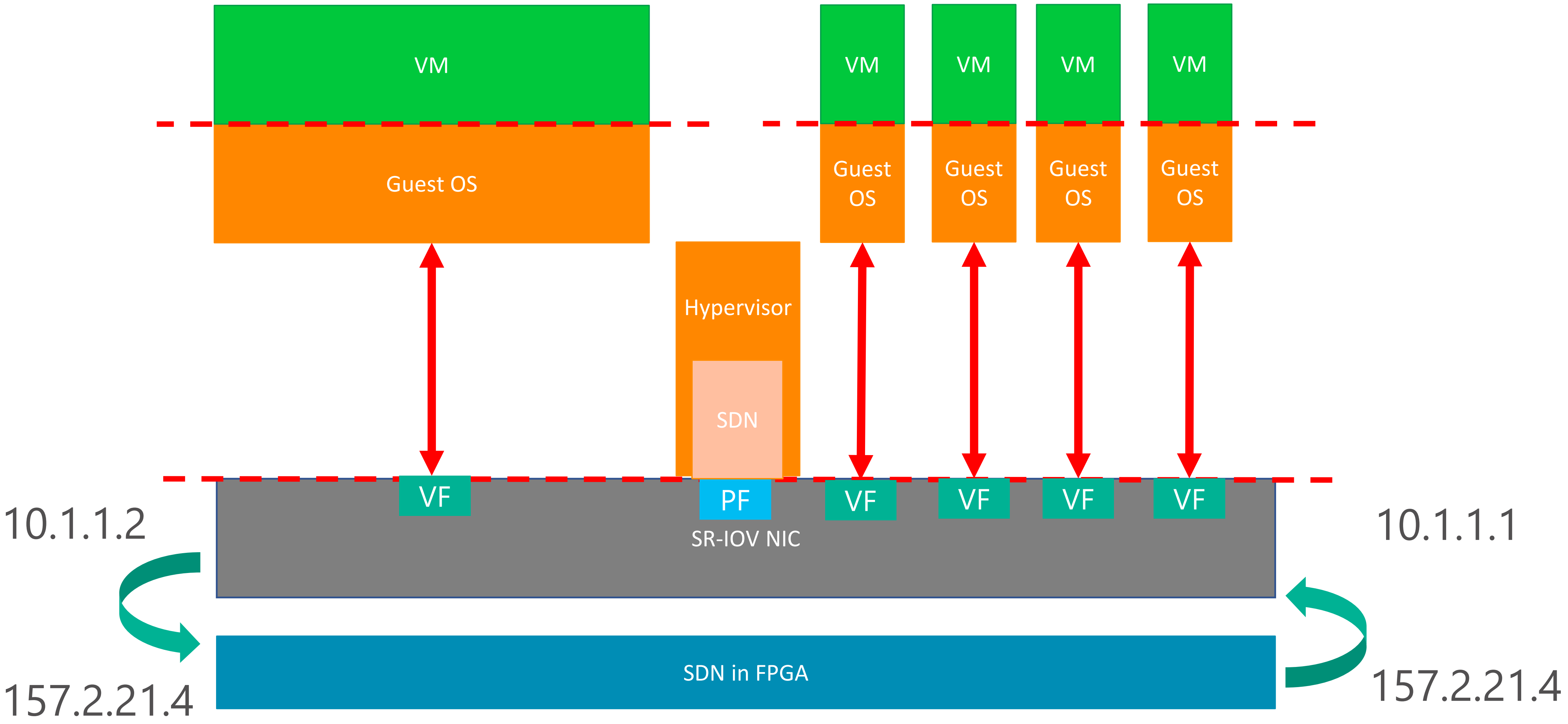
Dealing with Virtualization



Dealing with Virtualization



Dealing with Virtualization

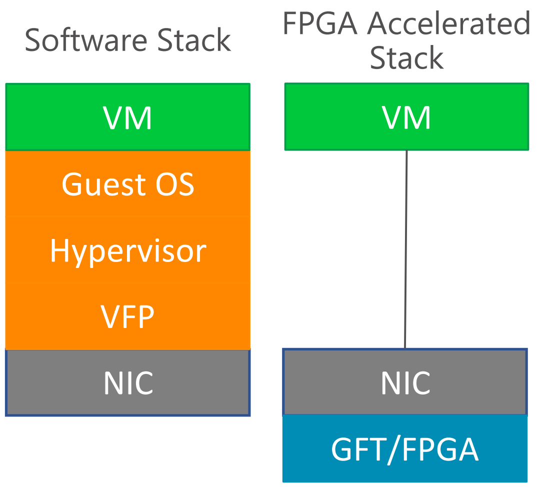


Infrastructure Acceleration

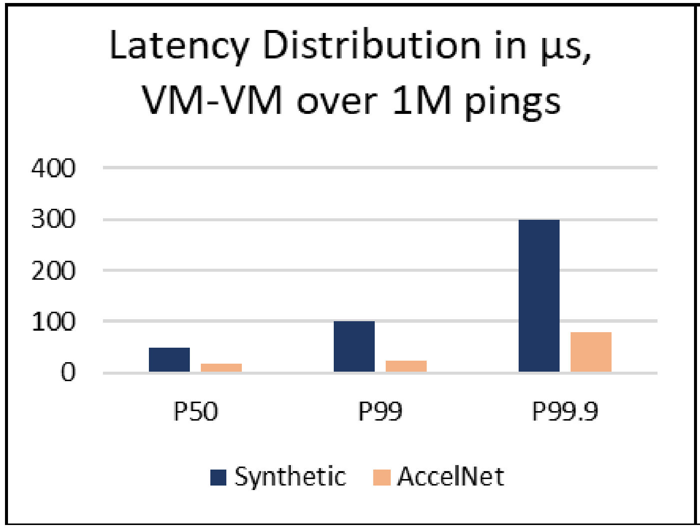
Microsoft Azure

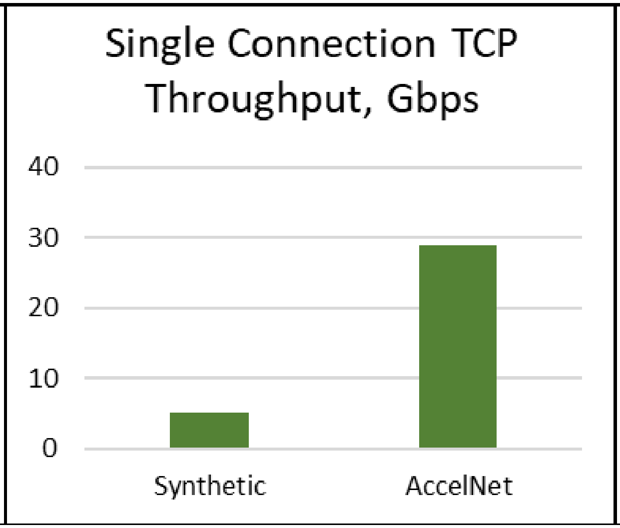
SmartNIC: SDN and Crypto offload

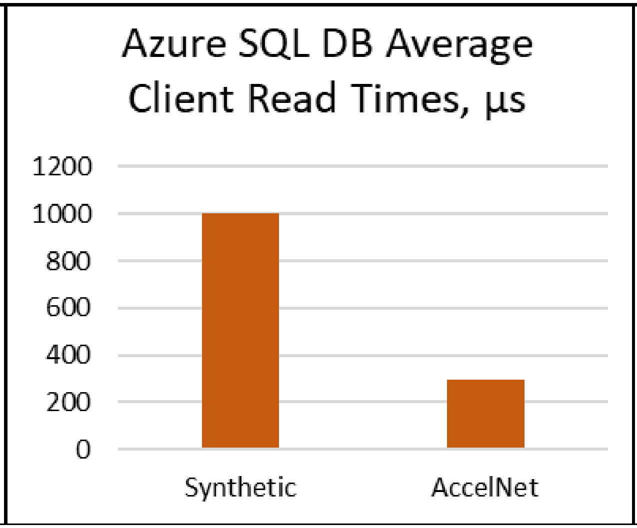
- Cut out most of the software stack
- Generic Flow Table (GFT) rule based packet rewriting
- Enhanced network security
- 10x latency reduction vs software
- >25Gb/s throughput at 10s of µs latencies – **the fastest cloud network**
- Free to customers
- Storage works similarly

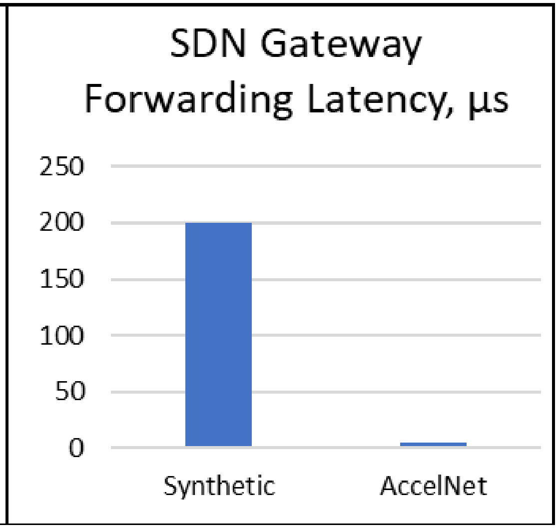


AccelNet Performance







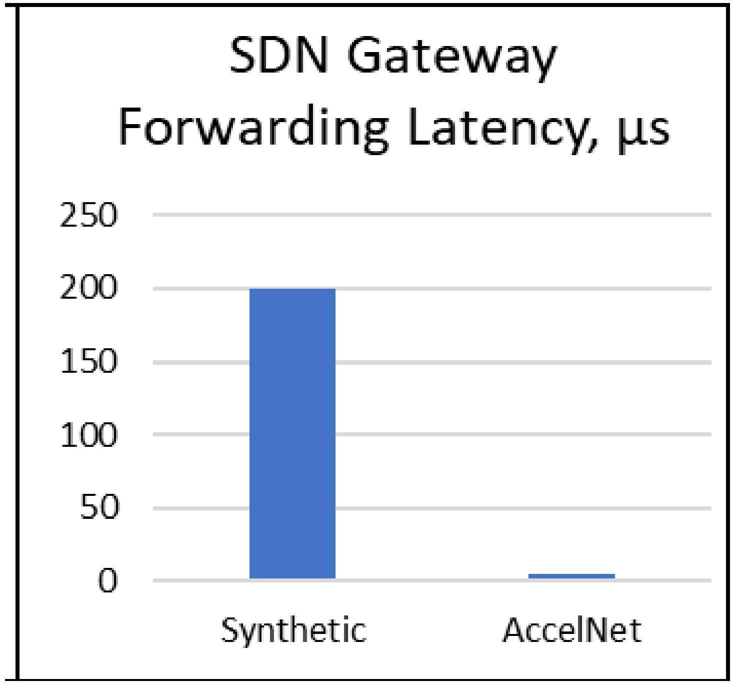


Questions from HPC to the Cloud

- ✓Is the cloud really big enough to handle Exascale HPC?
- ✓Are there examples of really big applications scaling on the cloud?
- ✓Won't virtual machines kill performance?
- ✓Doesn't HPC need fast specialized networks?
- Can clouds support specialized hardware?

Adding Heterogeneous Processing

- Fast network means that nearly any machine in the data is accessible in under 22 us
- Most datacenters will have HPC clusters
  - GPUs are the most common
- Reach HPC clusters with low, predictable latency
- Azure Stack and ExpressRoute Gateways allow for hybrid clouds



Optimizing for FPGAs over CPUs

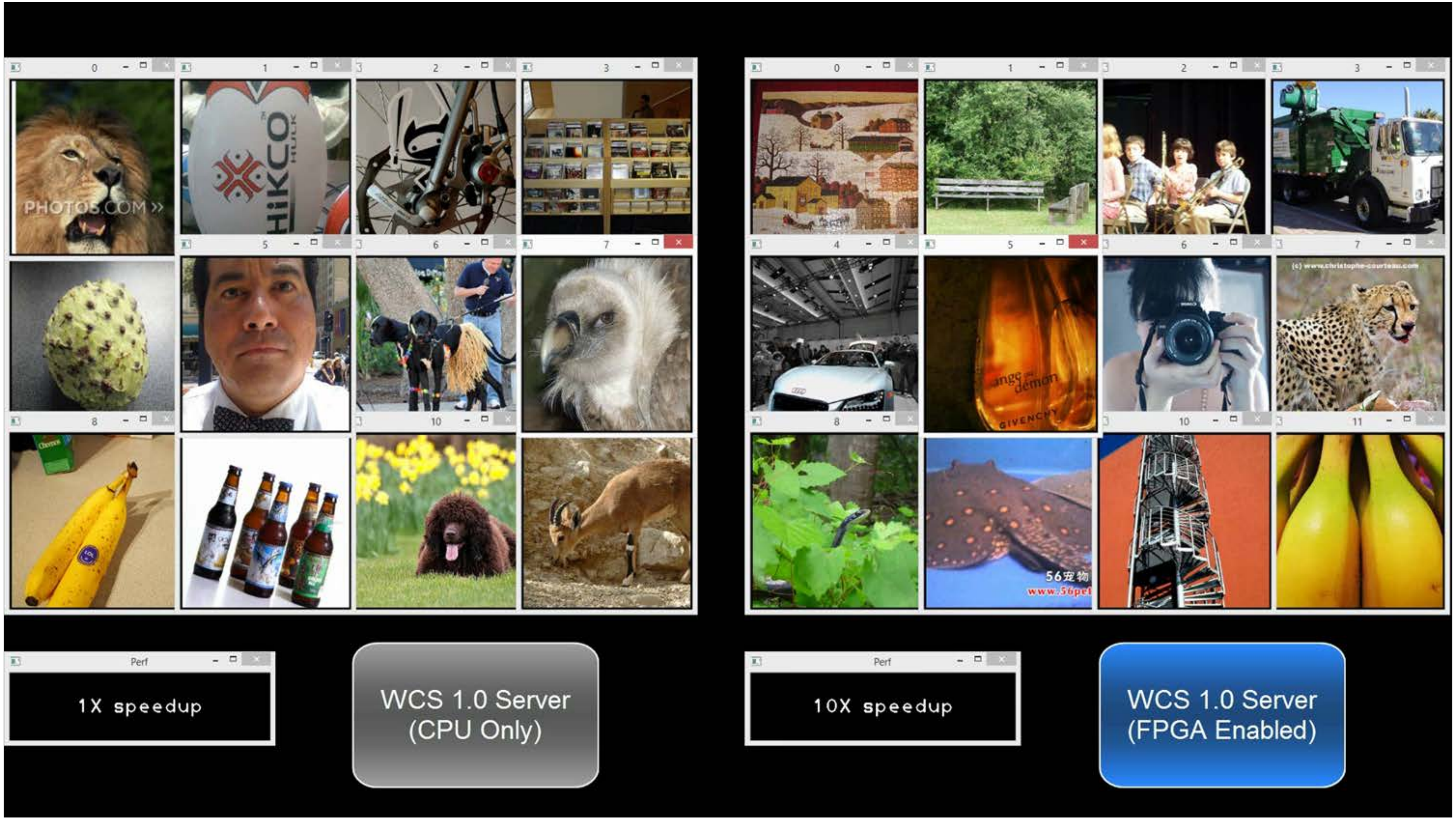
- No reason we can’t make FPGA-heavy HPC clusters
- Deploying multiple FPGAs per server allows for a higher than 1:1 ratio of FPGAs to compute
- Bing is starting to target this style of architecture, so HPC won’t be the first to try

Questions from HPC to the Cloud

- ✓Is the cloud really big enough to handle Exascale HPC?
- ✓Are there examples of really big applications scaling on the cloud?
- ✓Won’t virtual machines kill performance?
- ✓Doesn’t HPC need fast specialized networks?
- ✓Can clouds support specialized hardware?

What about AI / Deep Learning?

Deep Learning -- Image Classification via CNN



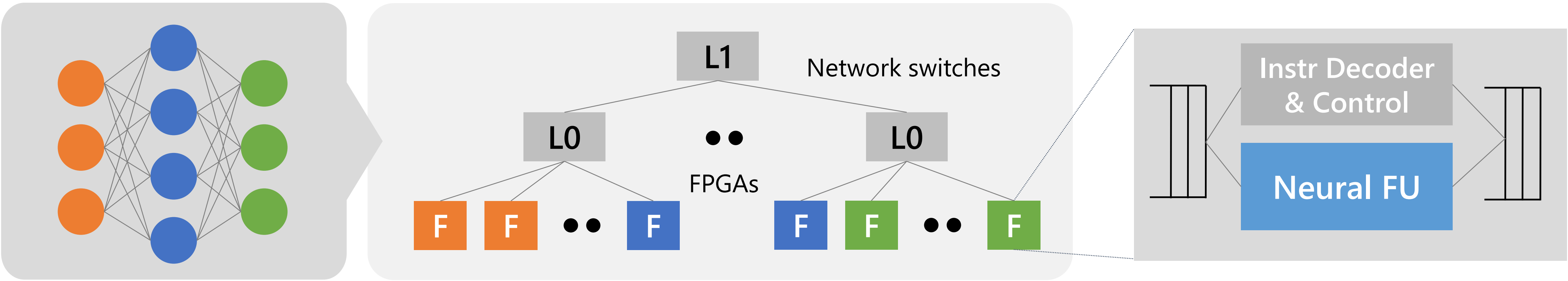
2x 8-core 2.10 GHz Xeon (95W TDP)

One Stratix V D5 FPGA (25 W)

Project BrainWave

A Scalable FPGA-powered DNN Serving Platform

Fast: ultra-low latency, high-throughput serving of DNN models at low batch sizes
Flexible: adaptive numerical precision and custom operators
Friendly: turnkey deployment of CNTK/Caffe/TF/etc

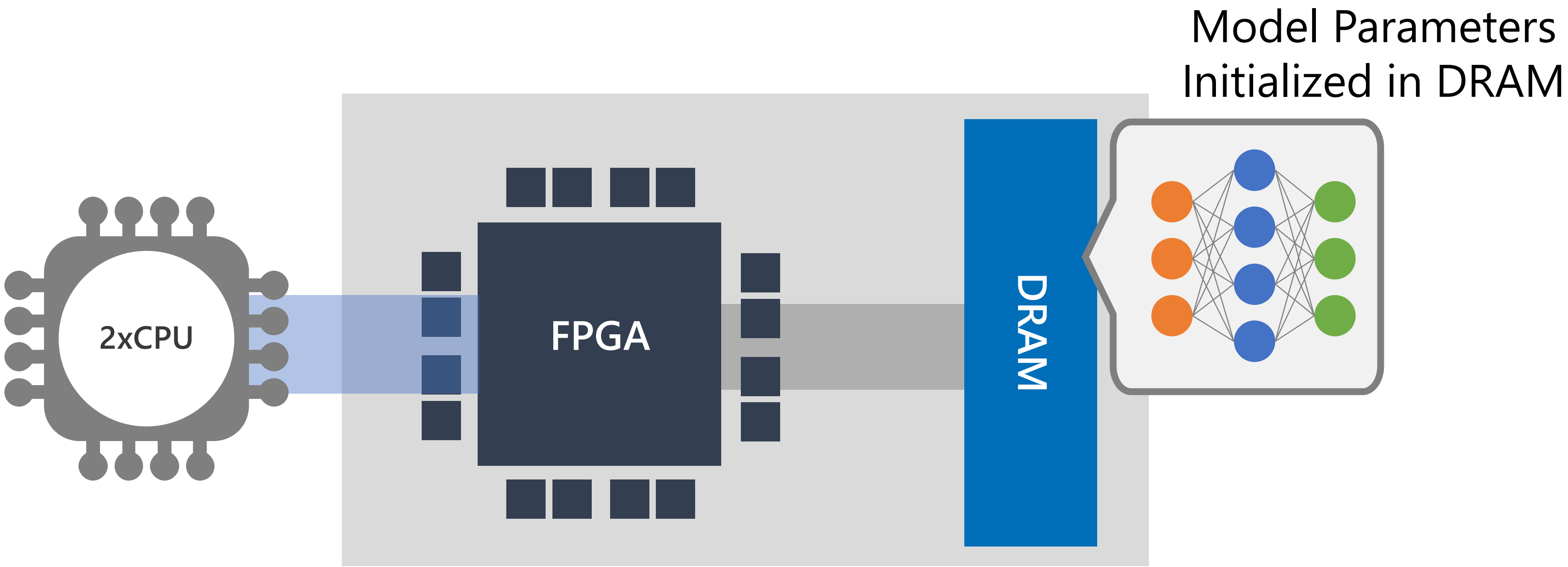


Pretrained DNN Model in CNTK, etc.

Scalable DNN Hardware Microservice

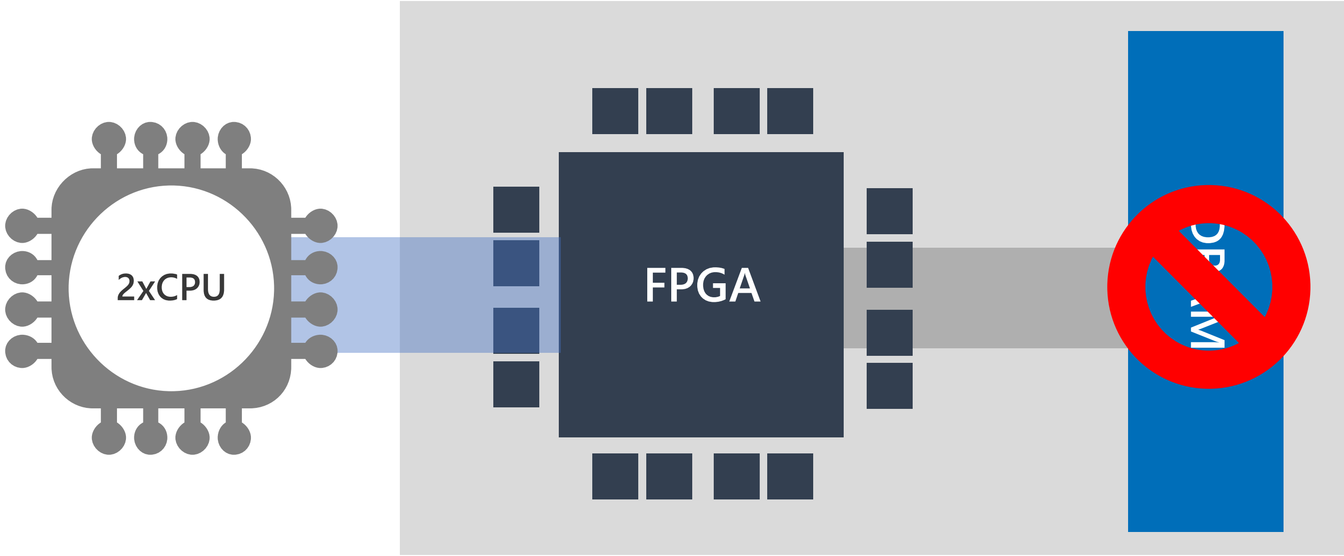
BrainWave Soft DPU

Conventional Acceleration Approach: Local Offload and Streaming

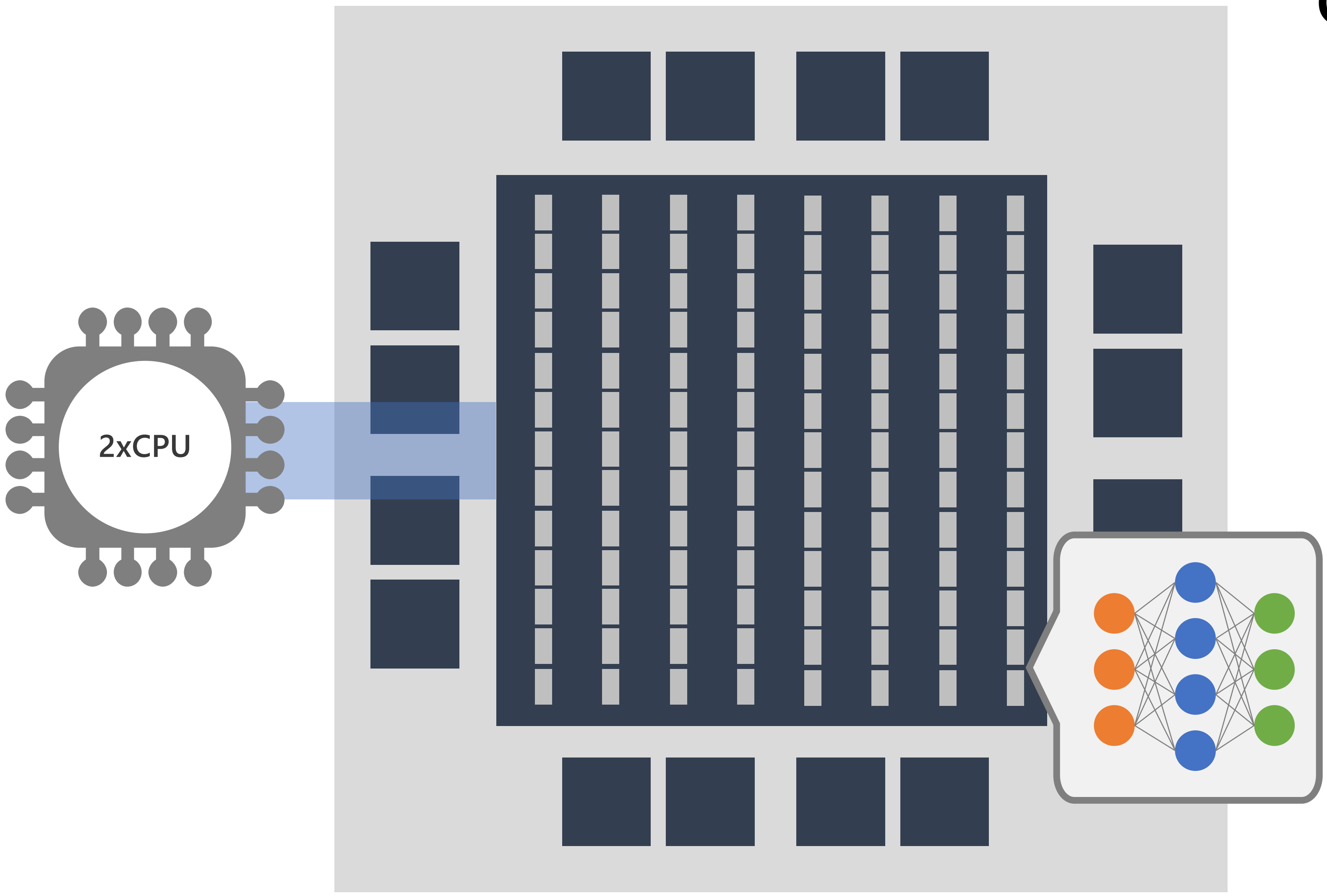


For memory-intensive DNNs with low compute-to-data ratios (e.g., LSTM), HW utilization limited by off-chip DRAM bandwidth

Alternative: “Persistent” Neural Nets



Alternative: "Persistent" Neural Nets



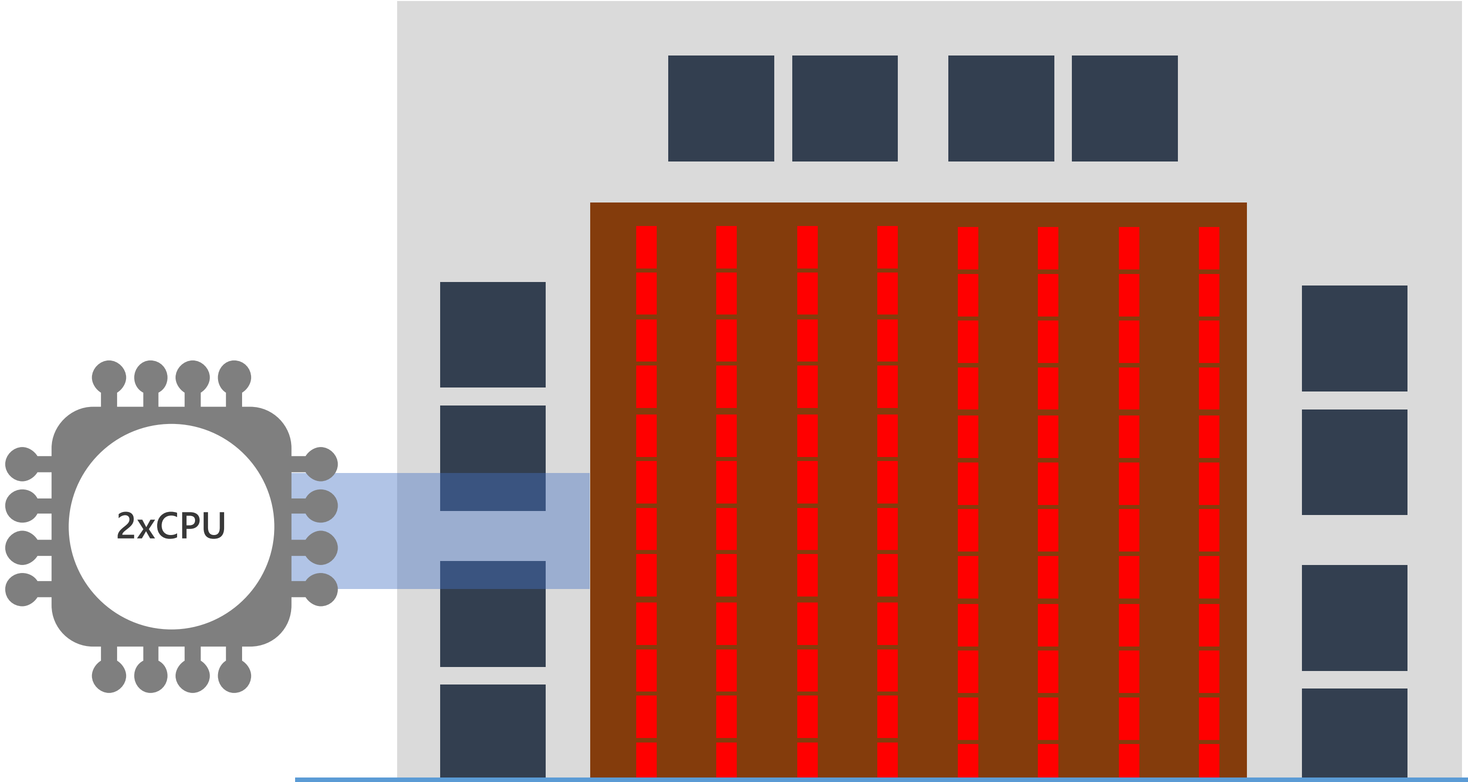
Observations

State-of-art FPGAs have O(10K) distributed Block RAMs O(10MB)
➔ Tens of TB/sec of memory BW

Large-scale cloud services and DNN models run persistently

Solution: persist all model parameters in FPGA on-chip memory during service lifetime

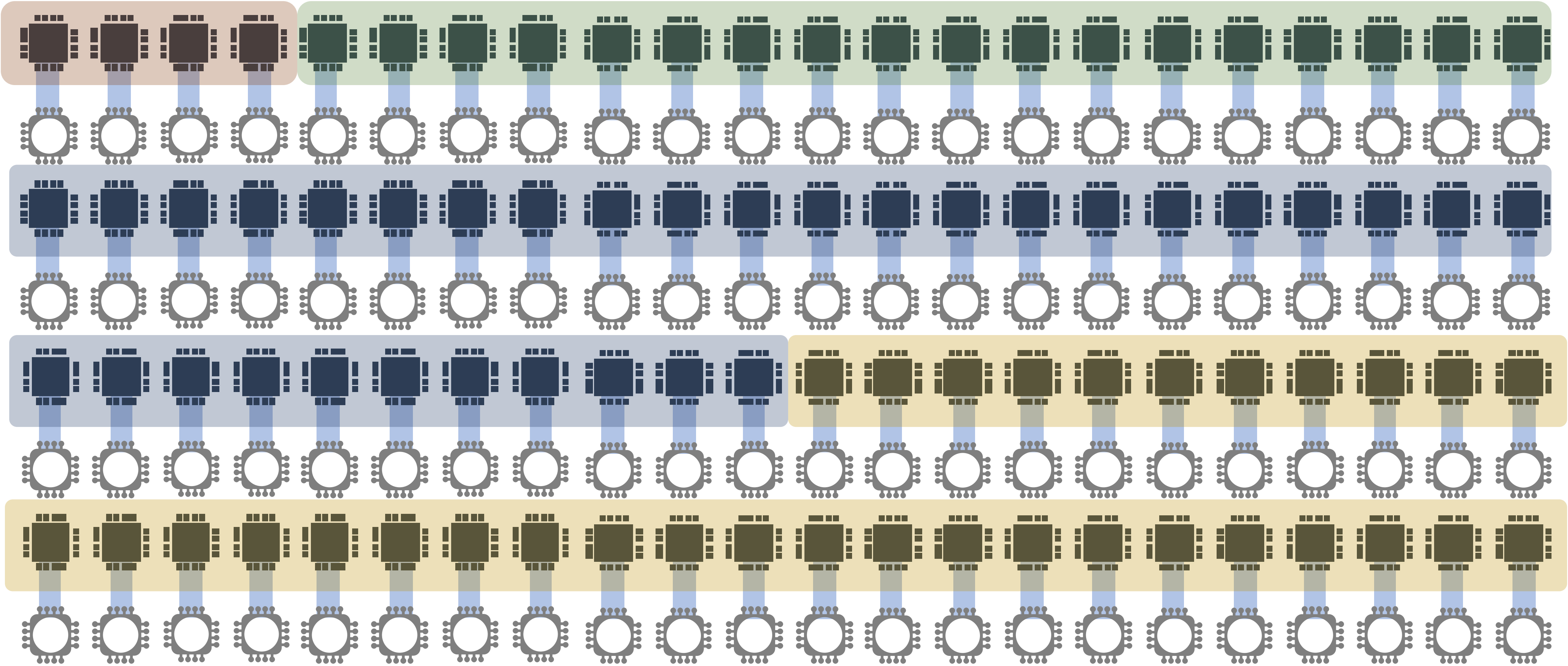
Alternative: "Persistent" Neural Nets



When single request arrives, all chip resources (on-chip memories and compute units) are used to process a single query (no batching required)

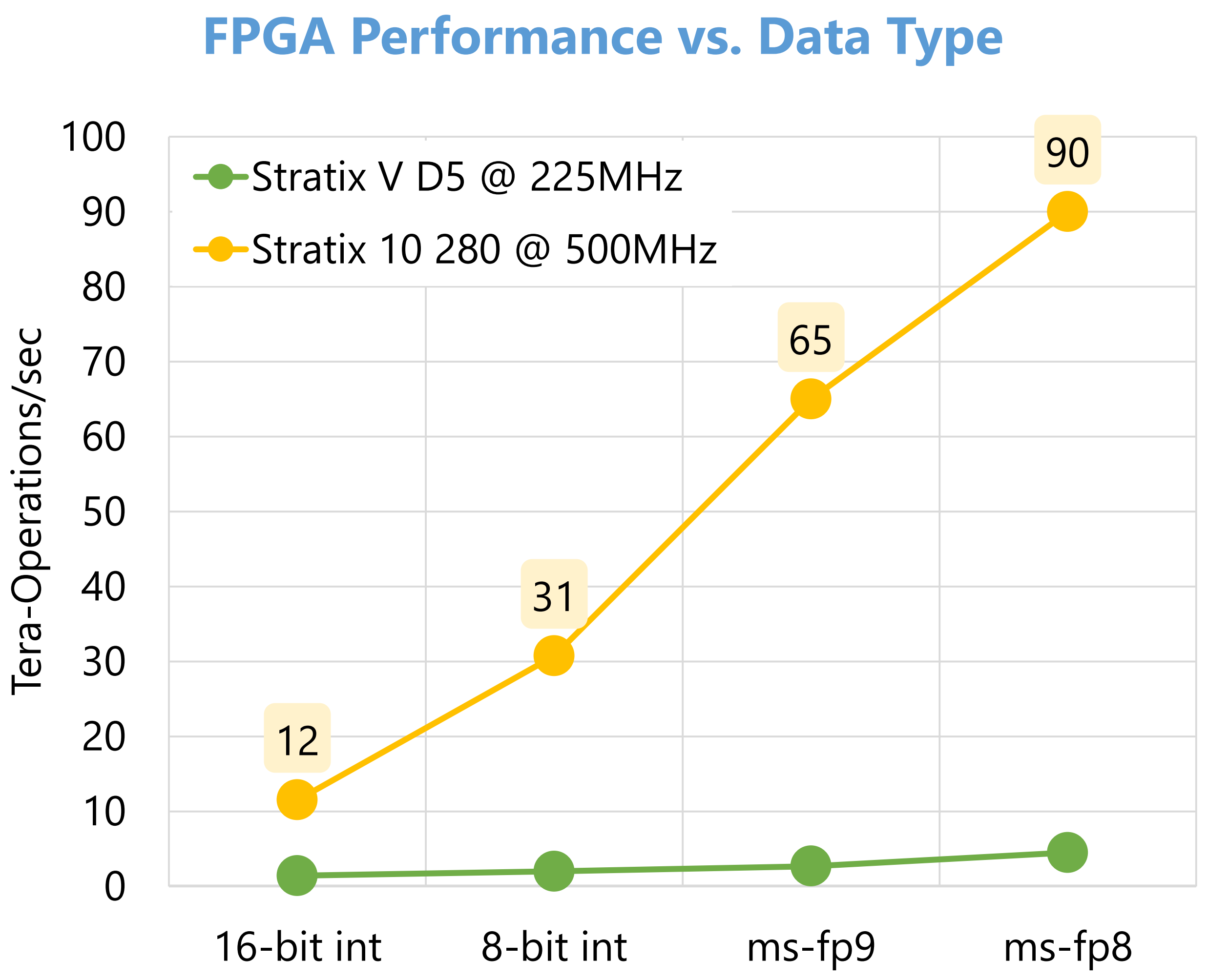
What if model doesn’t fit in single FPGA?

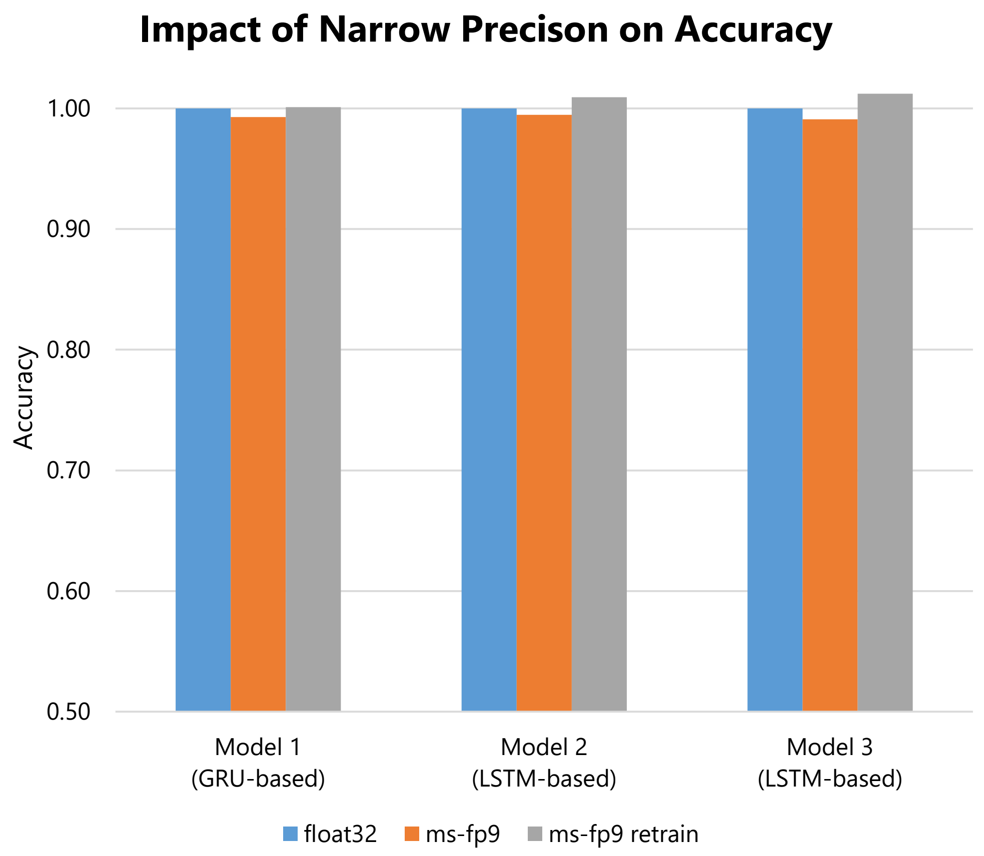
Solution: Persistency at Datacenter Scale



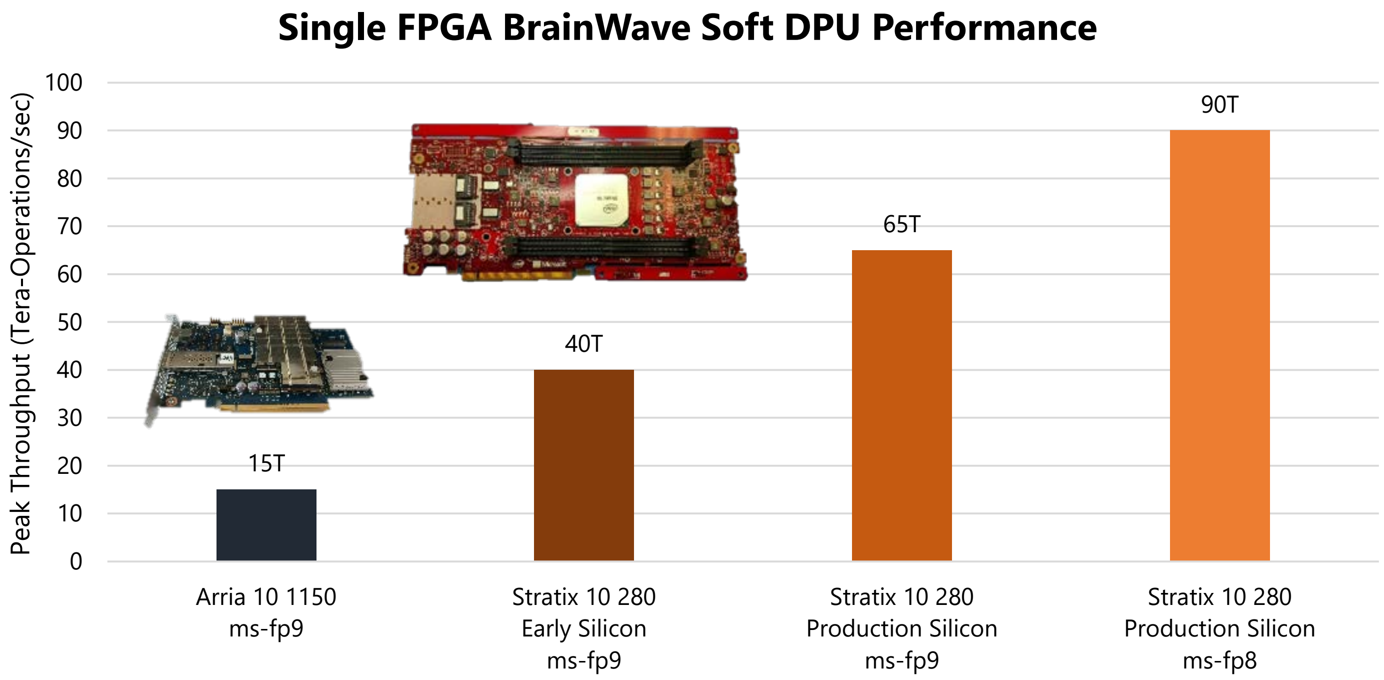
Multiple FPGAs at datacenter scale can form a persistent DNN HW microservice, enabling scale-out of models at ultra-low latencies

Narrow Precision Inference on FPGAs



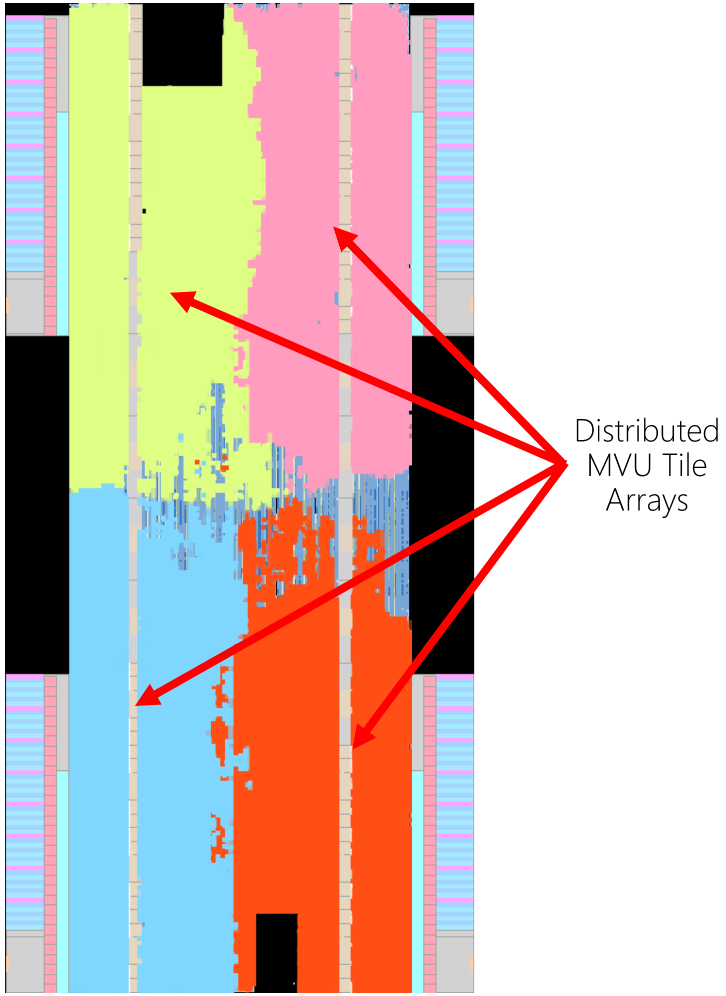


BrainWave Soft DPU Performance



Arria 10 1150 (20nm)
ms-fp9
316K ALMs (74%)
1442 DSPs (95%)
2,564 M20Ks (95%)
160 GOPS/W

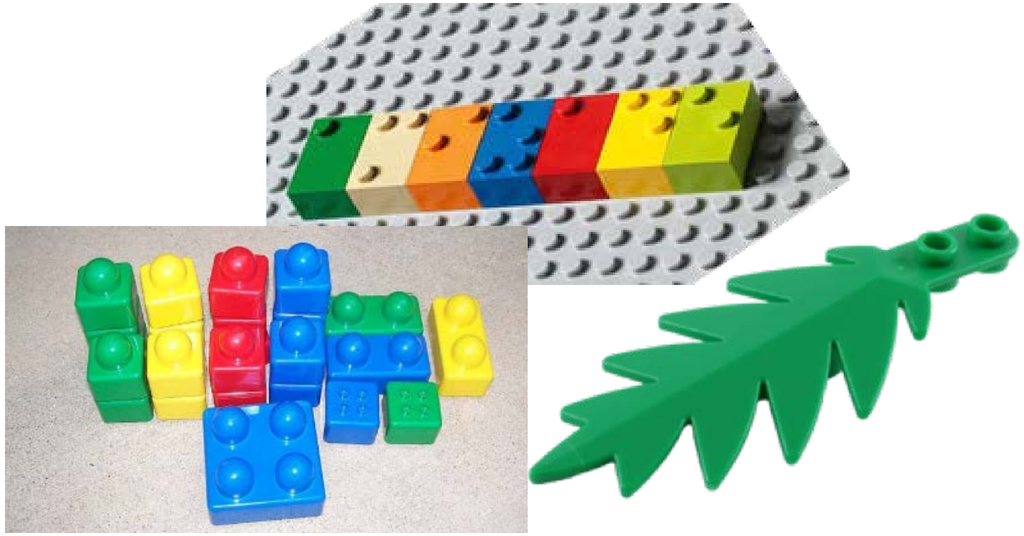
Stratix 10 280 Early Silicon (14nm)
ms-fp9
858K ALMs (92%)
5,760 DSPs (100%)
8,151 M20Ks (70%)
320 GOPS/W → 720 GOPS/W (production)



BrainWave Soft DPU
Floorplan on Stratix 10 280

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



“C-to-Gates” is *not* sufficient

- Concepts behind OpenCL/Vivado HLS are not new
  - 21+ tools called “C*” or “*C” targeting hardware
- Look back at dataflow architectures and CGRAs
- Integration of memory is critical
- Key question for each new tool/language:
  - Is this targeted at making hardware developers more productive?
  - Is this targeted at making software developers capable of using FPGAs?
  - If the answer is “both”, the answer is *neither*

Open-Source FPGA Development?

- Open-source projects build on libraries from a variety of places and times
  - Look how much Fortran code is still around for HPC!
- The Cloud can offer a *relatively*-stable HW platform
  - But FPGAs are light-years away from x86 code
- Dividing code into hardware microservices is the most scalable method
  - FFT, DGEMM, Smith-Waterman, etc
- Ripe area for research and development!
  - But do your homework. LOTS of existing work.

Conclusion

- The Cloud is larger and more powerful than the world's fastest supercomputers, and are still growing
- The FPGA Cloud enables huge increases in computing performance and efficiency, especially ML
- Network acceleration avoids virtualization overhead
- Still need work on software development for FPGAs

Microsoft