

FPGAs in Supercomputers: Opportunities or Folly?

FPGA'19 Banquet Panel

Deming Chen, ECE, University of Illinois at Urbana-Champaign

CHANGE
IS GOOD,
SON...

NOT WHEN
IT'S NOT
ENOUGH
FOR BUS
FARE.



Supercomputing applications

- quantum mechanics
- weather forecasting
- climate research
- oil and gas exploration
- molecular modeling
- physical simulations
 - early moments of the universe, airplane and spacecraft aerodynamics, the detonation of nuclear weapons, and nuclear fusion, etc.
- ...



The Summit Supercomputer

Sponsors	U.S. Department of Energy
Operators	IBM
Architecture	9,216 POWER9 22-core CPUs 27,648 Nvidia Tesla V100 GPUs ^[1]
Power	13 MW ^[2]
Storage	250 PB
Speed	200 petaflops (peak)
Purpose	Scientific research

Opportunity?

Engineering Letters, 16:3, EL_16_3_23

Maxwell – a 64 FPGA Supercomputer

University of Edinburgh, 2008

Microsoft's FPGA-powered supercomputers can translate Wikipedia faster than you can blink

The world doesn't have to long to wait for Microsoft's "A.I. supercomputers"; they're already here.

Paderborn University in Germany, 2018

Folly?

- **There is no FPGAs used in the top supercomputer systems yet**

Panelists

- **Hal Finkel**, *Lead, Compiler Technology and Programming Languages*, Argonne National Laboratory
- **Martin Herbordt**, *Professor*, ECE, Boston University
- **Wen-Mei Hwu**, *Sanders III AMD Endowed Chair Professor*, ECE, UIUC
- **Venkata Krishnan**, *Principal Engineer*, Intel
- **Viraj Paropkari**, *Senior Manager*, Global Data Center Marketing, Xilinx

Questions charged for the panelists

- Is there a need to bring FPGAs into supercomputers? Why or why not?
- Are there unique applications that are specifically suitable for FPGAs for supercomputing fields?
- What are the challenges and/or major issues facing FPGAs for supporting supercomputing?
- What and where are the opportunities? Who are the stakeholders?
- Name one thing that the FPGA industry should (or should not) do in the near term to facilitate FPGA's induction into supercomputers.



FPGAs in SUPERCOMPUTERS: OPPORTUNITY OR FOLLY

Venkata Krishnan
Intel Corporation

Feb 26, 2019



Disclaimer

Views, thoughts, and opinions expressed belong solely to the speaker, and not necessarily to the speaker's employer or any other group or individuals.

Any product information provided here is subject to change without prior notice.

Examples shown are for illustrative purposes and are to be treated as such.

DO FPGAs BELONG TO SUPERCOMPUTERS?

The answer is a **YES** but...

- Not as an alternative for Xeon/GPU/ASIC but rather complement them

FPGAs are relevant in Supercomputers

- For a well-defined set of “services” (e.g. specific/targeted applications)
- Such services need to quickly adapt to algorithmic changes/data & meet certain performance, cost, power requirements

One part of enabling this is making FPGAs 1st class citizens on the network

Note: There are times when FPGAs can be used in place of GPUs or help avoid ASICs. Also the term 1st class here broadly refers to FPGAs acting autonomously without host/OS involvement. It doesn't preclude FPGAs being deployed as a SmartNICs or as a NIC assist.

FPGAs can INDEED be FIRST CLASS CITIZENS on THE NETWORK!!

- EP300 PLD (1984) & XC2046 FPGA (1984)

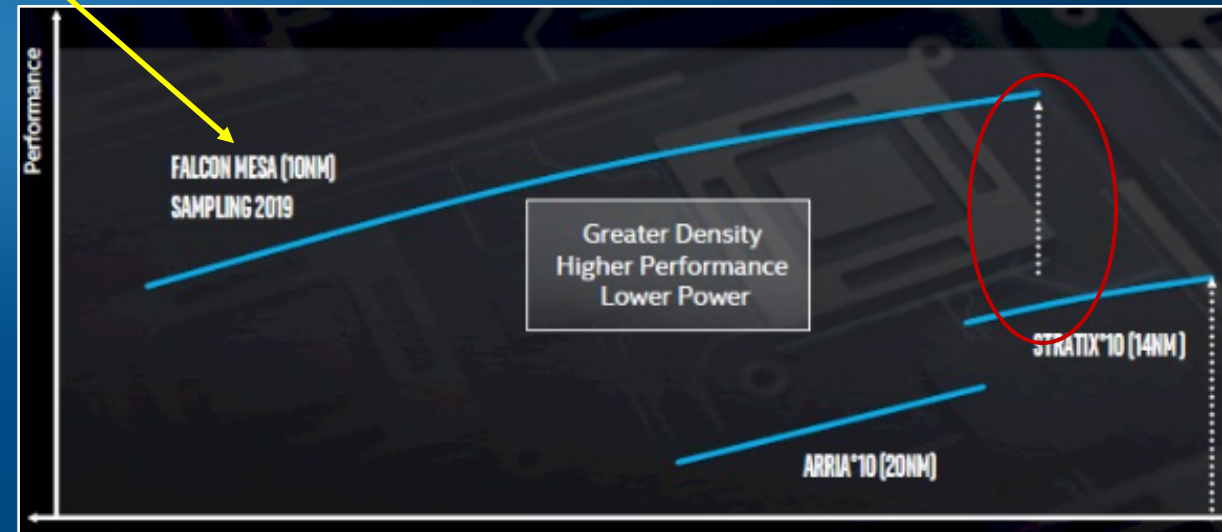
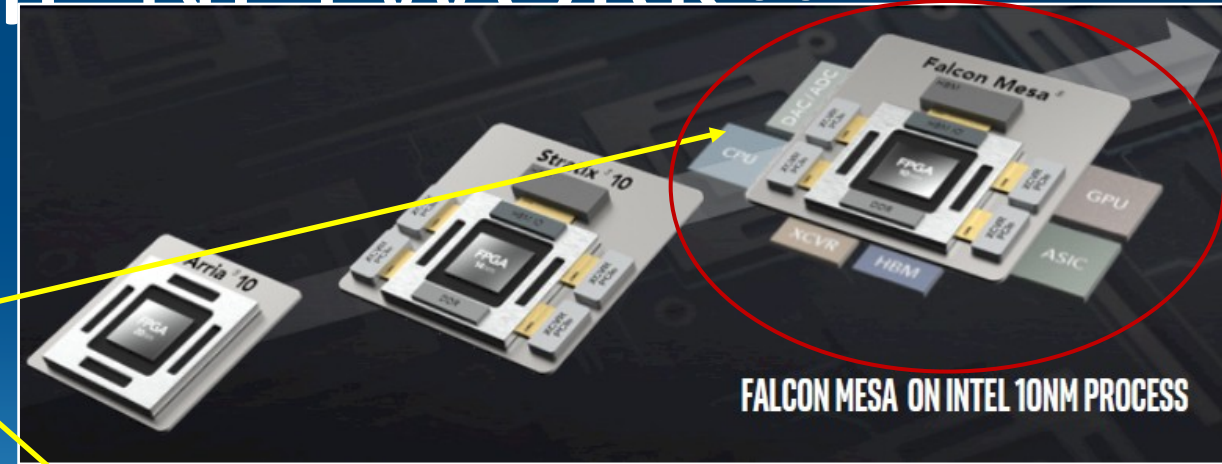
- 64 logic blocks with two 3-input LUT & 1 register

- Falcon Mesa (after Stratix10) in 2019

- Millions of logic blocks/registers

AND

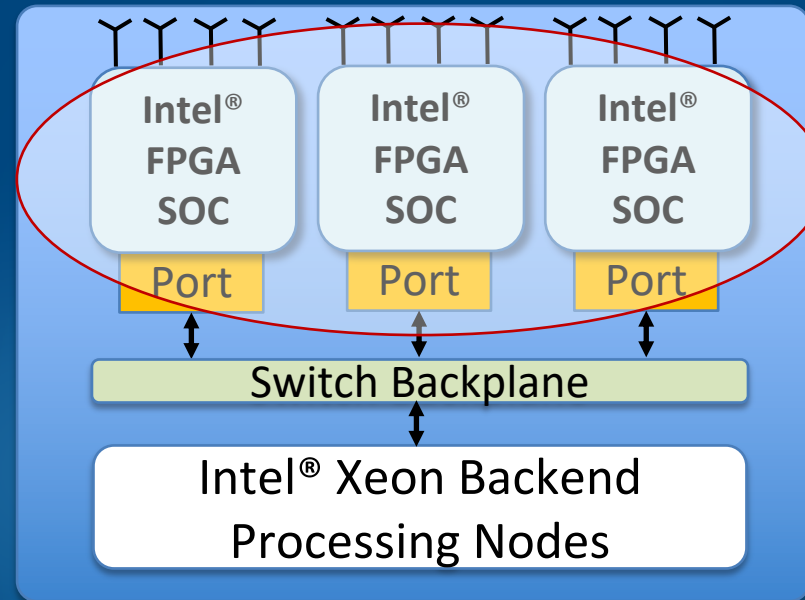
- Heterogeneous System-in-Package (SIP) Integration
- In package High Bandwidth Memory (HBM2/HBM3)
- Integrated high speed transceivers (112G PAM4)
- Quad-core ARM* SoC
- PCIe Gen4 IP, DSP, multipliers
- Support for DDR/DDR-T
- And much more...



Couple of examples (there are others) to show this usage in a large system follow...

EXAMPLE – Computing Near Sensors

FPGA SOC as Front End Processing Nodes

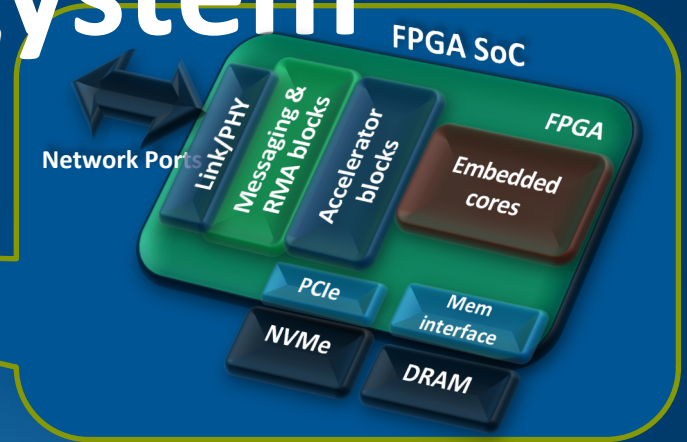
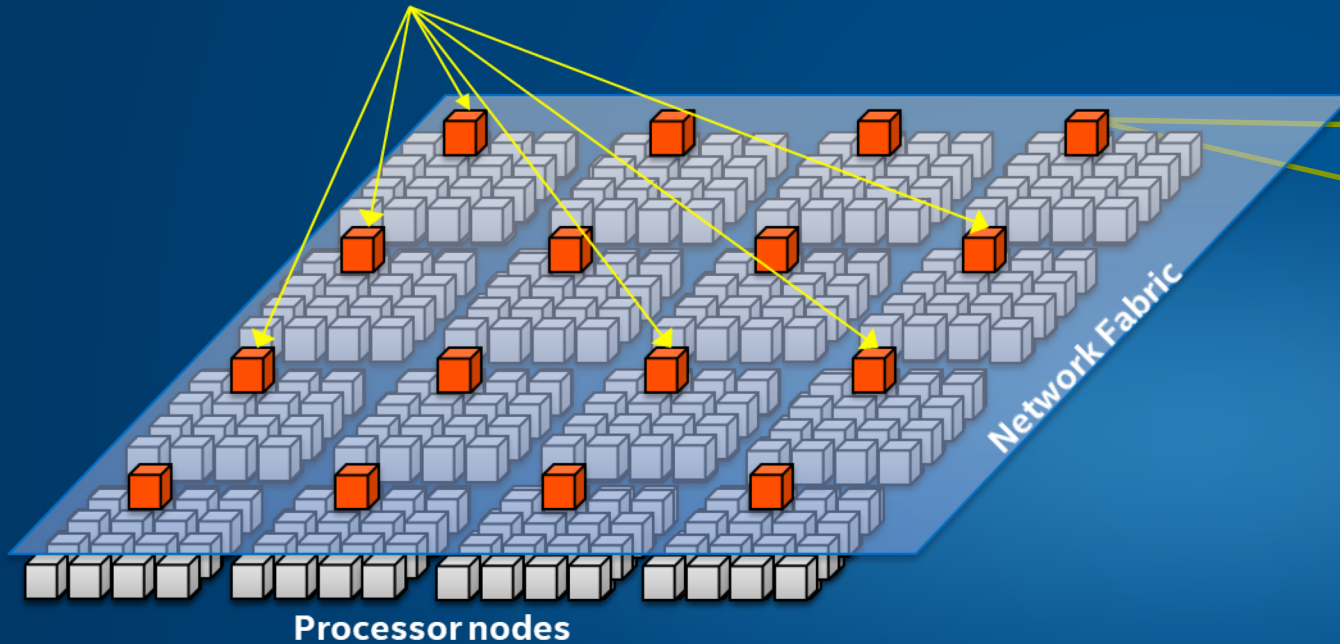


Frontend (Trigger) - Particle detectors, Radio Astronomy, Aerospace etc.

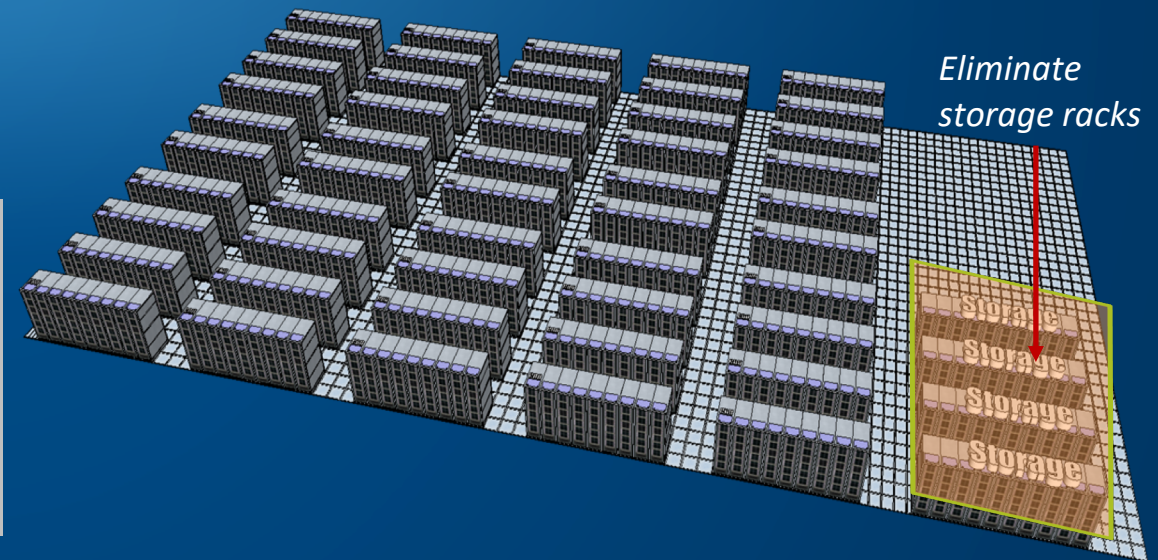
- “Filter” huge volume of data by performing compute at point of data acquisition
- Estimated reduction in backend nodes/fabric requirements **could be 10x-100x**
- Flexibility enables new/updates to algorithms

EXAMPLE – DISAGGREGATED STORAGE in a HPC system

Assumes a massively distributed high performance storage across entire system with FPGA SOC as storage nodes (for accelerating storage services)



As a standalone (autonomous) node



INTEGRATION OF STORAGE "NODE" WITH SWITCHES

- Rack Savings
- Infrastructure Power & cost reduction
- Cabling cost
- Performance (acceleration) with opportunity to reconfigure based on changes to storage services



GrEAT. BUT HOW DO
WE GET THERE?

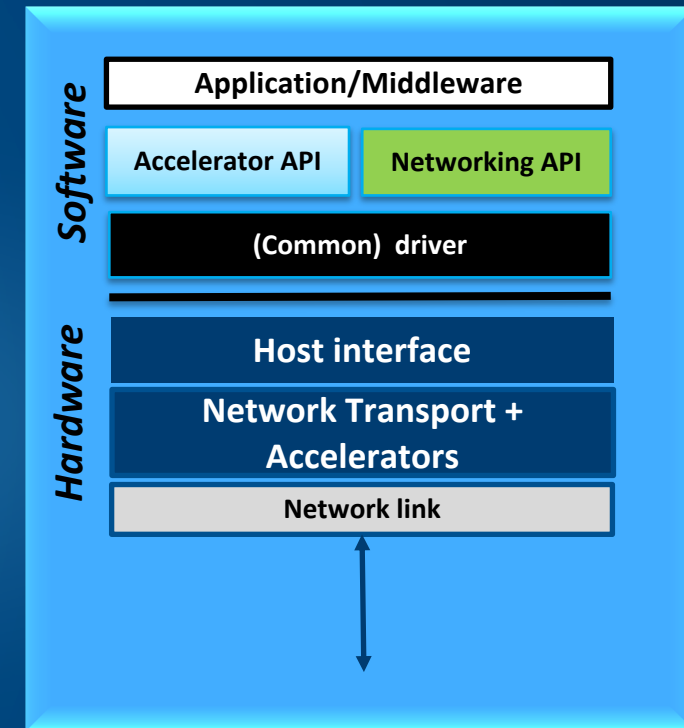
NEED A FOUNDATIONAL INFRASTRUCTURE

1. Basic infrastructure that integrates networking & accelerator support

- Modular architecture that allows ease of customization
- Familiar programming environment developed around open standards (not proprietary)

2. Customizing infrastructure based on targeted application or services

- Identify core libraries & provide necessary IP (accelerator) blocks
- Integrate them on the above infrastructure for a complete solution



High-level view of a conceptual software/hardware stack

STEP 1 IS A NECESSARY STEP

STEP 2 DRIVEN BY CUSTOMER OR APPLICATION REQUIREMENTS

Providing a “FAMILIAR” PROGRAMMING ENVIRONMENT

Extensions to OFI*

- OFI is a low-level networking API with extensive middleware support - applications can continue to use standard APIs
- Extending this network API with acceleration capabilities is easier than doing it the other way around (i.e. taking an acceleration API such as OpenCL and extending it with networking primitives)



OFI

Open standard. Can be extended for new features (e.g. acceleration)

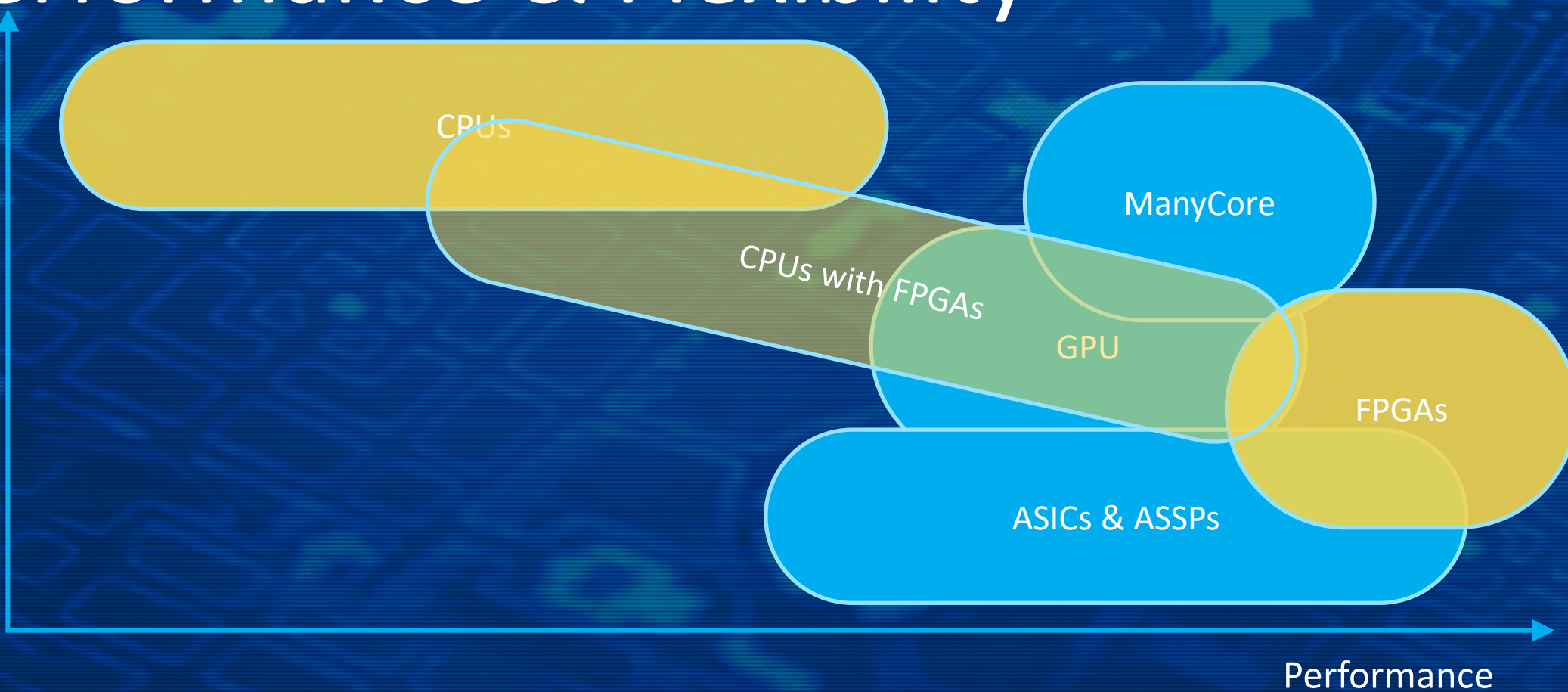
Defines a communication API and agnostic to networking protocols & HW implementation

Architected to support RDMA functionality



The Challenge: Balancing Performance & Flexibility

flexibility



OFI LIBFABRIC COMMUNITY

Intel® MPI
Library

MPICH

Open MPI
SHMEM

Sandia
SHMEM

GASNet

Charm++

Clang
UPC

Chapel

Others: rsockets, PMDK, Spark,
ZeroMQ, TensorFlow, MxNET, NetIO,
Intel MLSL, ...

libfabric Enabled Middleware

libfabric

Advanced application oriented semantics

Tag Matching

Scalable
memory
registration

Triggered
Operations

Remote
Completion
Semantics

Multi-
Receive
buffers

Shared
Address
Vectors

Unexpected
Message
Buffering

Streaming Endpoints

Reliable Datagram Endpoints

Sockets
TCP, UDP

Verbs

Cisco
usNIC

Cray
GNI

Intel
OPA, PSM

Shared
Memory

Mellanox
UCX

IBM Blue
Gene

Network
Direct

HPE
Gen-Z

RxM, RxD, Multi-
Rail, Hooks, ...

* Other names and brands may be claimed as the property of others

**OFI insulates applications
from fabric diversity**

HPC with FPGAs**

Martin Herbordt

Computer Architecture and Automated Design Laboratory
Department of Electrical and Computer Engineering
Boston University
<http://www.bu.edu/caadlab>

* This work supported, in part, by Red Hat, Microsoft, the U.S. NIH and NSF, and by donations from Xilinx, Intel-Altera, and Gidel

+ Thanks to Ahmed Sanaullah, Ethan Yang , Qingqing Xiong, Jiayi Sheng, Robert Munafo, Josh Stern, Tony Geng, Tianqi Wang

Question 1:

What's the status of FPGA/HPC right now?

Method:

- Look

FPGA/HPC is here!



MAXELER Technologies
MAXIMUM PERFORMANCE
www.maxeler.com

MPC-X
The MPC-X provides Maximum compute density holding eight DFEs in a single slot.

Open SPL
Open SPL is an open standard for a novel Spatial Programming Language. It is based on a novel perspective

Ultra Low Latency Dataflow Renderer
Dataflow Engines (DFEs) deliver the lowest latency solution for VR rendering.

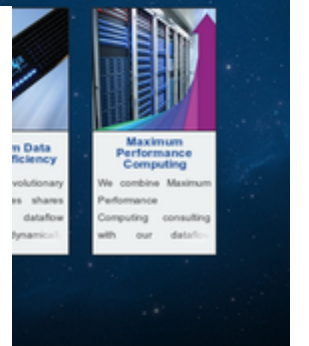
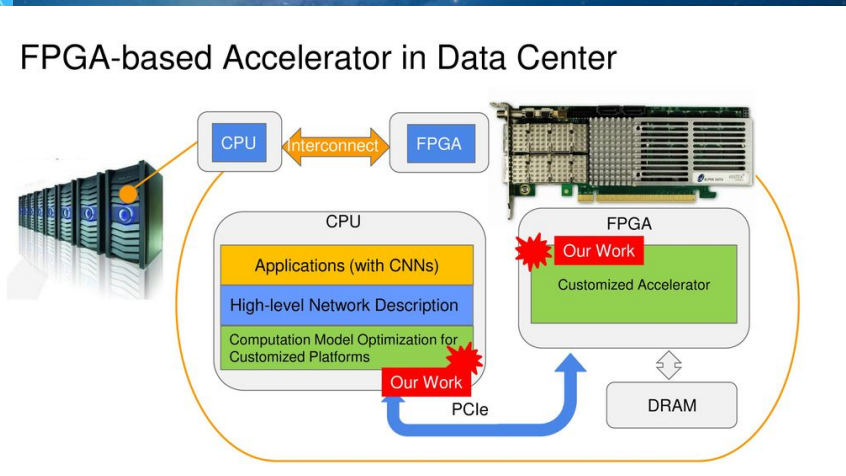
Low Latency
Maxeler provides a fully programmable network platform for low latency trading, including:

Maximum Performance Trading
Maxeler's MPT combines dataflow computing technology with optimized industrial

DFEs for Computational Fluid Dynamics
In August 2015 Rolls-Royce and Maxeler started a collaborative project supported by the UK government.

Open SPL Course
The aim of this course is to introduce you to the basics of Computing in Space.

Galva
Highly capable PCIe dataflow compute card, available for Universities to enable affordable



HPC is not here!

Select All on Page Sort By: Newest First

~~**MetaMorph: A Library Framework for Interoperable Kernels on Multi- and Many-Core Clusters**~~
Ahmed E. Helal; Virginia Tech; Paul Sathre; Wu-chun Feng
SC '16: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis
Year: 2016
Page s: 119 - 129
Cited by: Papers (1)
IEEE Conferences
▶ Abstract (html) PDF (480 Kb) CC

~~**Evaluating and Optimizing OpenCL Kernels for High Performance Computing with FPGAs**~~
Hamid Reza Zohouri; Naoya Maruyama; Aaron Smith; Motohiko Matsuda; Satoshi Matsuoka
SC '16: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis
Year: 2016
Page s: 409 - 420
Cited by: Papers (24)
IEEE Conferences
▶ Abstract (html) PDF (416 Kb) CC

~~**GRAPE-8 -- An accelerator for gravitational N-body simulation with 20.5Gflops/W performance**~~
Junichiro Makino; Hiroshi Daisaka
SC '12: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis
Year: 2012
Page s: 1 - 10
Cited by: Papers (2)
IEEE Conferences
▶ Abstract (html) PDF (414 Kb) CC

~~**Hardware/software co-design for energy-efficient seismic modeling**~~
Jens Klöpper; David Donofrio; John Shalf; Marghoob Mohiyuddin; Samuel Williams; Leonid Oliker; Franz-Josef Pfreundt
SC '11: Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis
Year: 2011
Page s: 1 - 12
Cited by: Papers (10) | Patents (2)
IEEE Conferences
▶ Abstract (html) PDF (847 Kb) CC

~~**SCAM: a scalable CAM-based algorithm for multiple pattern inspection**~~
Fabrizio Petrini; Prat Agarwal; Davide Pasetto
Proceedings of the Conference on High Performance Computing, Networking, Storage and Analysis
Year: 2009
Page s: 1 - 11
Cited by: Papers (3)
IEEE Conferences
▶ Abstract (html) PDF (1508 Kb) CC

~~**A preliminary investigation of a neocortex model implementation on the Cray XD1**~~

~~**Architectures and APIs: Assessing Requirements for Delivering HPC Performance to Applications**~~
Keith D. Underwood; K. Scott Hemmert; Craig Ulmer
SC '06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing
Year: 2006
Page s: 49 - 49
Cited by: Papers (7)
IEEE Conferences
▶ Abstract (html) PDF (284 Kb) CC

~~**Preliminary Investigation of Advanced Electrostatics in Molecular Dynamics on Reconfigurable Computers**~~
Ronald Scrofano; Viktor K. Prasanna
SC '06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing
Year: 2006
Page s: 45 - 45
Cited by: Papers (3)
IEEE Conferences
▶ Abstract (html) PDF (244 Kb) CC

~~**A Near-Optimal Real-time Hardware Scheduler for Large Cardinality Crossbar Switches**~~
Raymond R. Hoare; Zhu Ding; Alex K. Jones
SC '06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing
Year: 2006
Page s: 8 - 8
Cited by: Papers (3)
IEEE Conferences
▶ Abstract (html) PDF (338 Kb) CC

~~**Is High-Performance, Reconfigurable Computing the Next Supercomputing Paradigm?**~~
Tarek El-Ghazawi
SC '06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing
Year: 2006
Page s: xv - xv
Cited by: Papers (4)
IEEE Conferences
▶ Abstract (html) PDF (444 Kb) CC

~~**A Configurable Network Protocol for Cluster Based Communications using Modular Hardware Primitives on an Intelligent NIC**~~
R.G. Jaganathan; K.D. Underwood; R. Sass
SC '03: Proceedings of the 2003 ACM/IEEE Conference on Supercomputing
Year: 2003
Page s: 22 - 22
Cited by: Papers (1)
IEEE Conferences
▶ Abstract (html) PDF (170 Kb) CC

~~**Parallel Dedicated Hardware Devices for Heterogeneous Computations**~~
A. Marongiu; P. Palazzari; V. Rosato
SC '01: Proceedings of the 2001 ACM/IEEE Conference on Supercomputing
Year: 2001
Page s: 29 - 29
Cited by: Papers (1) | Patents (1)
IEEE Conferences
▶ Abstract (html) PDF (123 Kb) CC

Search for "FPGA" in SC Proceedings from 2001-2016 – 13 hits ☹️

~~**A preliminary investigation of a neocortex model implementation on the Cray XD1**~~
IEEE Conferences
▶ Abstract (html) PDF (298 Kb) CC

~~**A preliminary investigation of a neocortex model implementation on the Cray XD1**~~
Cited by: Papers (1) | Patents (12)
IEEE Conferences
▶ Abstract (html) PDF (368 Kb) CC

Question 1:

What's the status of FPGA/HPC right now?

Answer 1: Long way to go

... especially if we mean having teams in place at non-FPGA-specialized facilities.

Question 2:

Why FPGAs for HPC?

Answer 2:

- **Beat GPUs?**
- **Transceivers – cheap, flexible, high quality interconnects**
- **Co-location of compute and communication logic**
- **Flexible on-chip interconnects**

“We are good at low latency. We will stay good at low latency.” - keynote, FGPA 2019

“Data movement is everything” - heard at FPGA 2019

Question 3:

What will FPGAs in HPC look like?

Answer 3: FPGAs everywhere!

FPGA Enhanced Clouds & Clusters

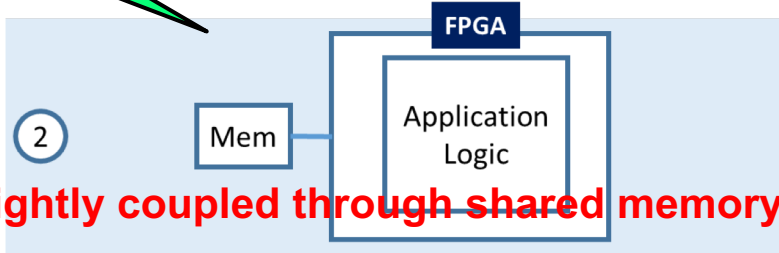
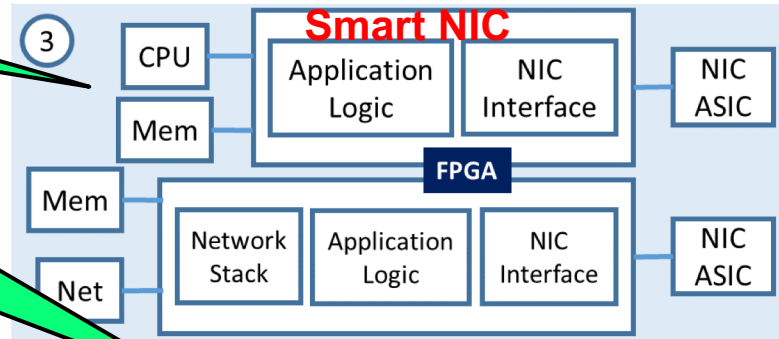
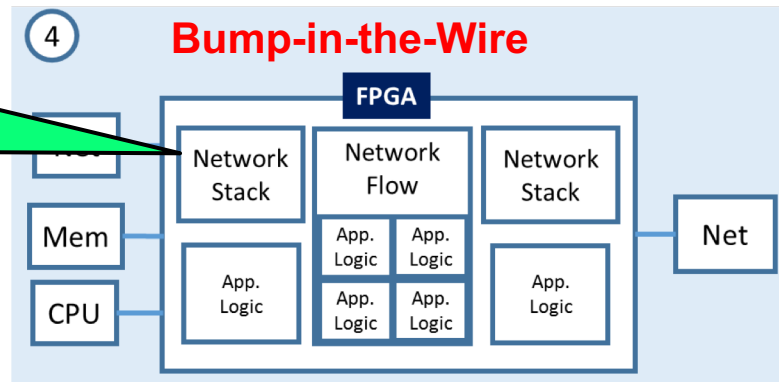
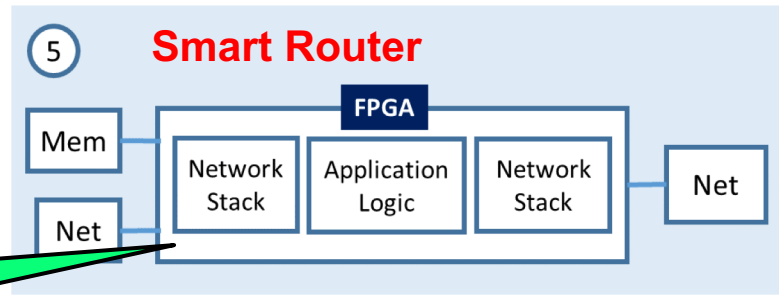
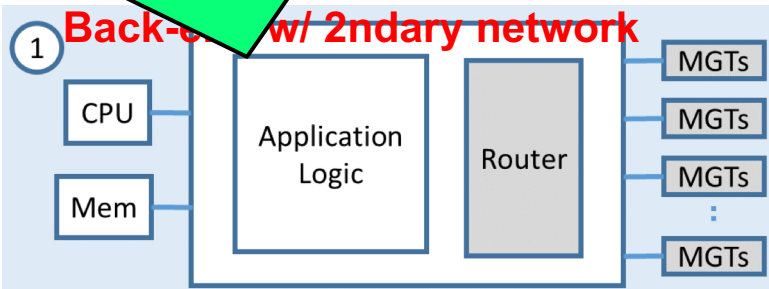
Killer HPC App: MPI Reductions, Reductions w/ user-defined ops/types

Killer HPC App: ML Training in the cloud – enables fine-grained pipeline to reduce storage – competitive with GPUs

Killer HPC App: Lossy compression – 100x speed-up over CPU

Killer HPC App: MPI Offload – SW into HW for 100x reduction in overhead

Killer HPC App: Anton w/ FPGA, long timescale Molecular Dynamics



Tightly coupled through shared memory

Question 4:

Can ordinary computer professionals make FPGA application cost-effective(ly)?

Method:

- Try OpenCL

Results

CPU : 14 core 2.4GHz Intel® Xeon® E7-4908v3
GPU: NVIDIA P100 (16nm), 3584 cores, HBM2
FPGA: Intel® Arria® 10 (20nm), 427K ALMs, 1.5K DSP blocks

	CPU (14 core)	Previous FPGA OpenCL	Our Work	GPU	Verilog
Average Speedup of Our Work	1.2x	2.5x	1.0x	0.3x	0.9x
Highest Speedup Achieved by Our Work	5.9x (NW) [1]	155x (Range Limited) [2]	1.0x	2.6x (SPMV) [3]	2.1x (SPMV) [4]

Why is 0.3x versus GPU good?

- GPU/CPU reference codes are (mostly) highly tuned by vendor, OpenCL is ours
- Apps are mostly very good on GPUs (e.g., versus CPUs)

We estimate a 4x increase in performance of our OpenCL designs using Intel® Stratix® 10

[1] S. Che *et al.* "Rodinia: A Benchmark Suite for Heterogeneous Computing," in *IISWC*, 2009.

[2] C. Yang *et al.* "OpenCL for HPC with FPGAs: Case study in molecular electrostatics," in *HPEC*, 2017.

[3] NVIDIA cuSparse

[4] L. Zhuo *et al.* "Sparse Matrix-Vector Multiplication on FPGAs," in *FPGA*, 2005.

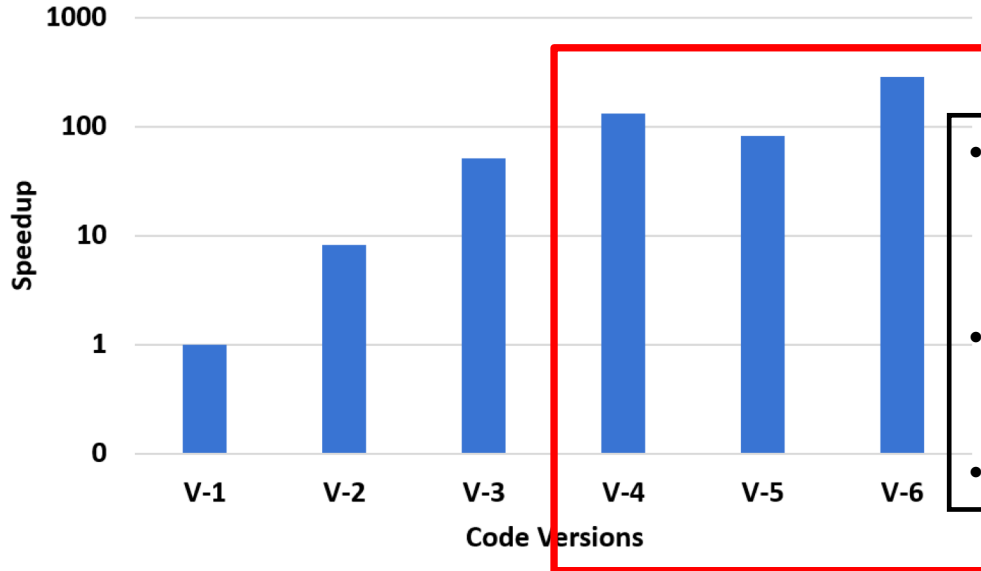
Related Work:

A. Sanaullah *et al.* "Unlocking Performance-Programmability by Penetrating the Intel FPGA OpenCL Toolflow," in *HPEC*, 2018

A. Sanaullah *et al.* "SimBSP: Enabling RTL Simulation for Intel FPGA OpenCL Kernels," in *H2RC*, 2018

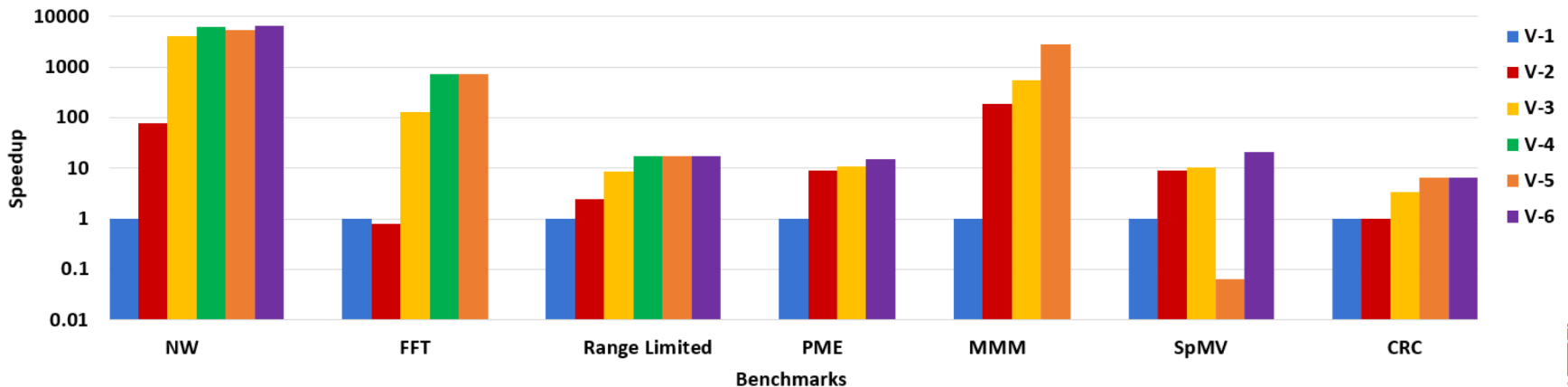
Characterization of Optimizations

Average Incremental Impact of Individual Optimizations



- All optimizations are application “unaware” and known to the compiler/autotuning communities
- However – many are not part of OpenCL “best practices,” esp. V-4 – V-6
- These account for 5x performance

Speedup of Code Versions Relative to Version 1 (Baseline)



Question 4:

Can ordinary computer professionals make FPGA application cost-effective(ly)?

Answer 1: This SHOULD work!

- **Currently requires skill/knowledge**
- **Standard practices not there yet, but should be**
- **Should be integrated into OpenCL compilers**

Summary

FPGA/HPC is here and spreading

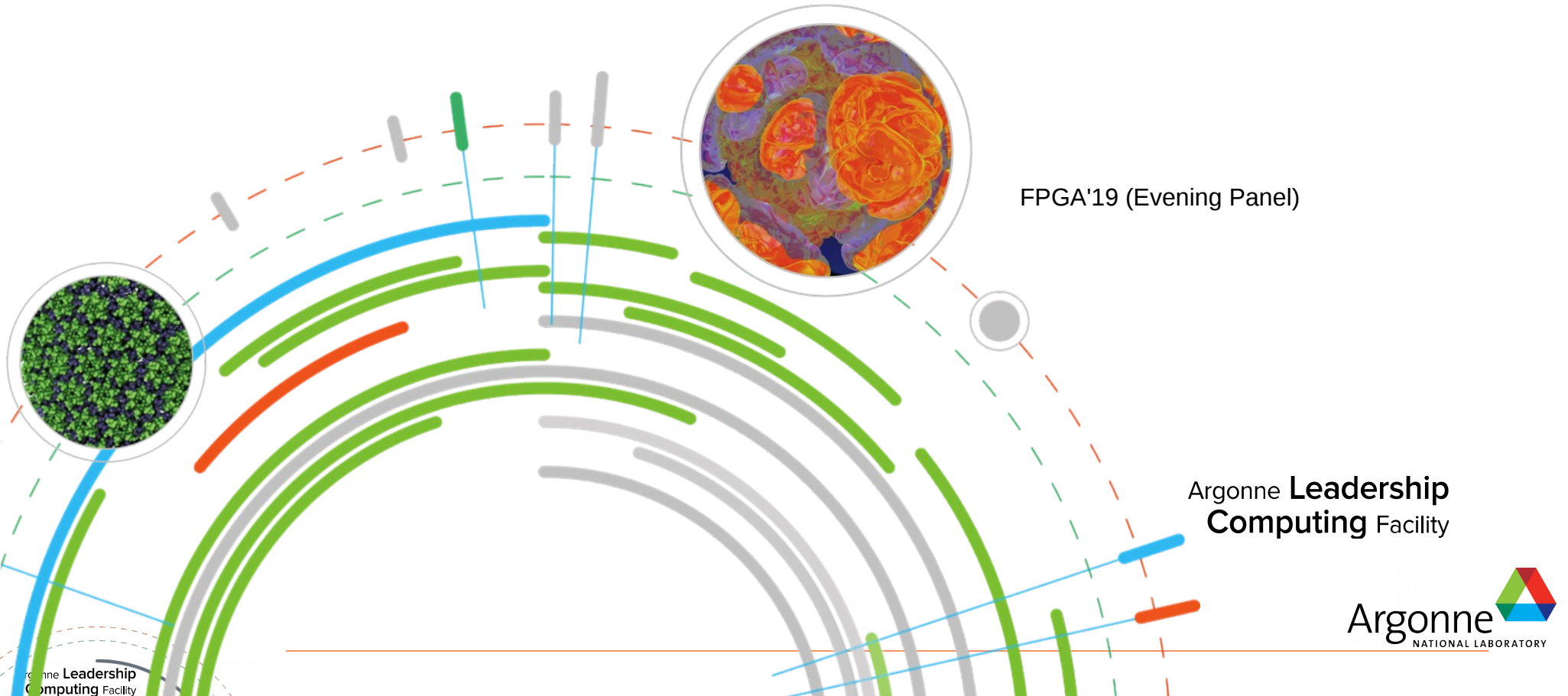
No reason it shouldn't extend further –

- System/Provider functions written by FPGA engineers w/ massive potential impact
- Applications written using HLS

Might take a while before many traditional HPC production apps run primarily on FPGAs

FPGAs for Supercomputing: Already Here and Yet So Far Away?

Hal Finkel (hfinkel@anl.gov)



FPGA'19 (Evening Panel)

Argonne **Leadership**
Computing Facility

Is there a need to bring FPGAs into supercomputers?

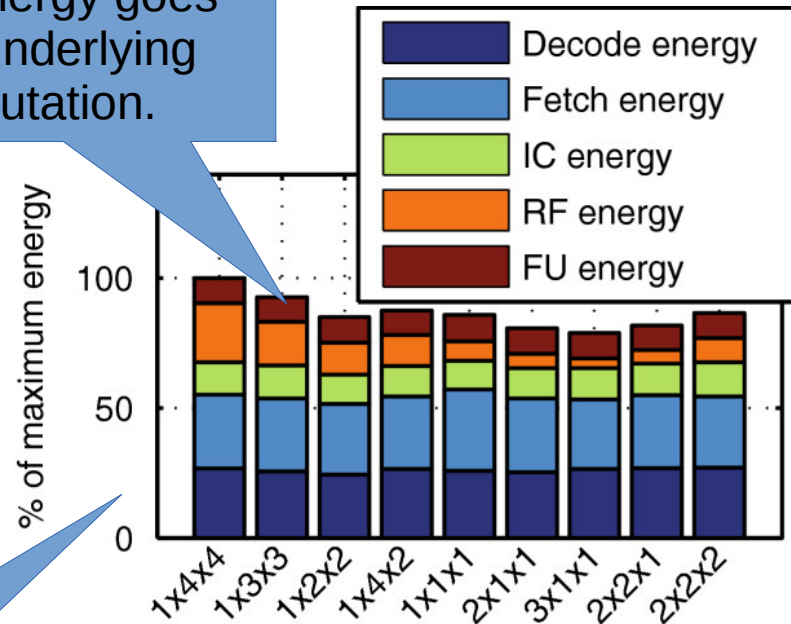
Operation	Energy (pJ)
64-bit integer operation	1
64-bit floating-point operation	20
256 bit on-die SRAM access	50
256 bit bus transfer (short)	26
256 bit bus transfer (1/2 die)	256
Off-die link (efficient)	500
256 bit bus transfer (across die)	1,000
DRAM read/write (512 bits)	16,000
HDD read/write	$O(10^6)$

Do FPGA's perform less data movement per computation?

Courtesy Greg Asfalk (HPE) and Bill Dally (NVIDIA)

Is there a need to bring FPGAs into supercomputers? (cont.)

Only a small portion of the energy goes to the underlying computation.



More centralized register files means more data movement which takes more power.

Fetch and decode take most of the energy!

(Model with (# register files) x (read ports) x (write ports))

<http://link.springer.com/article/10.1186/1687-3963-2013-9>

See also: <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-2008-130.pdf>

Are there unique applications that are specifically suitable for FPGAs for supercomputing fields?

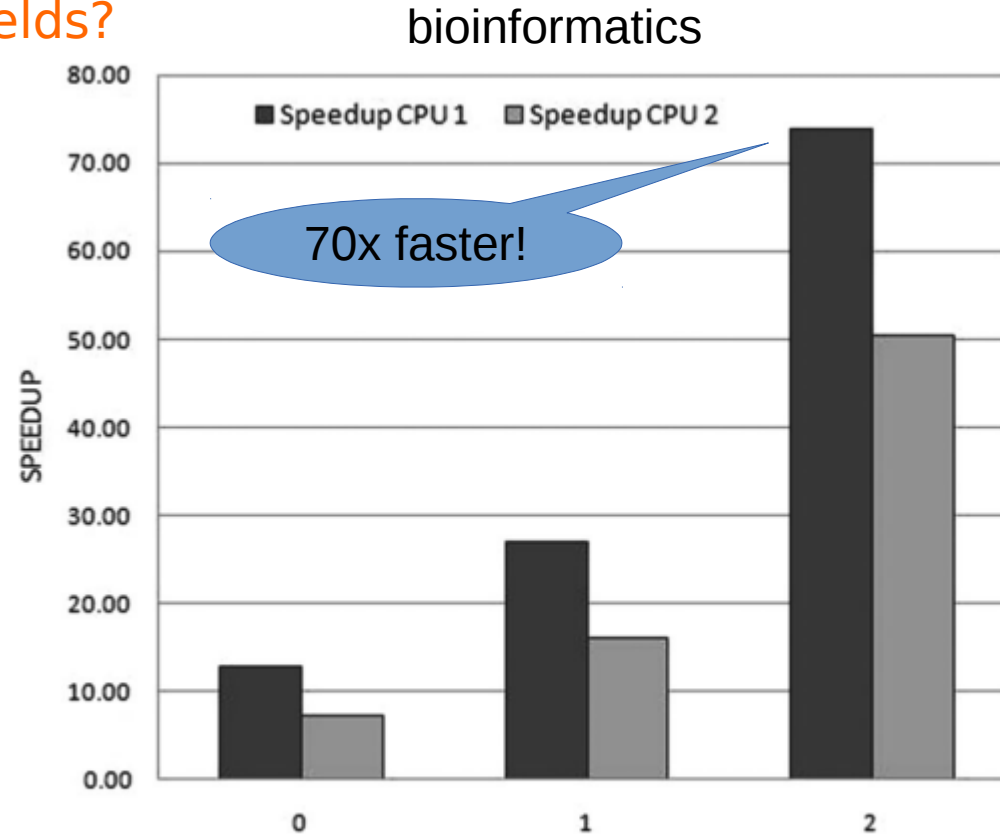


Fig. 9. Speed up of FFAST compared to BOWTIE for exact matches, one and two mismatches.

Are there unique applications... (cont.)

machine learning and neural networks

FPGA is faster than both the CPU and GPU, 10x more power efficient, and a much higher percentage of peak!

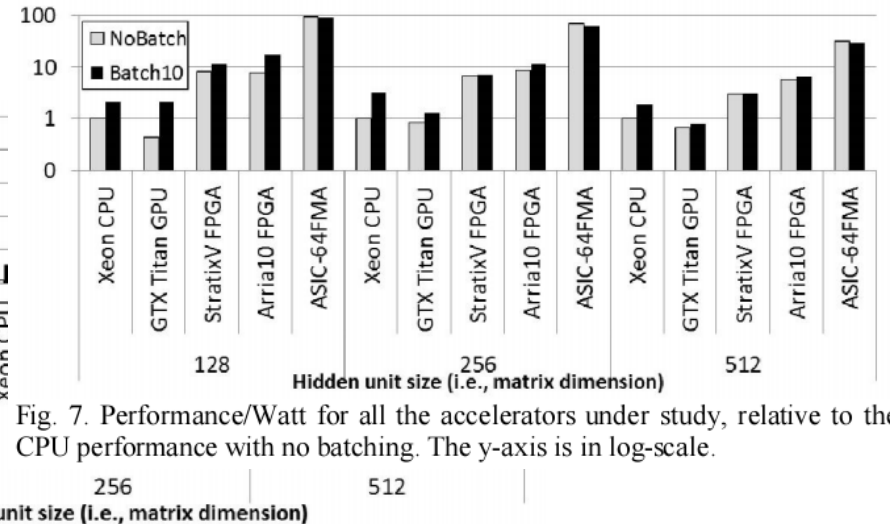
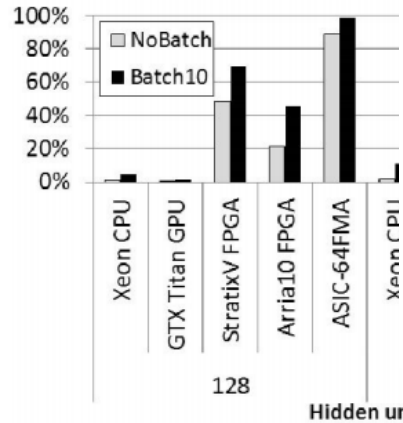
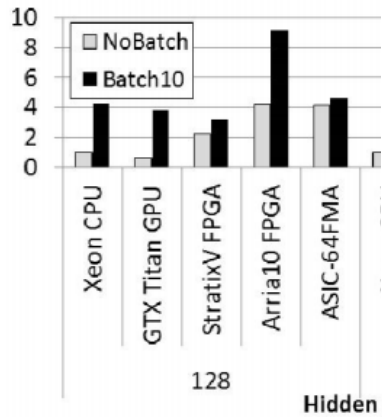


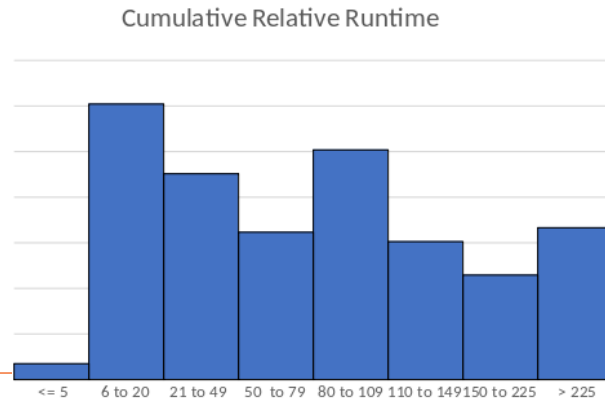
Fig. 5. Performance for all the accelerators under study, relative to CPU performance with no batching.

Fig. 6. Achieved performance relative to peak performance. E.g., 10% means the system is underutilized, where the achieved GFLOP/s is only at 10% of the available peak GFLOP/s. On the other hand, 100% means full utilization.

Fig. 7. Performance/Watt for all the accelerators under study, relative to the CPU performance with no batching. The y-axis is in log-scale.

What are the challenges and/or major issues facing FPGAs for supporting supercomputing?

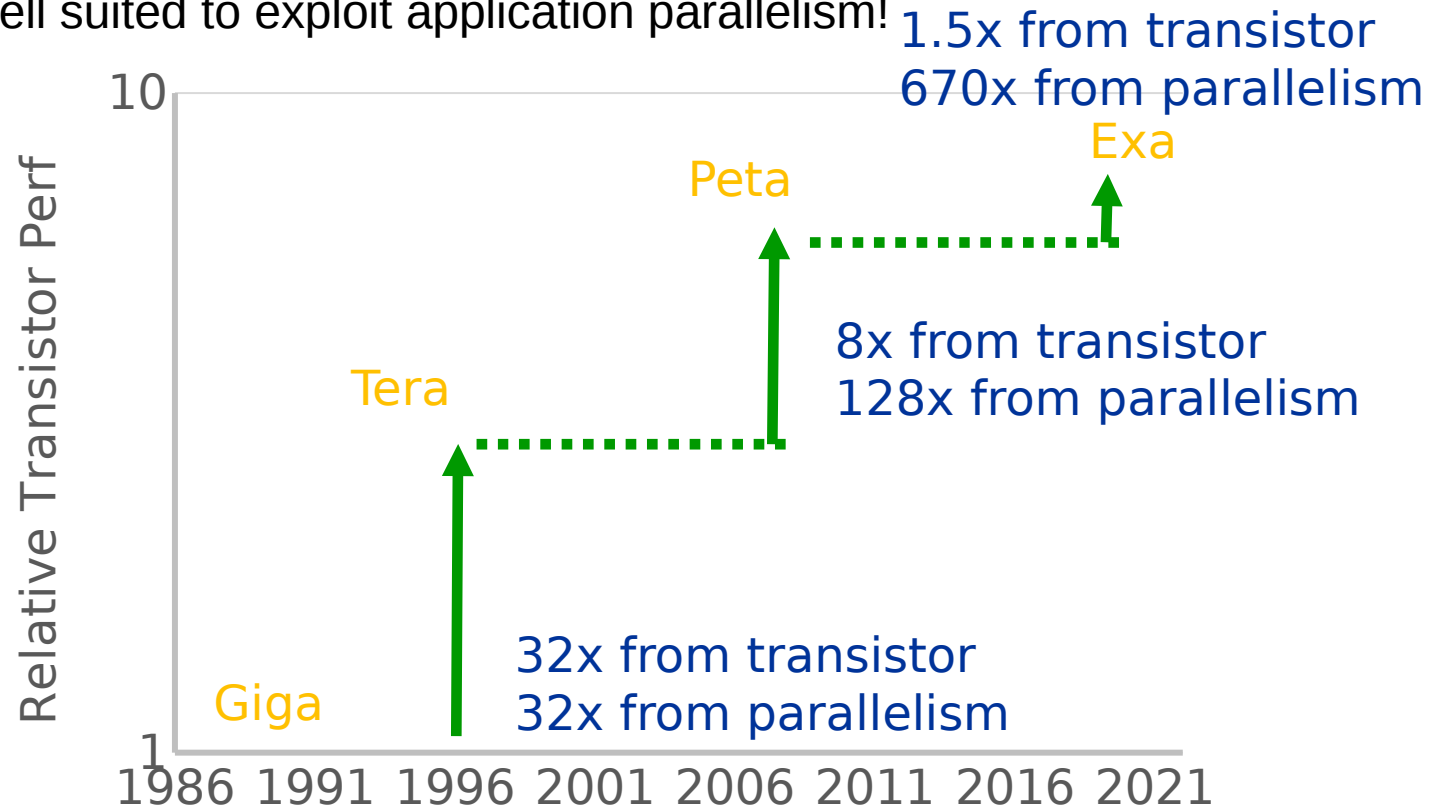
- Compile Time: Place & Route for a high-end FPGA can take hours or days, and that might be *per kernel* for a larger application. Large applications might have hundreds of kernels. Thus, FPGAs will have difficulty functioning as *general purpose* accelerators.
- Double-precision floating-point support. Most HPC applications currently require it.
- Code Size: It's unclear how many of the kernels used by real applications will fit on even a large FPGA.



Lines of code in kernels in DOE proxy apps: ASPA, SNAP, SNBone, SW4lite, Nekbone, SWFFT, miniFE, Lulesh, miniAero, MiniGMG, CoMD, CoSP2, HPCCG, miniTri, PENNANT, Pathfinder, RSBench, SimpleMOC, XSBench – Data collected by Brian Homerding (ALCF).

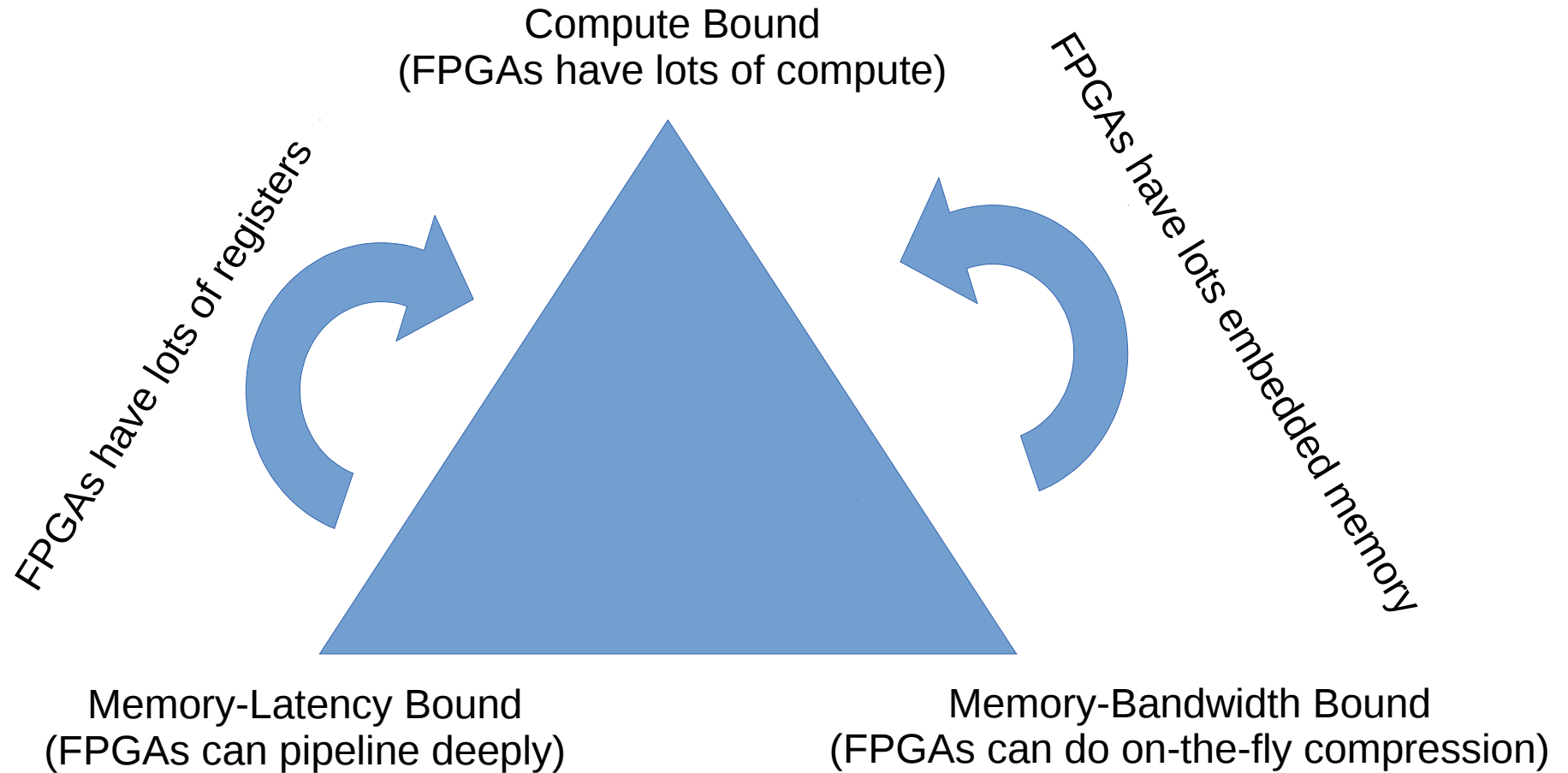
What and where are the opportunities? Who are the stakeholders?

FPGAs are well suited to exploit application parallelism!



System performance from parallelism

What and where are the opportunities? Who are the stakeholders? (cont.)



Name one thing that the FPGA industry should (or should not) do in the near term to facilitate FPGA's induction into supercomputers.

- Build integrated CGRAs: Addressing both compile-time and space constraints, while maintaining the advantages of FPGAs, will require increasing the amount of variety of hardened logic.



A Brief History of CUDA GPUs in Supercomputers

- CUDA started in 2007
- First major adoption - Tienhe-1 in 2009
 - David Kirk and Wen-mei Hwu tutorial in 2008
- First major US adoption - Blue Waters and Titan in 2013
 - 3072 → 4224 Kepler GPUs in Blue Waters
 - NSF PAID program helps science teams to use GPUs
 - Numerous courses and tutorials along the way
- As of 2018, about 20% of the Blue Waters applications use GPUs in a significant way
- Most top supercomputers use GPUs in 2018

What is limiting the GPU use today?

- Data transfer cost and GPU occupancy
 - Large granularity of the off-loaded compute is needed to get significant benefit
 - Level of data reuse must be high
- Insufficient DRAM/HBM capacity
 - Limited types of computation
- Programming effort
 - Application code often needs to be refactored
 - FORTRAN support is still weak

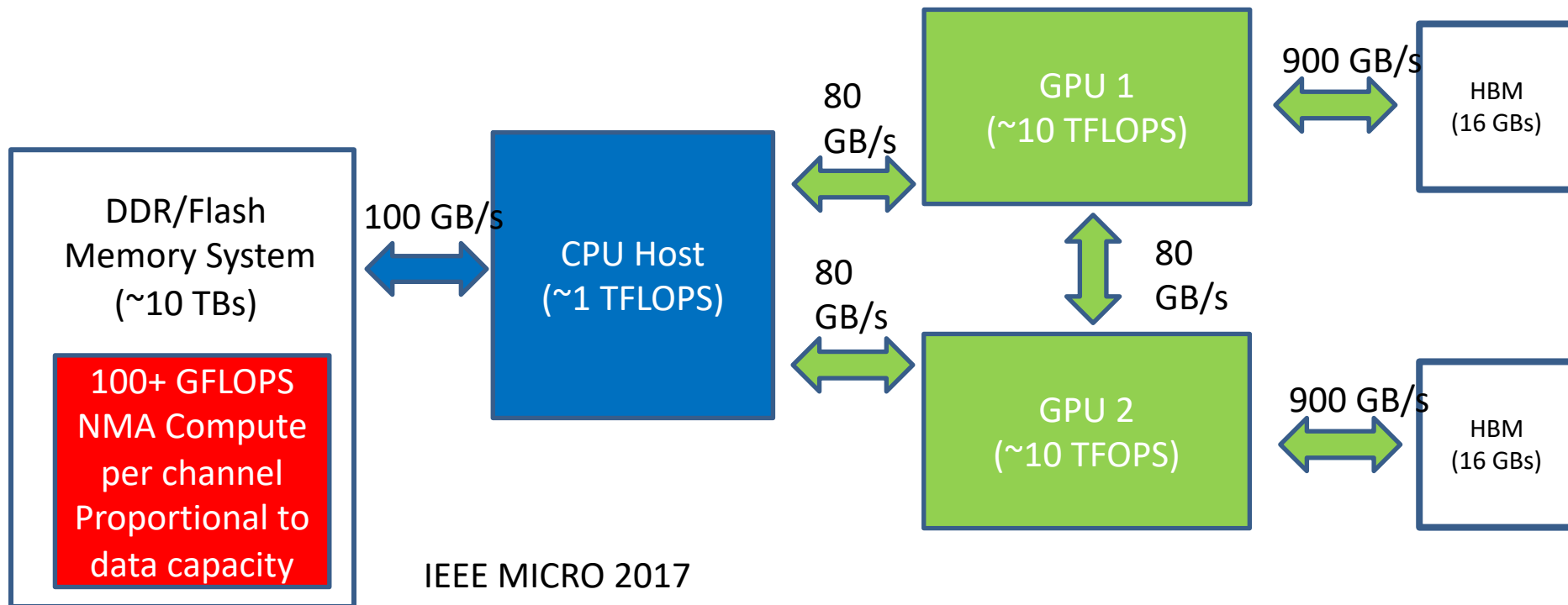
FPGA as a Computing Device

- Strength
 - Logic and integer arithmetic
 - Flexible data reuse patterns
 - Energy efficiency
- Weakness
 - Clock frequency
 - Floating-point compute throughput
 - DRAM bandwidth
 - Program loading (reconfiguration)
 - DRAM latency tolerance

Q: One suggestion for the FPGA industry

- FPGA in intelligent memory/storage controllers
 - Low-data-reuse compute
 - Data access pattern adjustment between memory bank and channels (e.g. strided accesses)
 - On-the-flight data compression/encryption
- Seamless collaboration with CPUs and GPUs crucial

Erudite: placing NMA compute inside storage-class memory controllers



FPGAs in Supercomputing -Challenges and Opportunities

- **Viraj Paropkari , Senior Manager, Global DC Marketing**

25th February 2019

1. Is there a need to bring FPGAs into supercomputers? Why or why not?

- > Absolutely **YES..YES..YES**
- > **Next gen supercomputer will do many complex applications (not killer workload with AI integrated)** When one looks at Exa-scale level and beyond systems ; those platform architecture needs to change dramatically to accommodate power efficiency and flexibility in a HPC center. We believe FPGAs have potential to get there. We don't expect overnight developers will switch from CPUs/GPUs to FPGAs but there is enough traction in HPC community and we are working closely to define next gen requirements – e.g Versal with it's capabilities such as AI engines, programmable HW , SW maturity also Alveo board level products with standardized stack. It is completely possible to co-exist CPU+GPU+FPGA system.
- > **FPGAs have great practical Performance :-** GPUs gives about 70% (in many cases 50%) of peak performance. This is waste about half of the HW. CPUs have high peak but high overhead as well. FPGAs offer higher of peak performance with low overhead. This is inherent architecture of logic units with interconnected memory system where data movement is less which consumes more power in CPUs and GPUs.
- > **Emphasis great Power efficiency using FPGAs:-** 10x power efficient over GPU and 50x power efficient over CPU
- > **Use of FPGAs in HPC in not new but was limited:-** If you look at various work in HPC done a few years ago, you will realize many promising performance results using HDL - hand tuned programmed codes. This work remained to very limited set of developers as it requires specialized HW knowledge to tune the algorithm to HW level in HDL. Now with the maturity of SDAccel – HLS/OpenCL/C++ support from vendors such as Xilinx ; the performance is encouraging to hand tunes HDL codes.

Traditional HPC applications Requirements & FPGAs advantages

> Compute Bound

- >> FPGAs have many compute DSP blocks, lot of registers, flexible precision e.g Alveo U250
- >> Future FPGAs such as Versal getting more powerful for Flops rate– adaptable HW engines, Intelligent engines, SW programmable engines

> Memory Bandwidth Bound

- >> High bandwidth HBM FPGA devices becoming recently available e.g Alveo U280
- >> Lot of embedded memory and bandwidth compared to GPUs
 - 2x on chip memory and 4x on chip memory BW compared to Volta
- >> Other techniques such as compression/ decompression on-the-fly

> Memory Latency Bound

- >> FPGAs can pipeline deeply
- >> Algorithms such as FFTs get benefitted by deeper execution pipelines
- >> Versal will push latency advantage further with high throughput
- >> Massive array of intelligent cores with local memory tightly coupled to adaptable HW



FAST

Faster than CPUs & GPUs
Latency advantage over GPUs



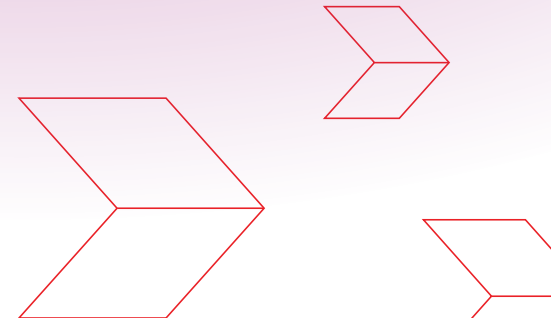
ADAPTABLE

Optimized for any workload
Adapt to changing algorithms



ACCESSIBLE

Deploy in the cloud or on-premises
Applications available now



Alveo Product Table

	Product Name	Alveo U200	Alveo U250	Alveo U280
Dimensions	Width	Dual Slot	Dual Slot	Dual Slot
	Form Factor, Passive Form Factor, Active	Full Height, ¾ Length Full Height, Full Length	Full Height, ¾ Length Full Height, Full Length	Full Height, ¾ Length Full Height, Full Length
	Weight (Passive/Active)	1066g/1122g	1066g/1122g	1000g/1144g
DRAM Memory	DDR Format	4x 16GB 72b DIMM DDR4	4x 16GB 72b DIMM DDR4	2x 16GB 72b DIMM DDR4
	DDR Total Capacity	64GB	64GB	32GB
	DDR Max Data Rate	2400MT/s	2400MT/s	2400MT/s
	DDR Total Bandwidth	77GB/s	77GB/s	38GB/s
	HBM2 Total Capacity	–	–	8GB
	HBM2 Total Bandwidth	–	–	460GB/s
Internal SRAM	Total Capacity	35MB	54MB	41MB
	Total Bandwidth	31TB/s	38TB/s	30TB/s
Interfaces	PCI Express®	Gen3 x16	Gen3 x16	Gen4 x8 w/ CCIX
	Network Interface	2x QSFP28	2x QSFP28	2x QSFP28
Power and Thermal	Thermal Cooling	Passive, Active	Passive, Active	Passive, Active
	Typical Power	100W	110W	100W
	Maximum Power	225W	225W	225W
Logic Resources	Look-Up Tables	892K	1,341K	1,079K
	Registers	1,831K	2,749K	2,179K
	DSP Slices	5,867	11,508	8,490
Compute Performance	INT8 TOPs	18.6	33.3	24.5
	Machine Learning	Machine Learning Solution Brief		
	Acceleration Applications	Acceleration Application Solutions		

Alveo Data Center Accelerator Cards

2. Are there unique applications that are specifically suitable for FPGAs for supercomputing fields?

> For compute :

- >> For following traditional HPC scientific workloads; **FPGAs are very good**
 - Bioinformatics / gene sequencing –Illumina as use case example
 - Material simulation – Quantum espresso ; Maxeler use case
 - Weather forecast –Maxeler use case with University
 - Oil and Gas imaging – Maxeler use case with Eni
 - For emerging application such as machine learning and neural networks **FPGAs are awesome**

> For Storage:

- >> Storage acceleration as data explosion in experiments in scientific computing
- >> FPGAs provide end to end compression / decompression

> For networking :

- >> Many architecture to use FPGAs in inline network acceleration cards
- >> Ultra low latency communication with partial offload

Realizing the Promise of Personalized Medicine

26 Hour Genome

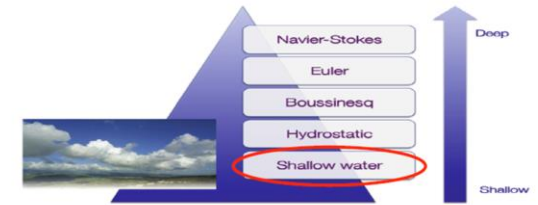
Ultra-Rapid Whole Genome Diagnosis for Critically Ill Newborns



Global Weather Simulation in China

Imperial College London

Simulating Atmosphere
Shallow Water Equation



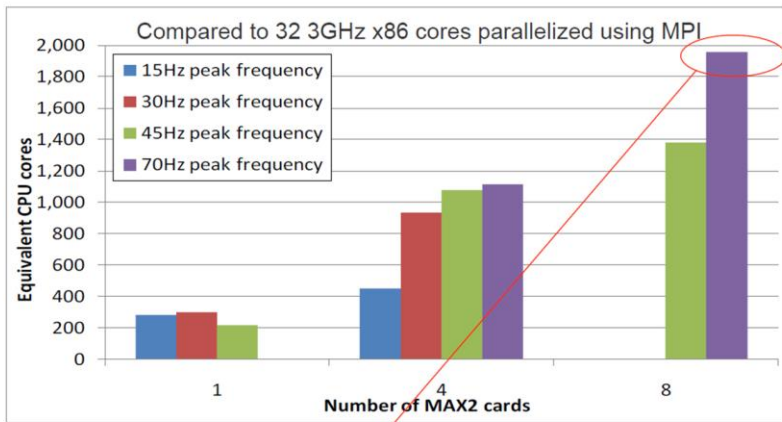
[L. Gan, H. Fu, W. Luk, C. Yang, W. Xue, X. Huang, Y. Zhang, and G. Yang. Accelerating solvers for global atmospheric equations through mixed-precision data flow engine, FPL Conference 2013]

Platform	Performance	Speedup	Efficiency	Energy Improvement
6-core CPU	4.66K	1	20.71	1
Tianhe-1A node	110.38K	23x	306.6	14.8x
MaxWorkstation	468.1K	100x	2.52K	121.6x
Maxeler MPC-X	1.54M	330x	3K	144.9x

14x (between 100x and 330x)
9x (between 121.6x and 144.9x)

3000³ Seismic Imaging

presented by the Italian National Oil Company at the Annual SEG Conference, 2010.



100kWatts of Intel cores => 1kWatt of FPGA Computing

Quantum Espresso FFT & ZGEMM on VU9P FPGAs

Comparison of migrated code to reference system

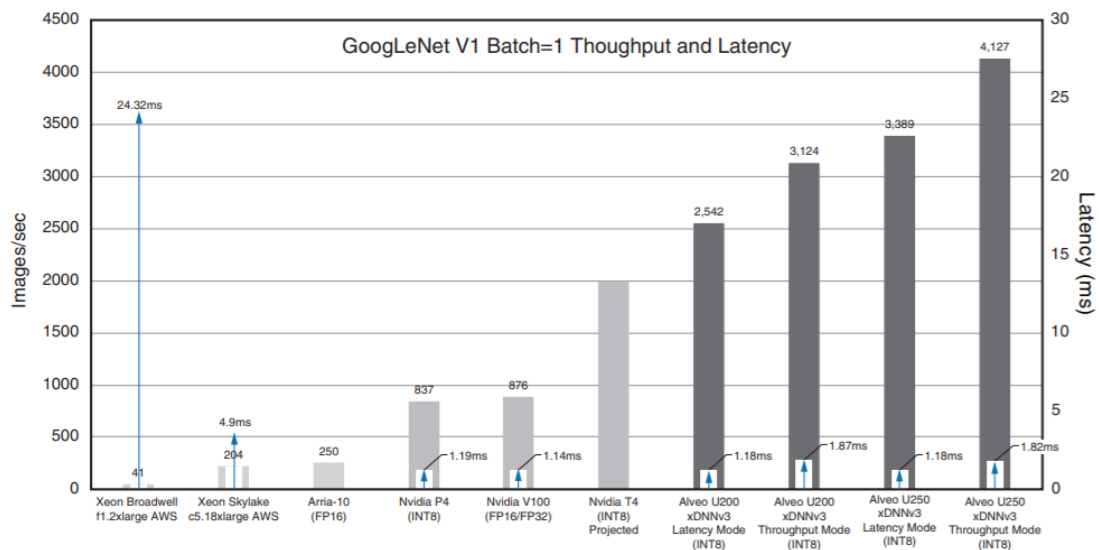


System	1 rack of BlueGene/Q	8 Xilinx VU9P cards	Comparison
Space	205,920 in ³	1,520 in ³	135x
Power	192 kW	1 kW	192x
Performance	338 cubes/s	2,237 cubes/s	6.6x

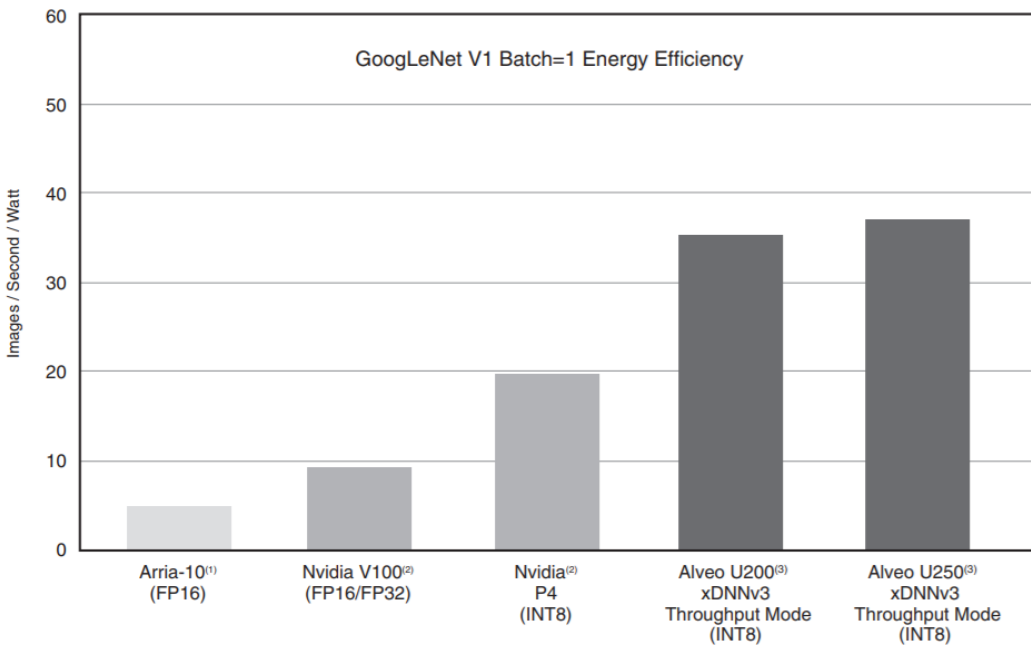
- BlueGene/Q contains significant water cooling and communication
 - FFT divided to 256 nodes
- Maxeler solution running on 8 VU9P devices in parallel
 - FFT in a single node
- 900x total improvement in compute/space & 1,300x improvement in compute/power



For Machine learning and neural networks , Xilinx FPGAs are really good



- Xilinx FPGA faster than both CPU ,GPU, Intel FPGAs
- 20x speedup over CPUs
- 4x speedup over NVIDIA GPUs
- 16x speed up over Intel FPGAs



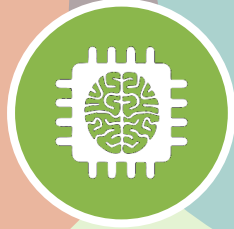
- Xilinx FPGA power efficient than CPU,GPU, Intel FPGAs
- 2x efficient over NVIDIA GPUs
- 7x efficient over Intel FPGAs

Infuse Machine Learning with other accelerations

Database



Machine Learning



HPC & Life Sciences



Video



Financial



3. What are the challenges and/or major issues facing FPGAs for supporting supercomputing?

- > **Double precision floating point myth created by GPU and other architectures**
 - >> Important metric is achieved performance % ; one can't just look at theoretical flop/s and rule out FPGAs
- > **Scalability on cluster level**
 - >> Beyond 1 node is not mainstream yet
- > **Ease of programming**
 - >> If you look at various work in HPC done a few years ago, you will realize many promising performance results using HDL - hand tuned programmed codes. This work remained to very limited set of developers as it requires specialized HE knowledge to tune the algorithm to HW level in HDL. Now with the maturity of SDAccel – HLS/OpenCL/C++ support from vendors such as Xilinx ; the performance is encouraging to hand tunes HDL codes.
 - >> Support for FPGA in higher level languages such as OpenMP, OpenACC
 - Still at very nascent stages

Solution Stack



Developer

100%

QoQ Growth of Published Applications in FY18

Hundreds

of Developers Trained Every Quarter

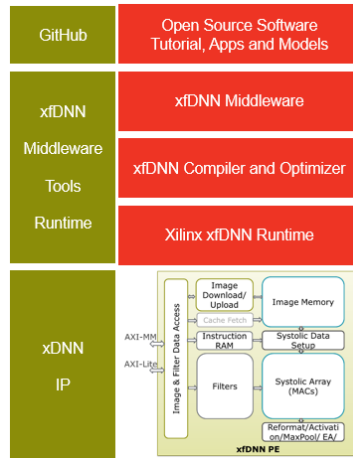
RTL, C, C++, OpenCL



Accelerated Solutions

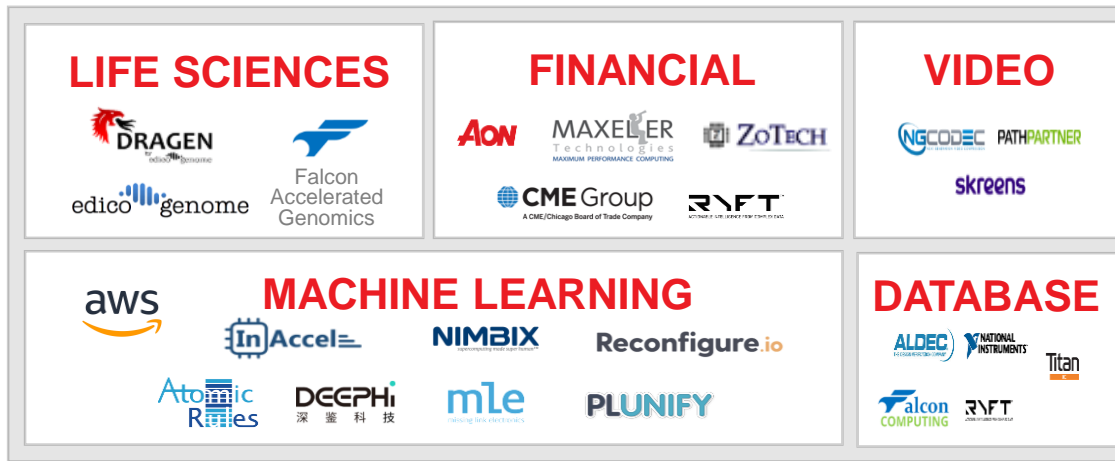
Developer Package

Xilinx ML Suite



End user

Framework, API, Python/Java/C++ Programmability



Solutions Xilinx ISVs

Platforms



Cloud






On-premise

Platform

Xilinx platform supports wider set of use cases vs GPU

Versal Expands To Even More Use Cases

		Xilinx Platforms	GPU
 <p>Compute</p>	<ul style="list-style-type: none"> ML Inference Database / Big Data Video Financial Services Genomics 	<ul style="list-style-type: none"> ✓ Low batch performance ✓ Low / variable precision performance ✓ Flexible datapath/memory for power efficiency ✓ Hardware & software programmable 	<ul style="list-style-type: none"> ✗ Parallel architecture poor fit for low-batch ML ✗ Fixed and inflexible datapath and memory ✗ SIMD architecture → inflexible & power hungry
	<ul style="list-style-type: none"> High Precision HPC ML Training 	<ul style="list-style-type: none"> ✗ Optimized for fixed point training and HPC 	<ul style="list-style-type: none"> ✓ Optimized for high precision floating point
 <p>Storage</p>	<ul style="list-style-type: none"> Compression Encryption Key-Value Store Database / Big Data ML Inference 	<ul style="list-style-type: none"> ✓ Processing near memory / storage ✓ Flexible low latency in-line processing ✓ Adaptable parallel memory hierarchy 	<p>Poor Fit</p>
 <p>Networking</p>	<ul style="list-style-type: none"> IPSec/SSL OVS offload Bare Metal Services Security Monitoring 	<ul style="list-style-type: none"> ✓ Optimization of latency and efficiency ✓ Rich I/O, flex datapath for inline processing ✓ Power efficient, flexible datapath/memory 	<p>Poor Fit</p>

4. What and where are the opportunities? Who are the stakeholders?

- > **Need participation from wide set of users – ranging from developers to OEMs**
- > **HPC Domain scientists**
 - >> Willing to put efforts in algorithm optimization
 - >> Library developers
- > **HPC Developers – traditional HDL/Verilog users**
 - >> Converting HDL codes to higher level
- > **CPU and FPGA vendors**
 - >> Fast cache coherent networks that help HPC platform architecture e.g CCIX (AMD), OpenCAPI (IBM Power)
 - >> Alveo U280 board is CCIX compliant
- > **HPC OEMs- Cray , Atos , Dell**
 - >> Effort to create HPC tool chain and communication optimized libraries
 - >> Resource management SW e.g Slurm

5. Name one thing that the FPGA industry should (or should not) do in the near term to facilitate FPGA's induction into supercomputers.

- > **The future of FPGAs in mainstream HPC is bright and there is path to it with recent advancements in HW and SW**
- > **Do not expect it will solve all application problems overnight – for low power , single precision workloads it can yield immediate benefits if efforts are put in by developers**
- > **Focus on targeted domains where FPGAs have shown value proposition**
- > **Focus on Building Tools, SW libraries, middleware**
- > **For Double precision workloads -**
 - >> Key is precision control and developers/scientists should be willing to experiment with lower precision/mixed precision with acceptable accuracy

Adaptable.
Intelligent.

