Interconnect Your Future

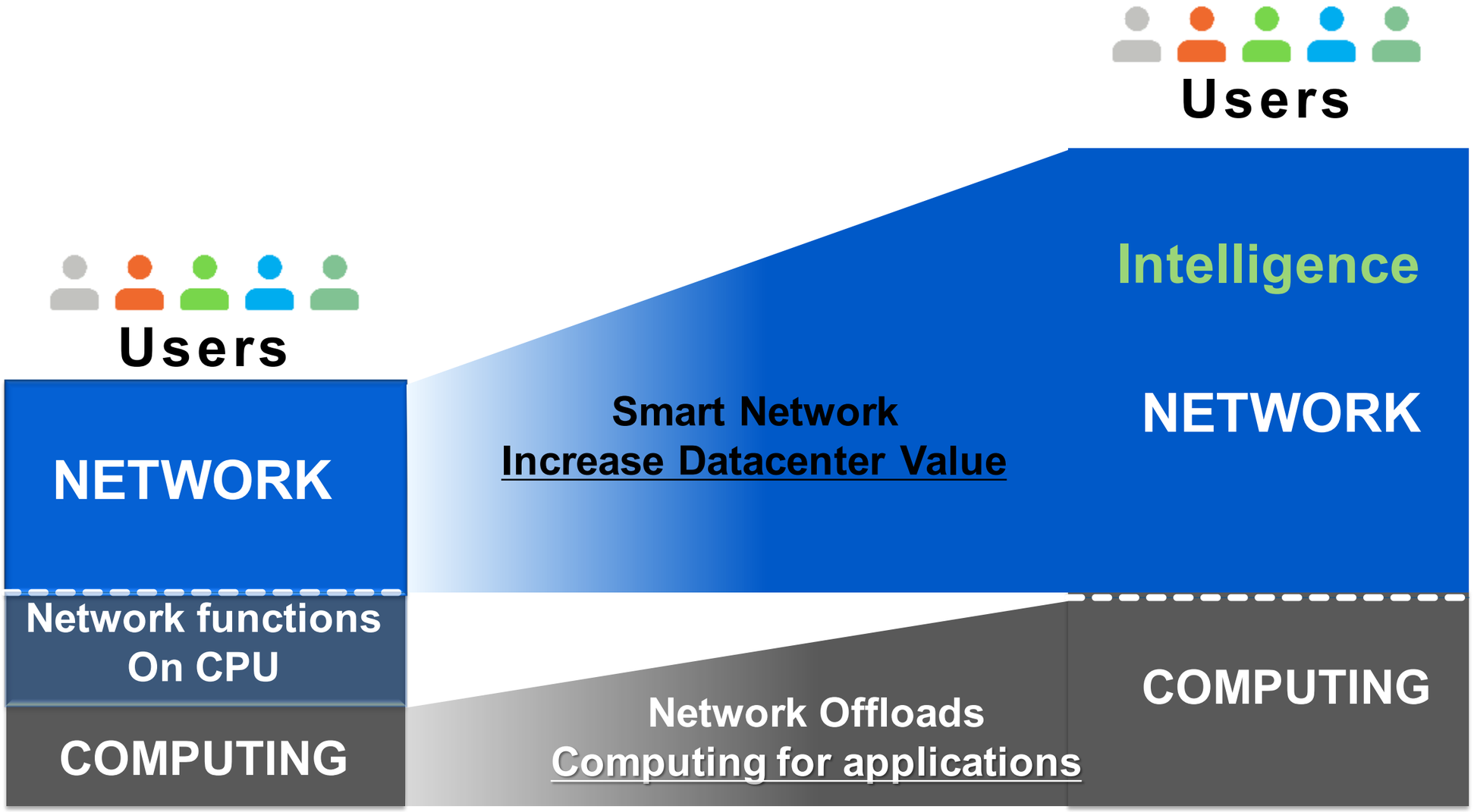
Achieving the next phase of performance evolution in Supercomputing

Boris Neiman - October 2016

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™

Strategy: More Value to Customer



Mellanox Connects the World's Fastest Supercomputer

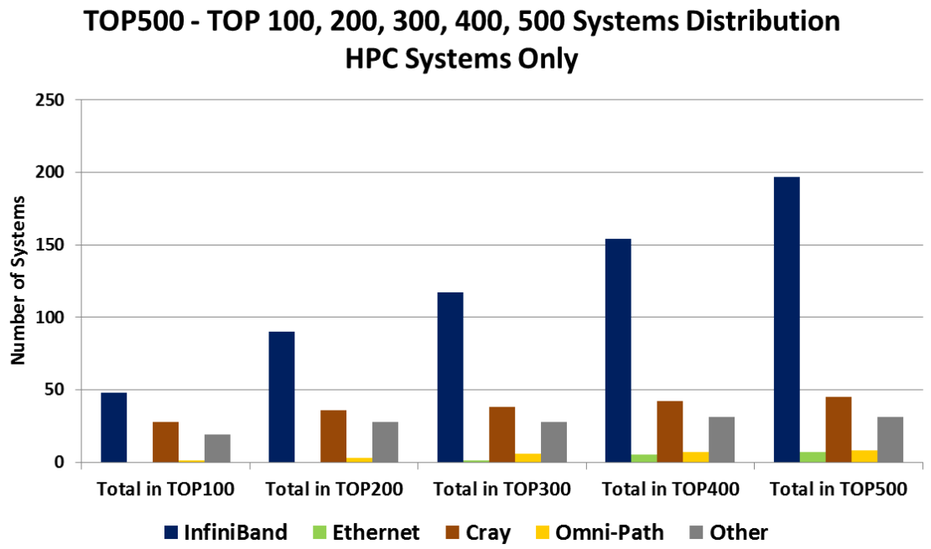
National Supercomputing Center in Wuxi, China

#1 on the TOP500 Supercomputing List

- 93 Petaflop performance, 3X higher versus #2 on the TOP500
- 40K nodes, 10 million cores, 256 cores per CPU
- Mellanox adapter and switch solutions

- The TOP500 list has evolved, includes HPC & Cloud / Web2.0 Hyperscale systems
- Mellanox connects 41.2% of overall TOP500 systems
- Mellanox connects 70.4% of the TOP500 HPC platforms
- Mellanox connects 46 Petascale systems, Nearly 50% of the total Petascale systems

InfiniBand is the Interconnect of Choice for HPC Compute and Storage Infrastructures



Accelerating HPC Leadership Systems

OAK RIDGE National Laboratory

“Summit” System

U.S. DEPARTMENT OF ENERGY

Lawrence Livermore National Laboratory

“Sierra” System

Proud to Pave the Path to Exascale

The Ever Growing Demand for Higher Performance

Performance Development

Terascale

Petascale

1st

"Roadrunner"

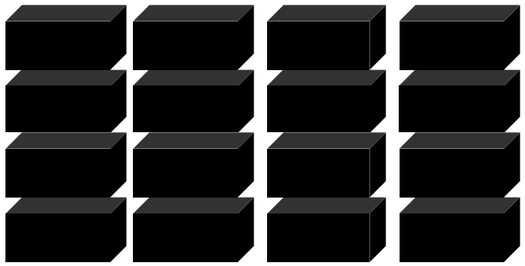
Exascale

OAK RIDGE National Laboratory
"Summit" System

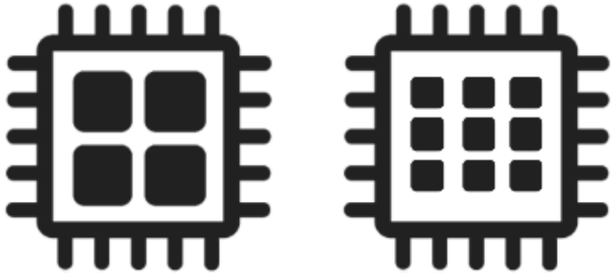
Lawrence Livermore National Laboratory
"Sierra" System

2000 2005 2010 2015 2020

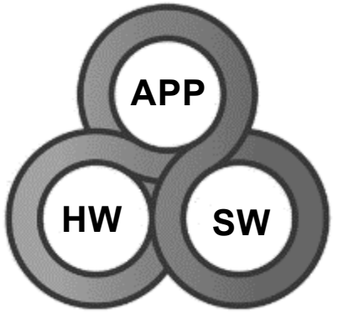
The Interconnect is the Enabling Technology



SMP to Clusters



Single-Core to Many-Core

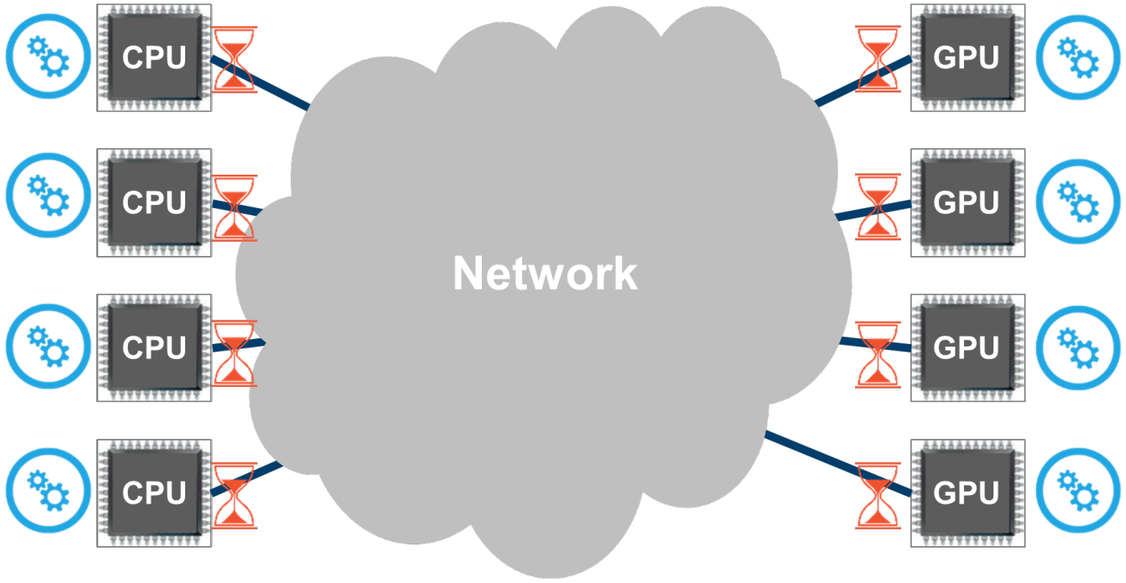


Application
Software
Hardware

Co-Design

The Intelligent Interconnect to Enable Exascale Performance

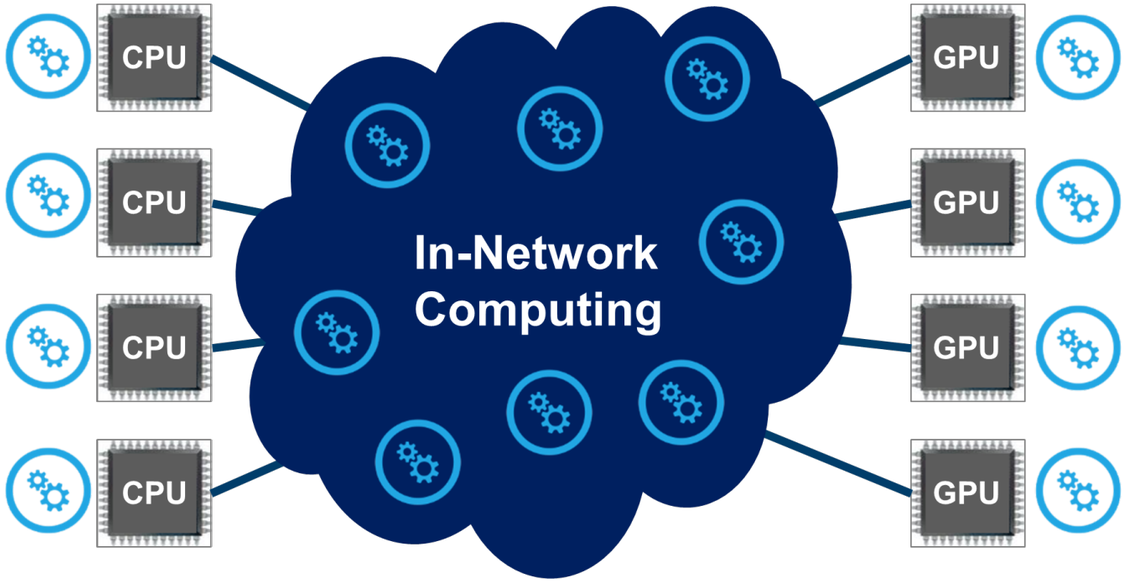
CPU-Centric



Limited to Main CPU Usage
Results in Performance Limitation

Must Wait for the Data
Creates Performance Bottlenecks

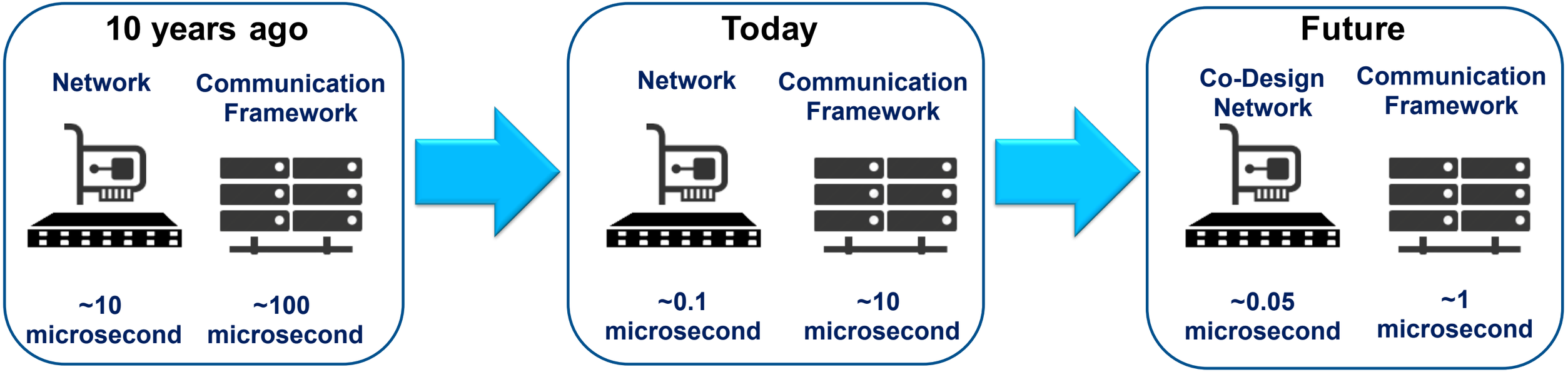
Co-Design



Creating Synergies
Enables Higher Performance and Scale

Work on The Data as it Moves
Enables Performance and Scale

Breaking the Application Latency Wall



- Today: Network device latencies are on the order of 100 nanoseconds
- Challenge: Enabling the next order of magnitude improvement in application performance
- Solution: Creating synergies between software and hardware – intelligent interconnect

Intelligent Interconnect Paves the Road to Exascale Performance

Switch-IB 2 and ConnectX-5 Smart Interconnect Solutions

SHArP Enables Switch-IB 2 to Execute Data Aggregation / Reduction Operations in the Network

Barrier, Reduce, All-Reduce, Broadcast
Sum, Min, Max, Min-loc, max-loc, OR, XOR, AND
Integer and Floating-Point, 32 / 64 bit

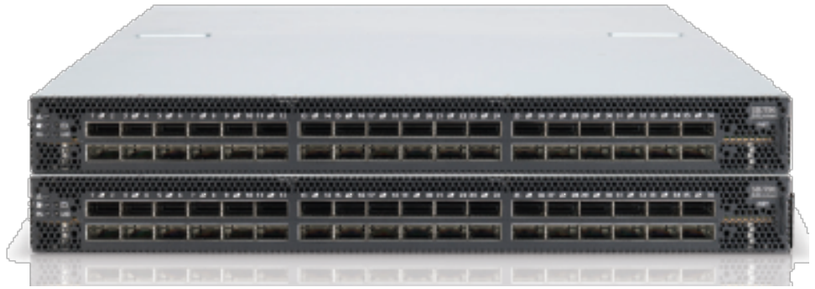
Delivering 10X Performance Improvement for MPI and SHMEM/PGAS Communications

100Gb/s Throughput
0.6usec Latency (end-to-end)
200M Messages per Second

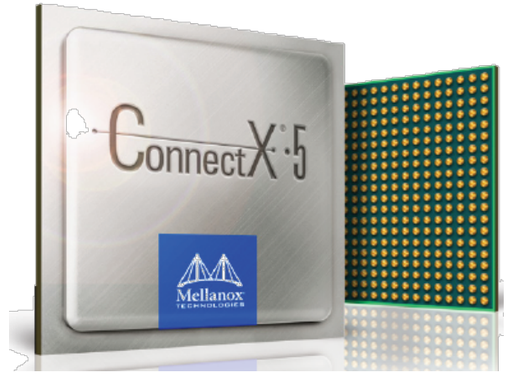
MPI Collectives in Hardware
MPI Tag Matching in Hardware
In-Network Memory

PCIe Gen3 and Gen4
Integrated PCIe Switch
Advanced Dynamic Routing

Switch-IB 2 SHArP



ConnectX-5

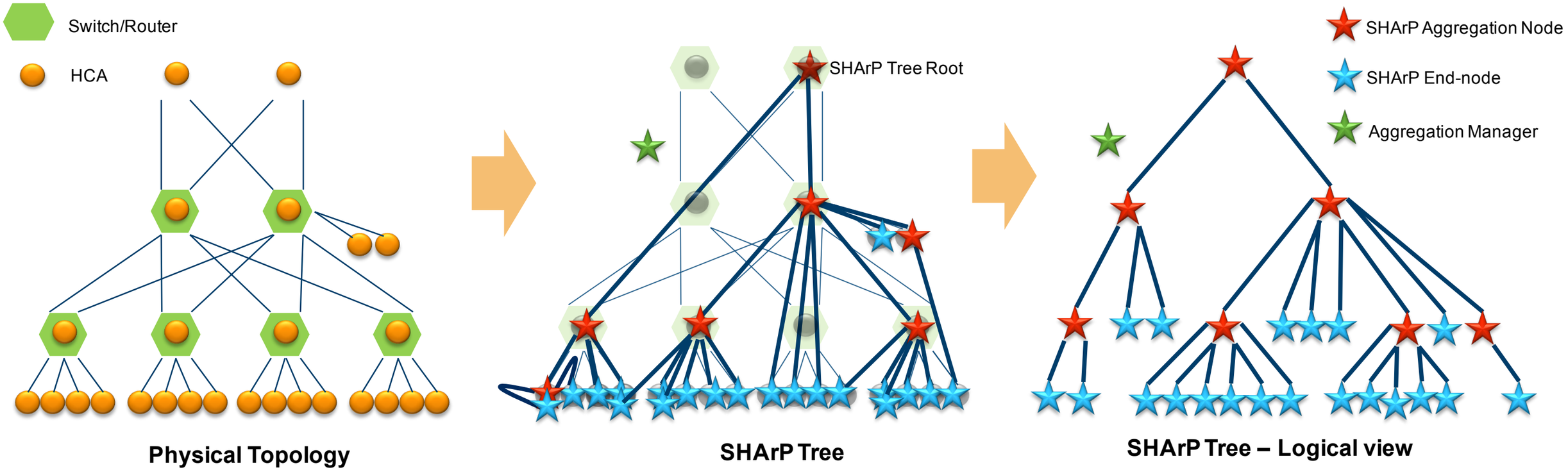


How does SHArP Works?

- SHArP Operations are executed by a SHArP tree defined on top of the physical fabric

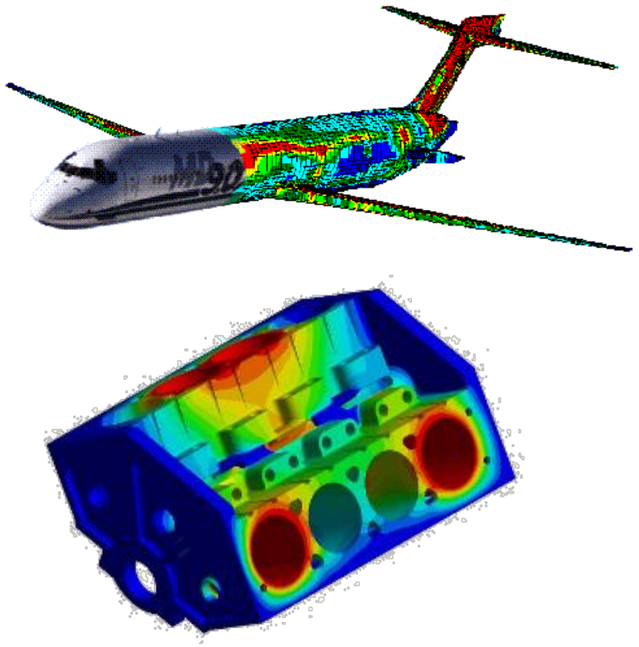
  Shortest overall path length from leafs to root

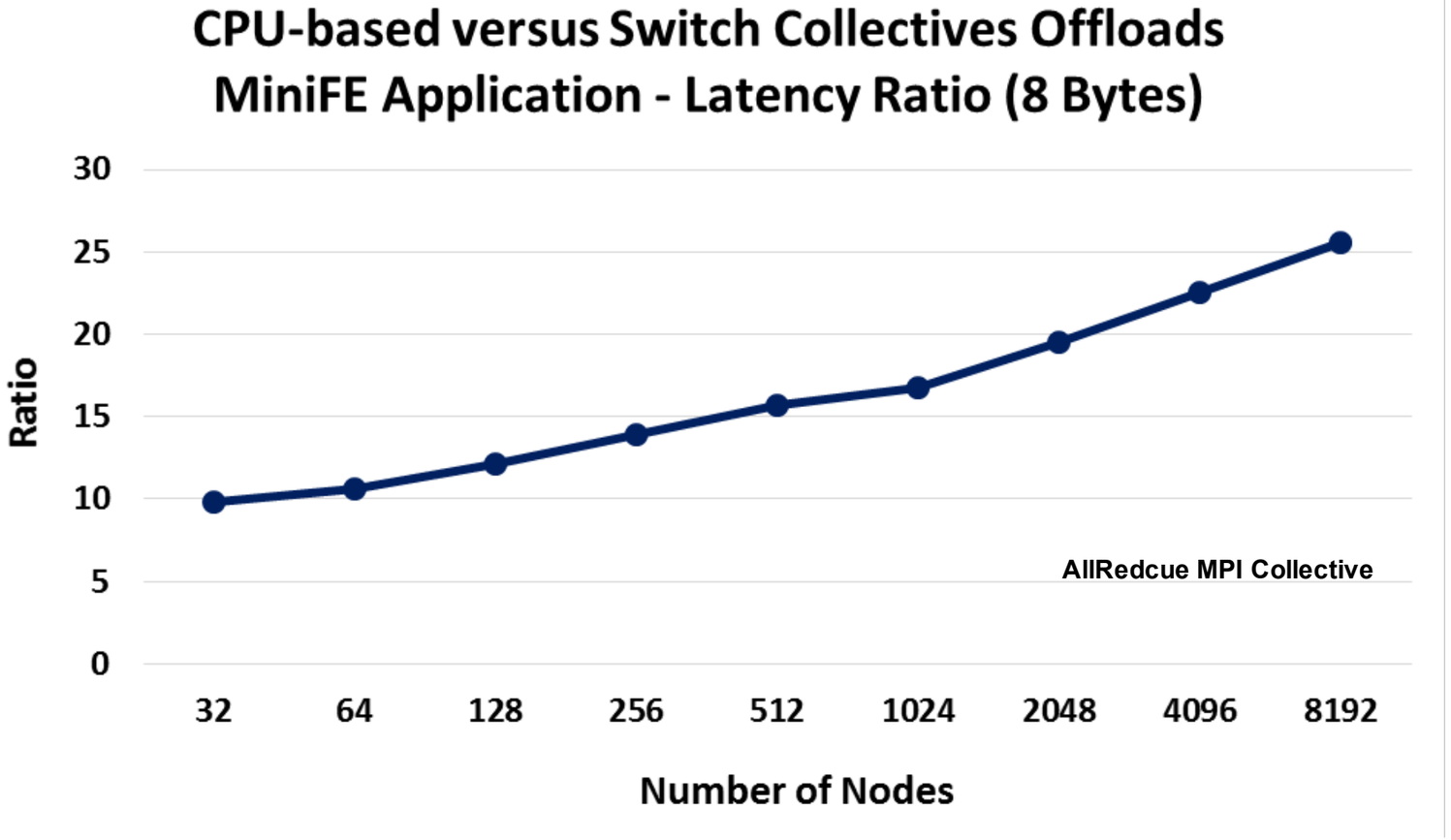
- Each SHArP Tree can handle Multiple Outstanding SHArP Operations



SHArP Performance Advantage

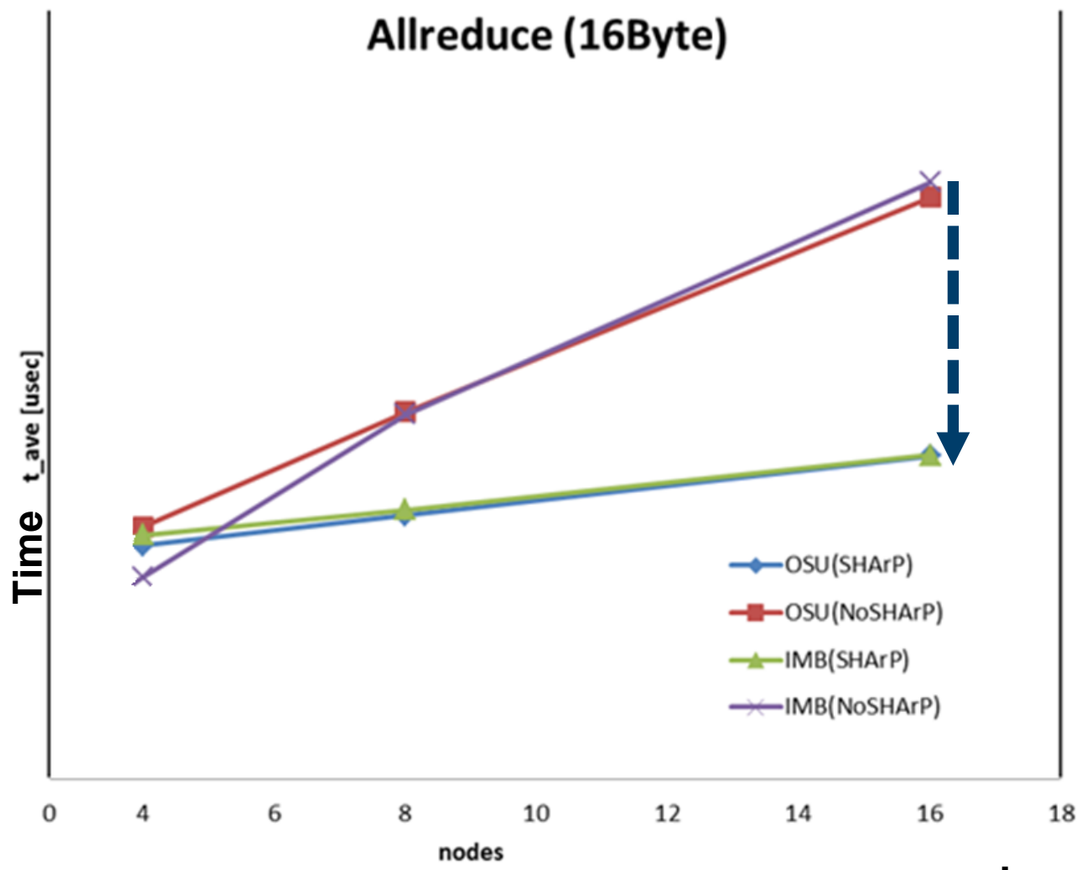
- MiniFE is a Finite Element mini-application
  - Implements kernels that represent implicit finite-element applications

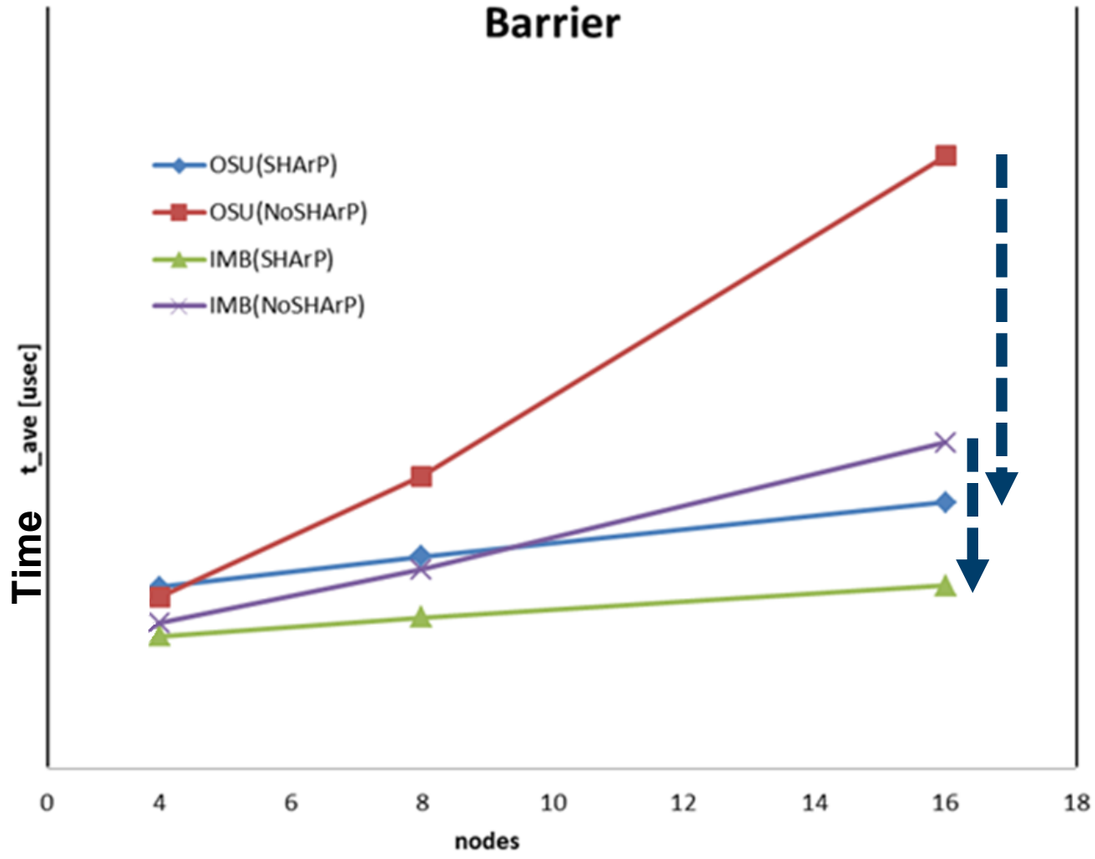




10X to 25X Performance Improvement!

SHArP Performance Advantage with Intel Xeon Phi Knight Landing





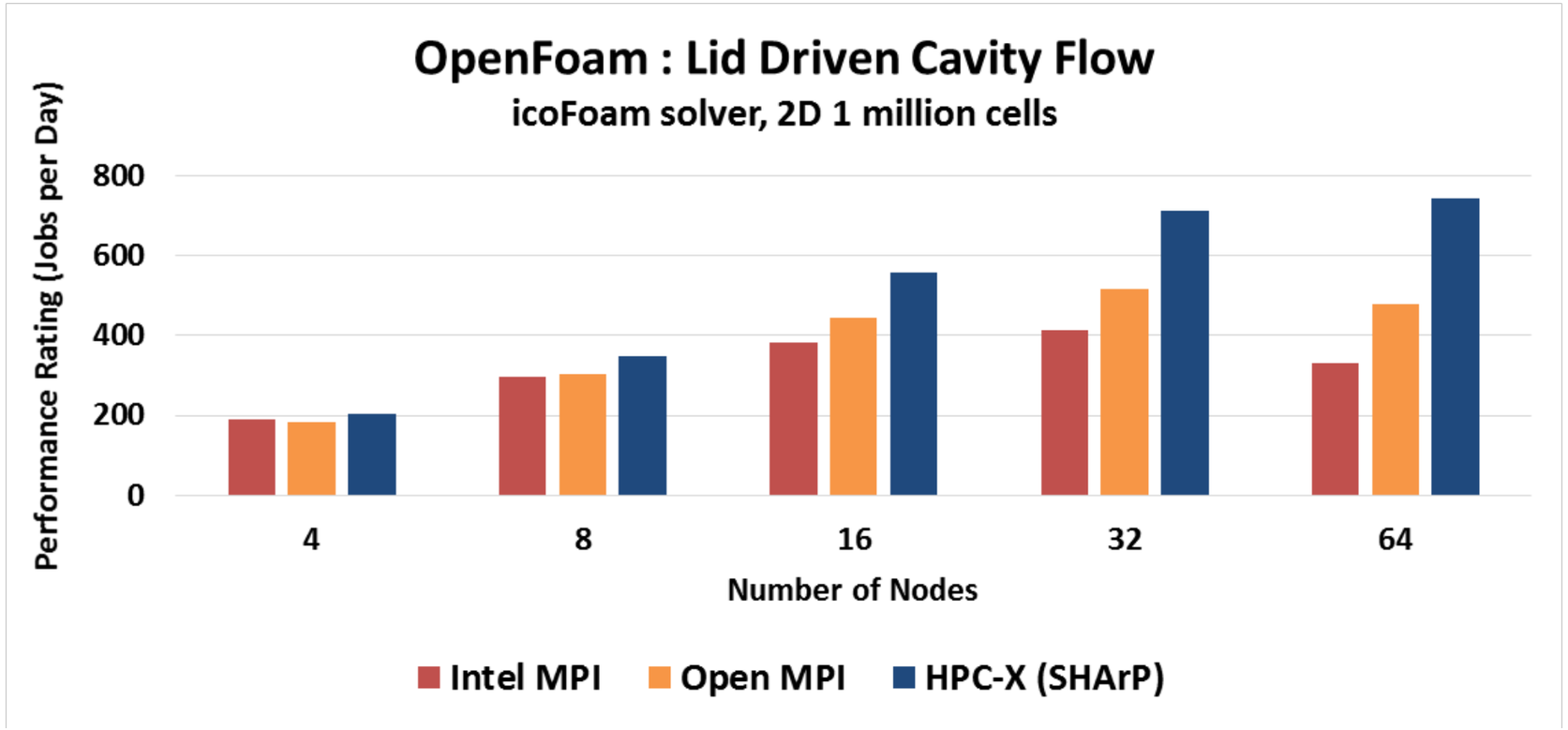
Lower is better

OSU - OSU MPI benchmark; IMB – Intel MPI Benchmark

Maximizing KNL Performance – 50% Reduction in Run Time (Customer Results)

HPC-X with SHArP Technology

OpenFOAM is a popular computational fluid dynamics application



HPC-X with SHArP Delivers 2.2X Higher Performance over Intel MPI

InfiniBand The Smart Choice for HPC Platforms and Applications

- *"We chose a co-design approach. This system was of course targeted at supporting in the best possible manner our key applications. The only interconnect that really could deliver that was Mellanox InfiniBand."*

JÜLICH FORSCHUNGSZENTRUM

Watch Video

- *"One of the big reasons we use InfiniBand and not an alternative is that we've got backwards compatibility with our existing solutions."*

UNIVERSITY OF BIRMINGHAM

Watch Video

- *"InfiniBand is the most advanced high performance interconnect technology in the world, with dramatic communication overhead reduction that fully unleashes cluster performance."*

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Watch Video

- *"InfiniBand is the best that is required for our applications. It enhancing and unlocking the potential of the system."*

CHPC CENTRE FOR HIGH PERFORMANCE COMPUTING

Watch Video

Highest-Performance 100Gb/s Interconnect Solutions

Adapters — ConnectX-5

100Gb/s Adapter, 0.6us latency
200 million messages per second
(10 / 25 / 40 / 50 / 56 / 100Gb/s)

Switch — SwitchIB-2

36 EDR (100Gb/s) Ports, <90ns Latency
Throughput of 7.2Tb/s
7.02 Billion msg/sec (195M msg/sec/port)

Switch — Spectrum

32 100GbE Ports, 64 25/50GbE Ports
(10 / 25 / 40 / 50 / 100GbE)
Throughput of 6.4Tb/s

Interconnect — LinkX

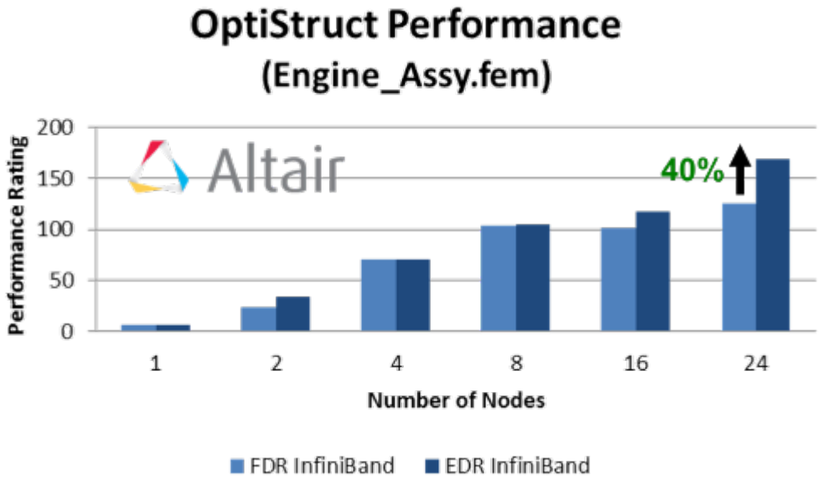
Transceivers
Active Optical and Copper Cables
(10 / 25 / 40 / 50 / 56 / 100Gb/s)

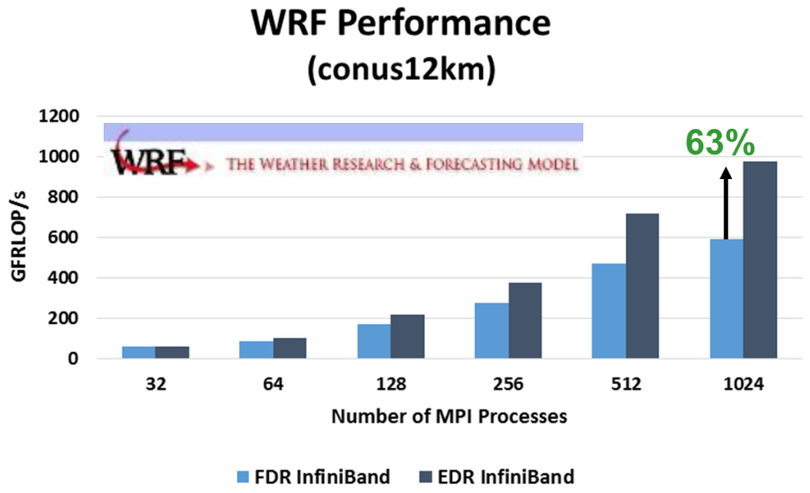
VCSELs, Silicon Photonics and Copper

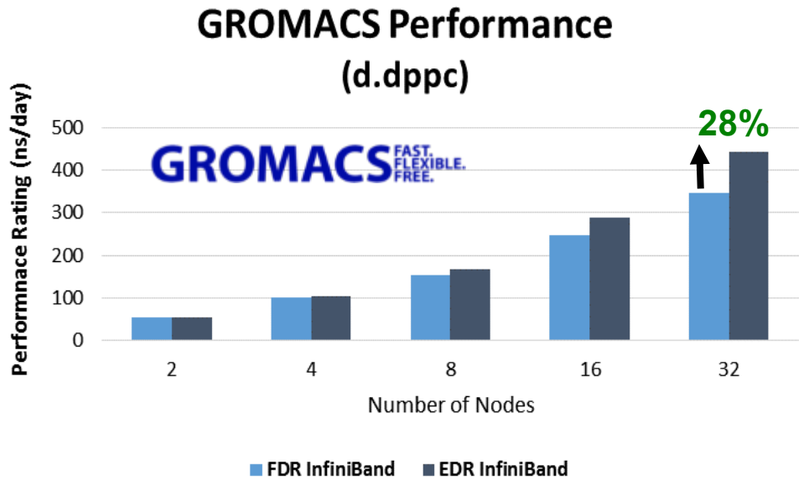
Software — HPC-X

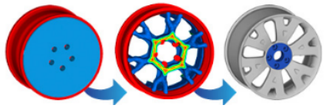
MPI, SHMEM/PGAS, UPC
For Commercial and Open Source Applications
Leverages Hardware Accelerations

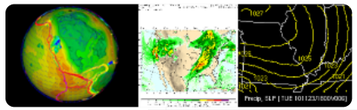
The Performance Advantage of EDR 100G InfiniBand (28-80%)

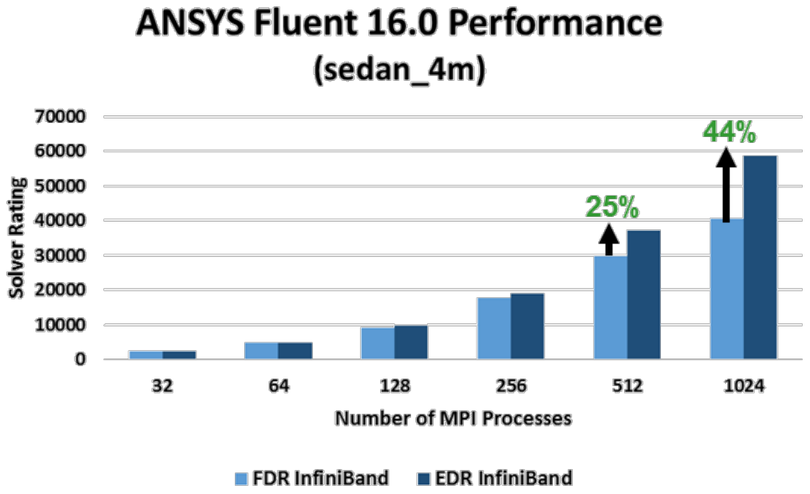


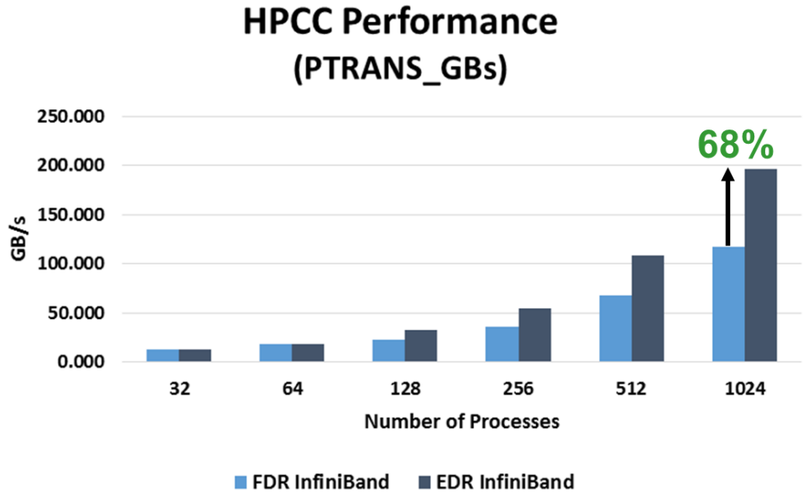


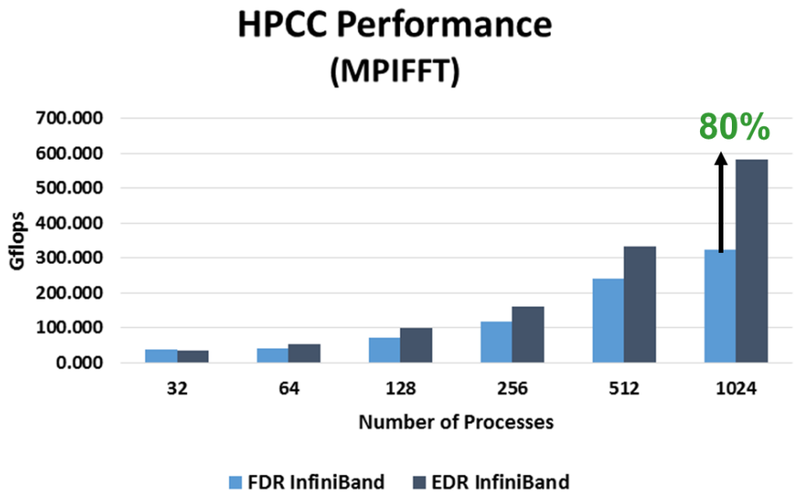






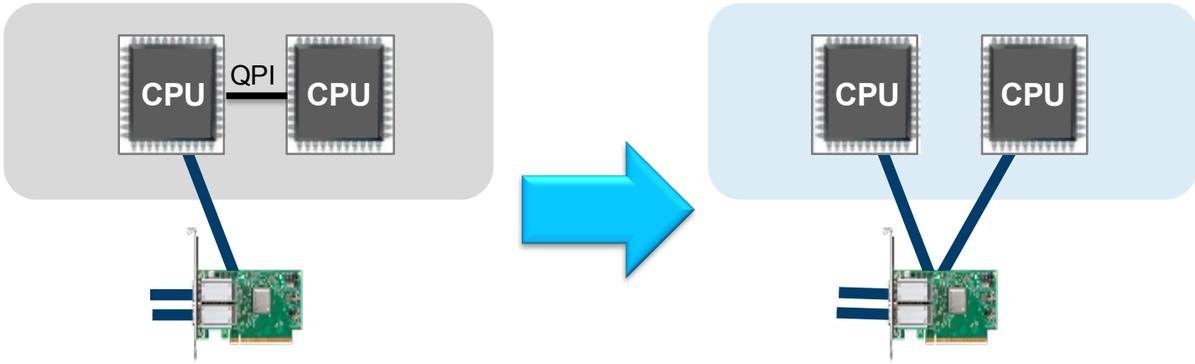




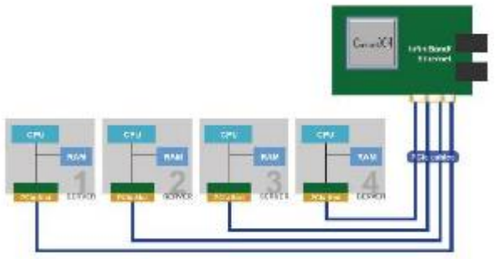


Multi-Host Socket Direct – Low Latency Socket Communication

- Each CPU with direct network access
- QPI avoidance for I/O – improve performance
- Enables GPU / peer direct on both sockets
- Solution is transparent to software



Multi Host Evaluation Kit



Multi-Host Socket Direct Performance

50% Lower CPU Utilization

20% lower Latency

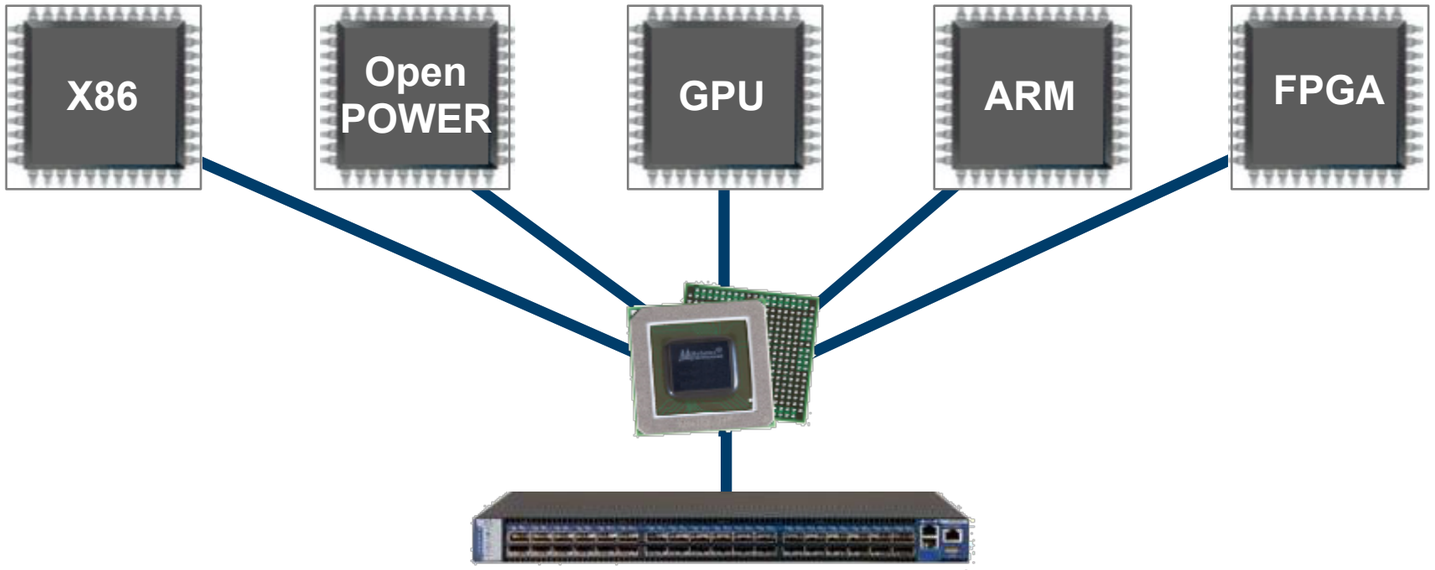
Lower Application Latency, Free-up CPU

End-to-End Interconnect Solutions for All Platforms

Highest Performance and Scalability for

X86, Power, GPU, ARM and FPGA-based Compute and Storage Platforms

10, 20, 25, 40, 50, 56 and 100Gb/s Speeds



Smart Interconnect to Unleash The Power of All Compute Architectures

Introducing Mellanox InfiniBand Router

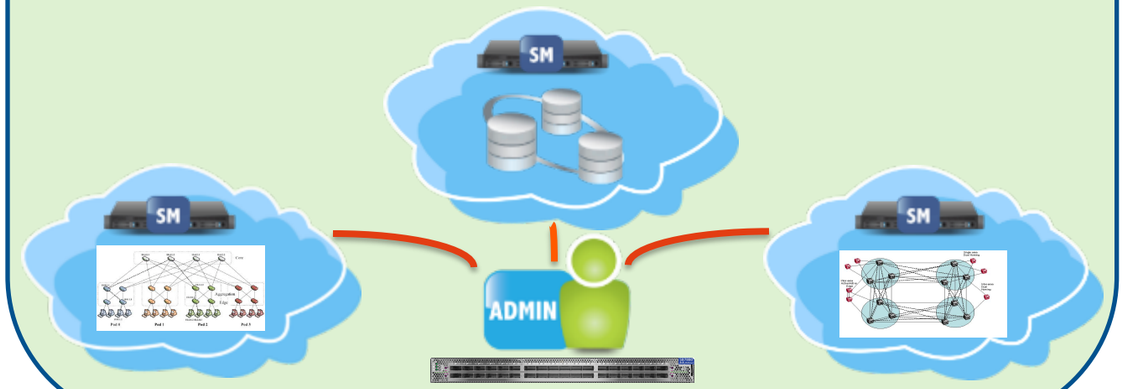
- 1U out-of-the-box router capability support
- Supports of up to 6 different subnets

Isolation

Separation and fault resilience between IB islands

Sharing common storage network by multiple subnets

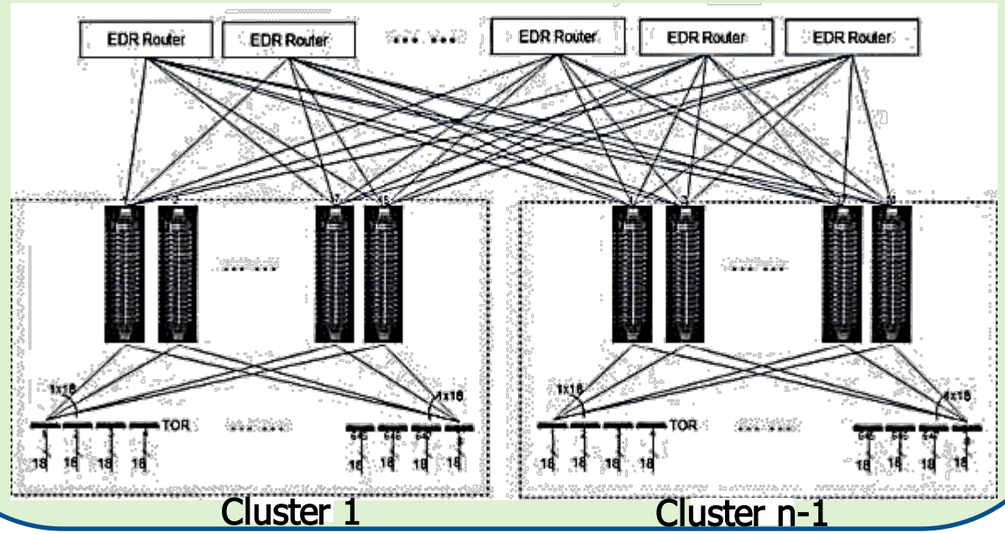
Connect different topologies by the different subnets



Scalability

Scale above 48K end ports

Running HPC/MPI jobs efficiently on the joint network

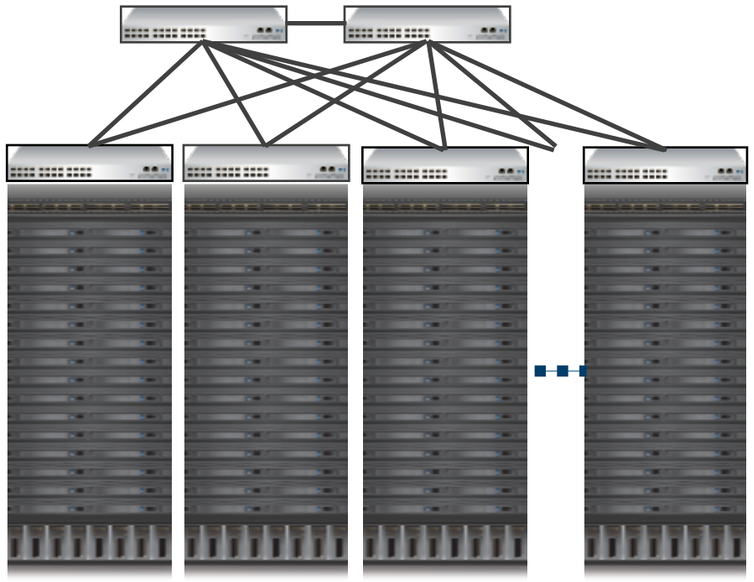


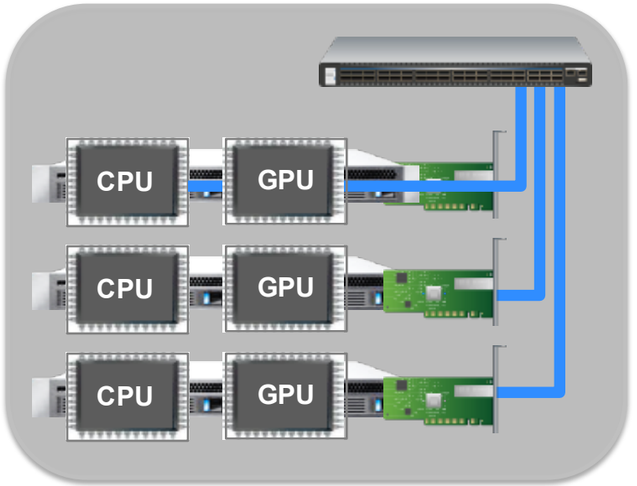
Machine Learning

GPUDirect Enables Efficient Training Platform for Deep Neural Network

(a) Face (b) Body (c) Cat

THE OHIO STATE UNIVERSITY



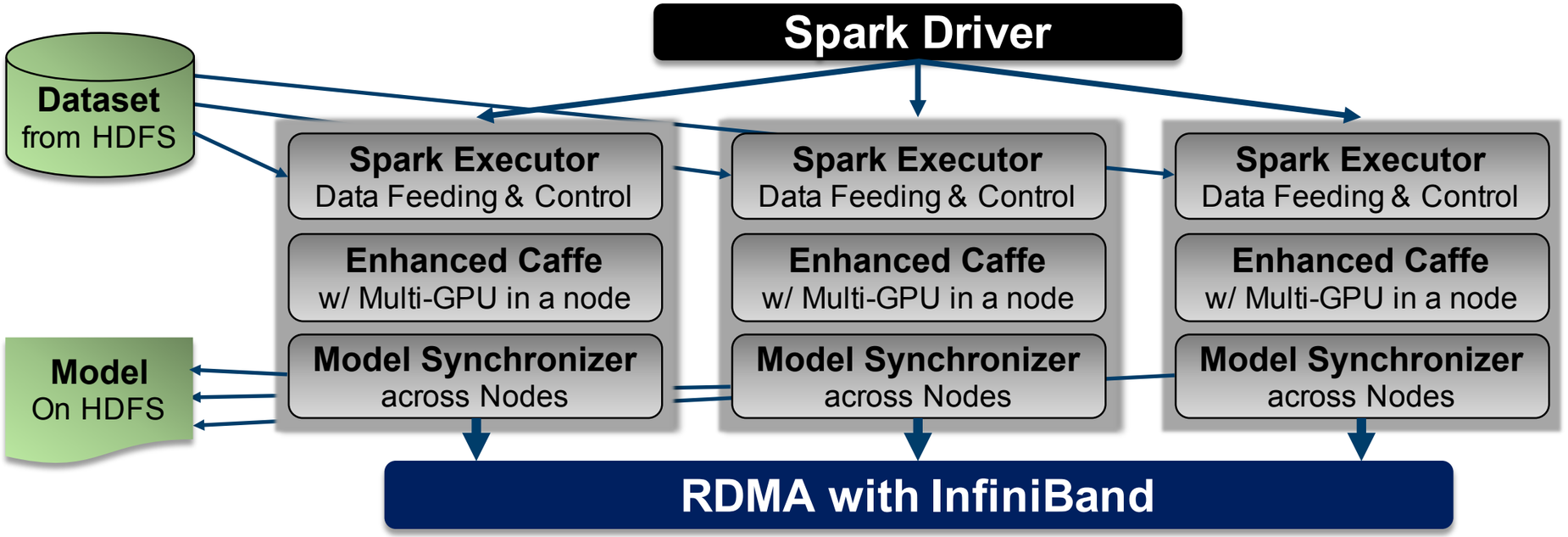


1K nodes (16K cores) for 1 week

3 Nodes with 3 GPUs for 3 days
Mellanox InfiniBand and GPU-Direct

Deep Learning - Transforming Data to Intelligence

RDMA Accelerated Deep Learning (Hadoop)



flickr from YAHOO!

Caffe

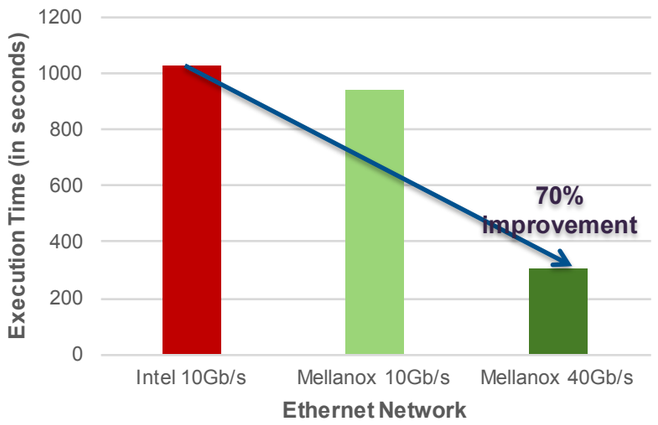
Large Scale Distributed Deep Learning on Hadoop Clusters - Yahoo Big ML Team [link]

- RDMA enables Deep Learning with Caffe + Hadoop
- 18.7x Overall Speedup, 80% Accuracy , 10 hours of training
  - 4 servers with 8 GPUs and Mellanox InfiniBand

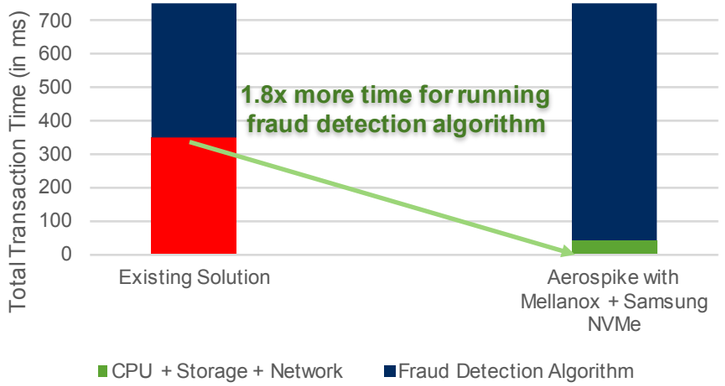
Enabling Advanced Predictive Analytics for Image Recognition

Enable Real-time Decision Making

TeraSort



Fraud Detection workload



Big Sur Machine Learning Platform

MAPR

Mellanox TECHNOLOGIES
Connect. Accelerate. Outperform.

SAMSUNG

Mellanox TECHNOLOGIES
Connect. Accelerate. Outperform.

AEROSPIKE

DELL

Mellanox TECHNOLOGIES
Connect. Accelerate. Outperform.

Innova flex

Cassandra

Couchbase

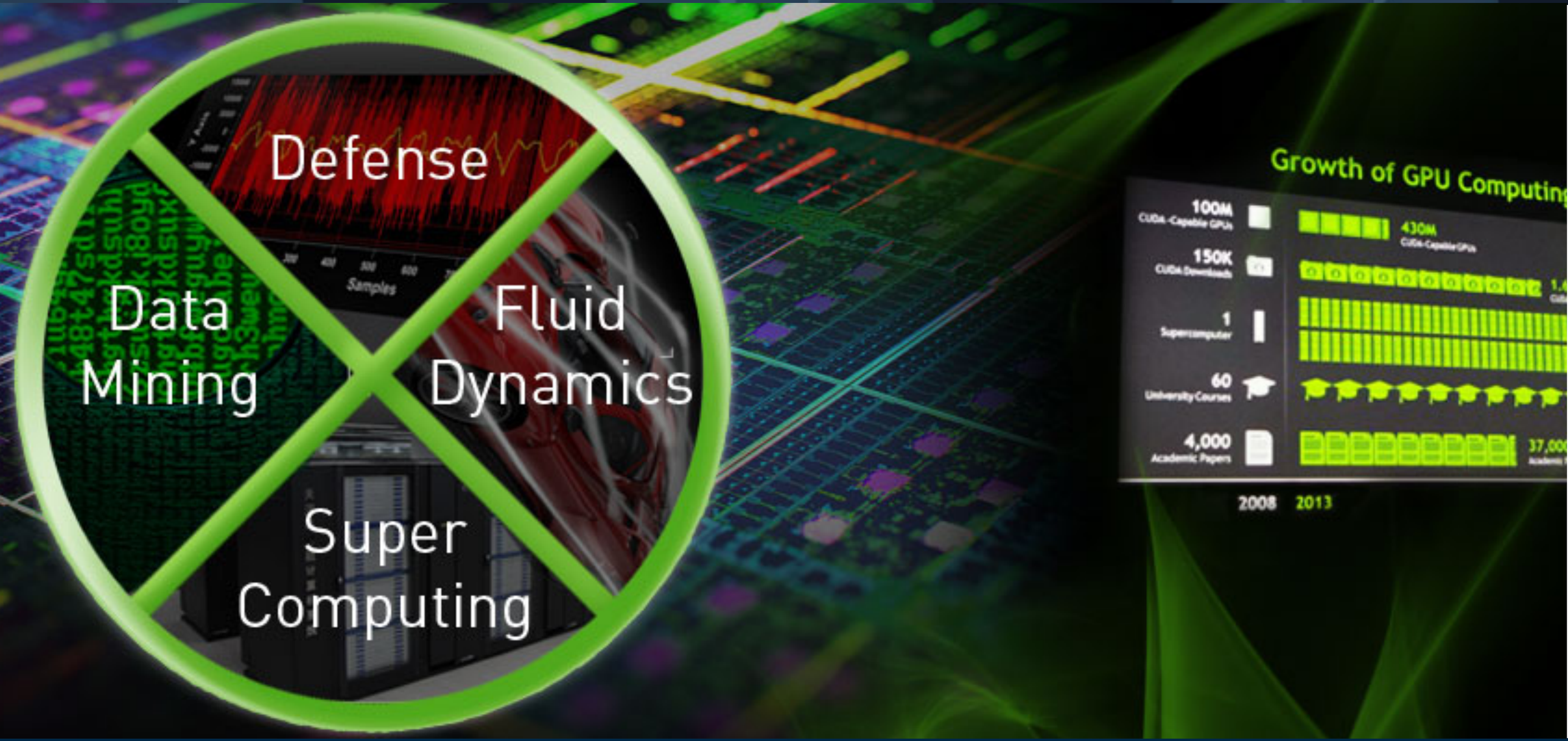
Hortonworks

MEMCACHED

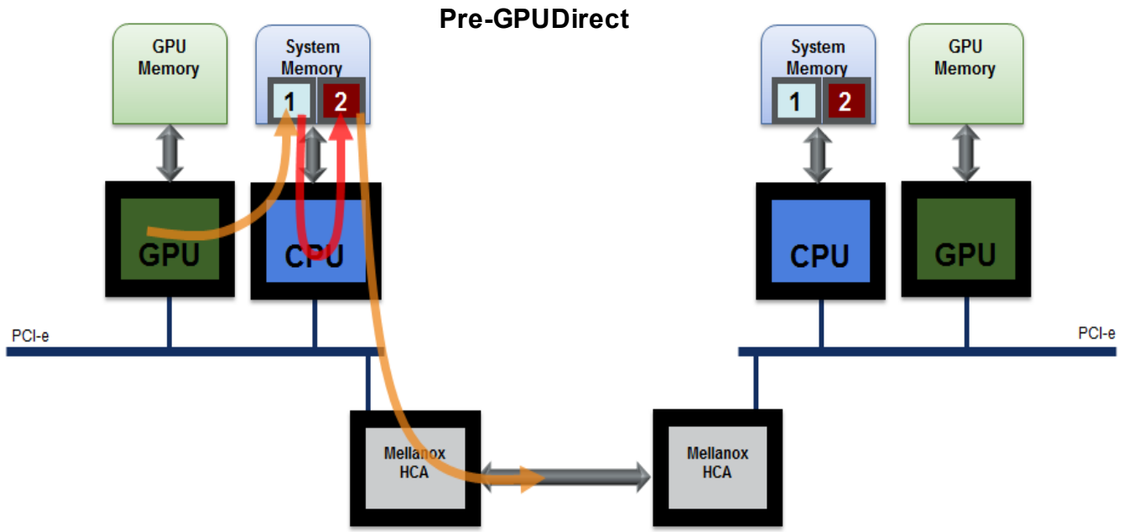
mongoDB

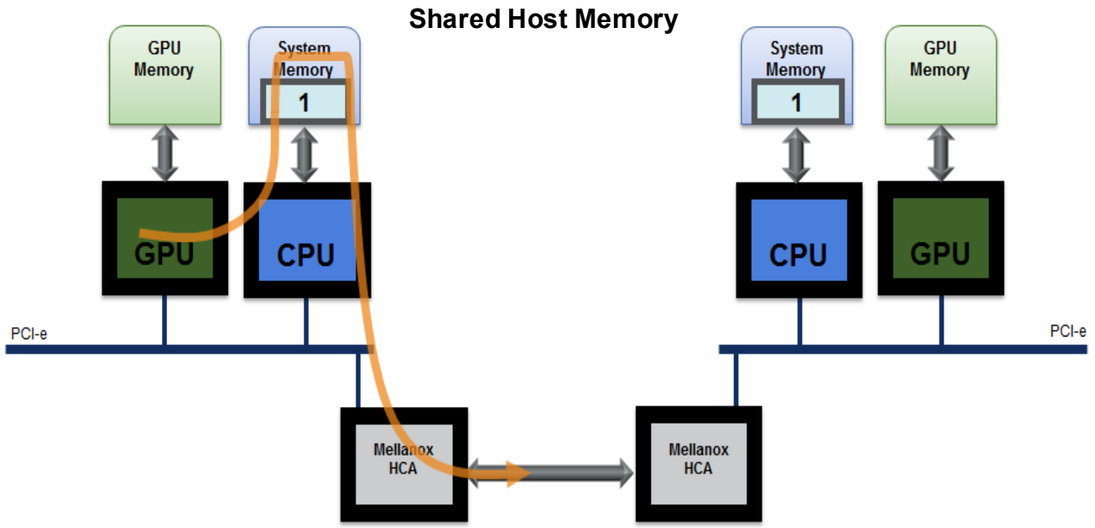
redis

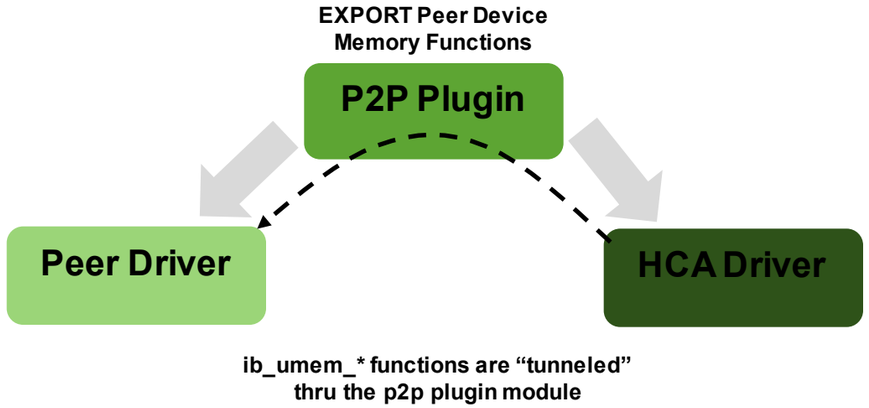
GPU Computing



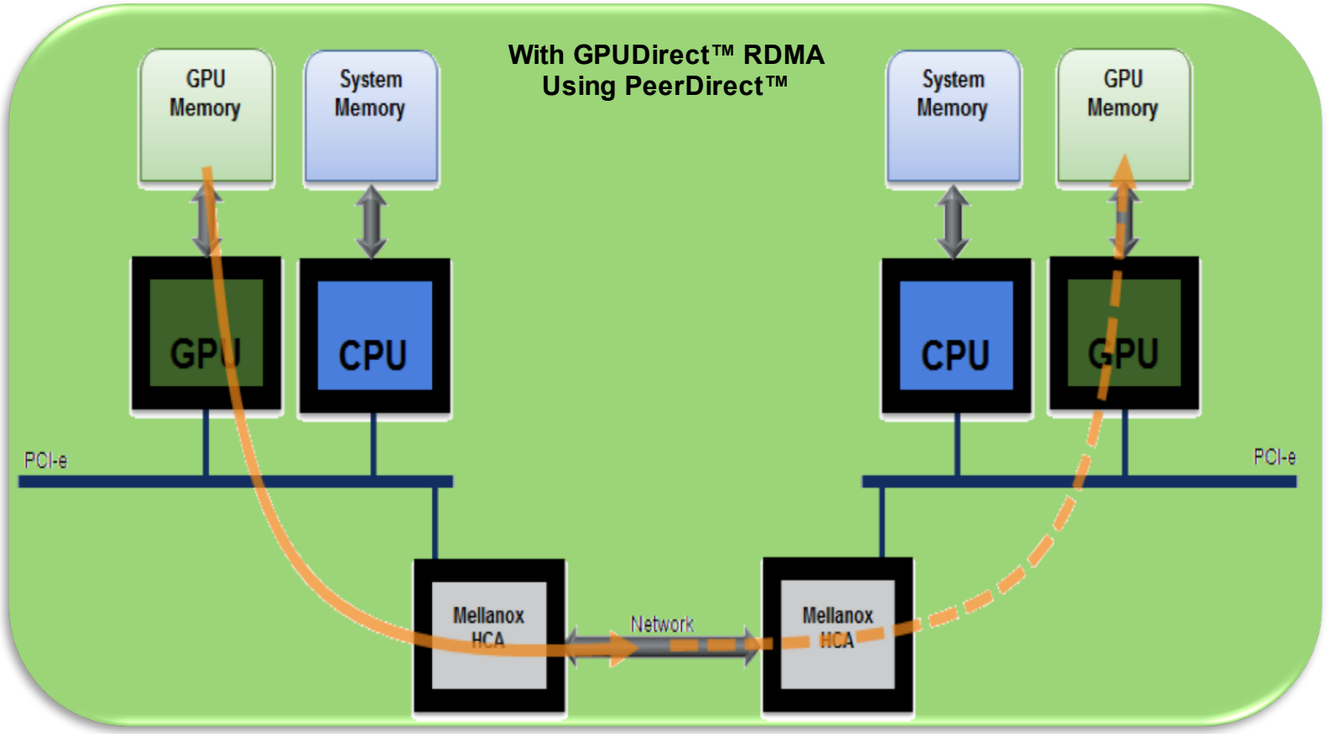
Evolution of GPUDirect RDMA





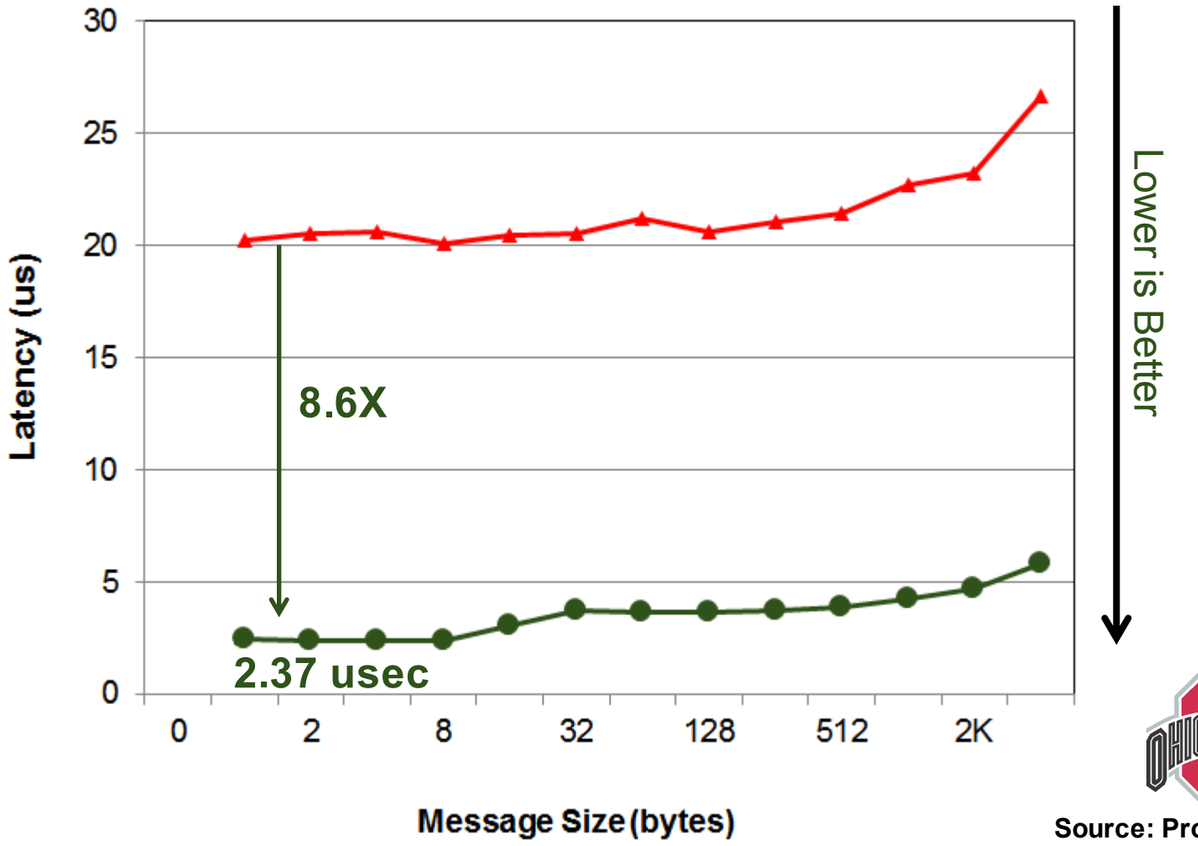


GPUDirect™ RDMA Ecosustem



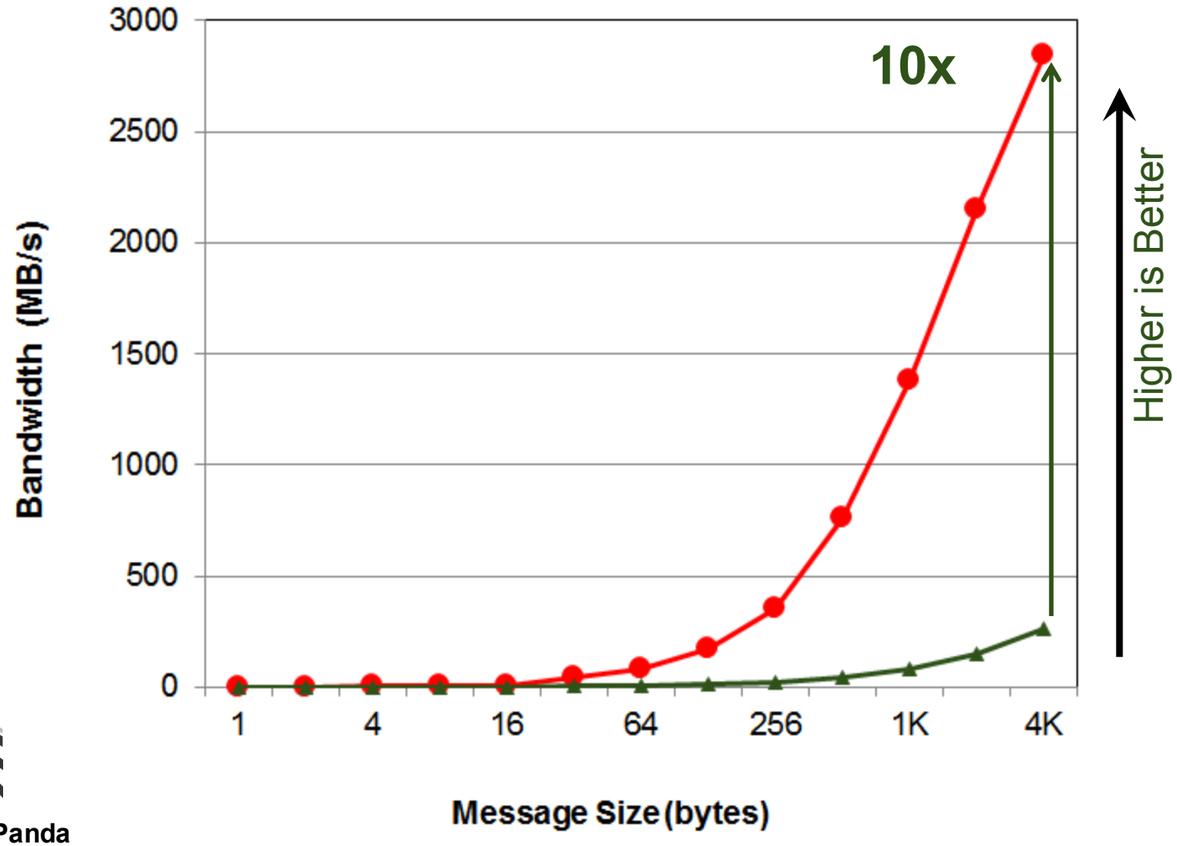
Performance of MVAPICH2 with GPUDirect RDMA

GPU-GPU Internode MPI Latency



Source: Prof. DK Panda

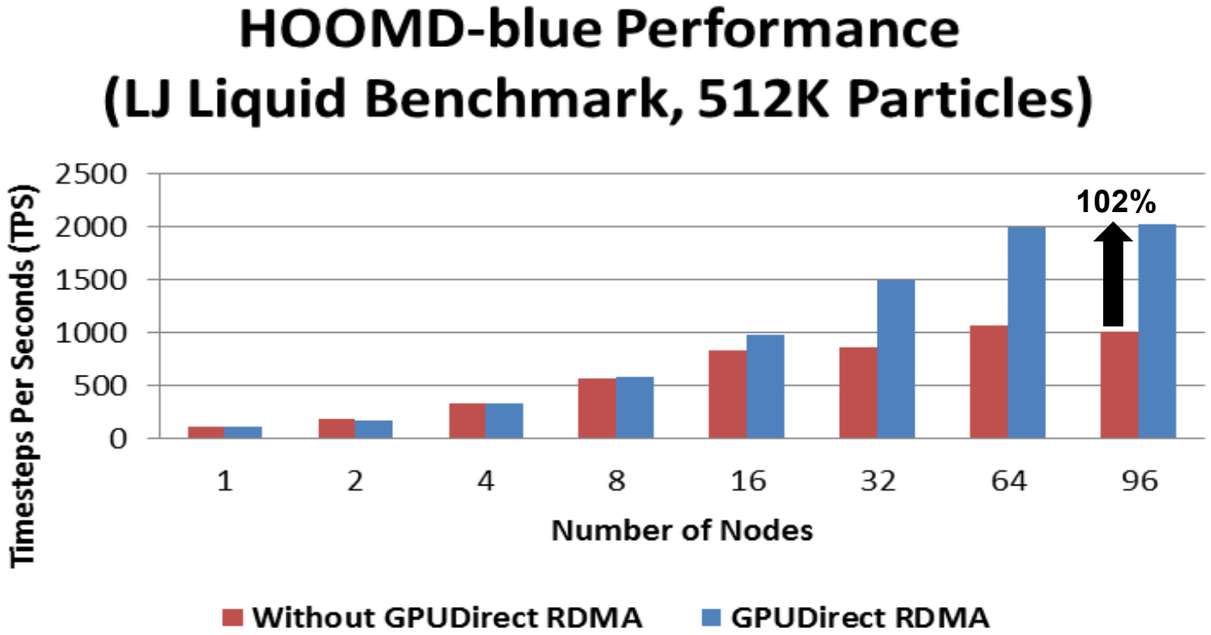
GPU-GPU Internode MPI Bandwidth

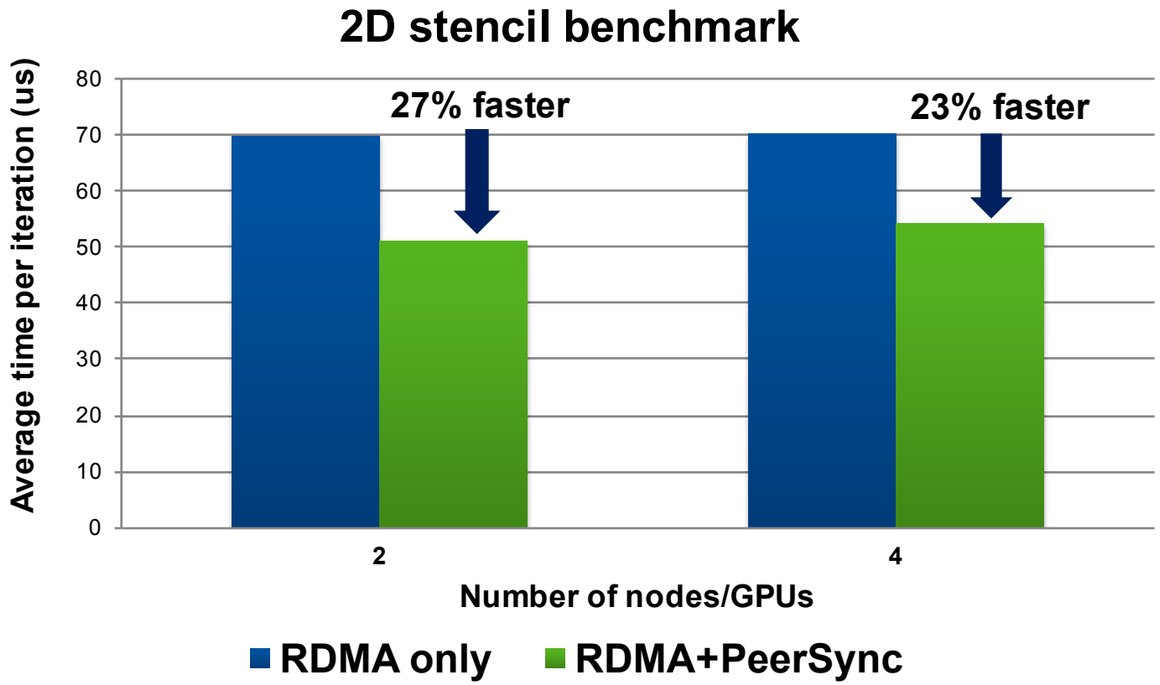


88% Lower Latency

10X Increase in Throughput

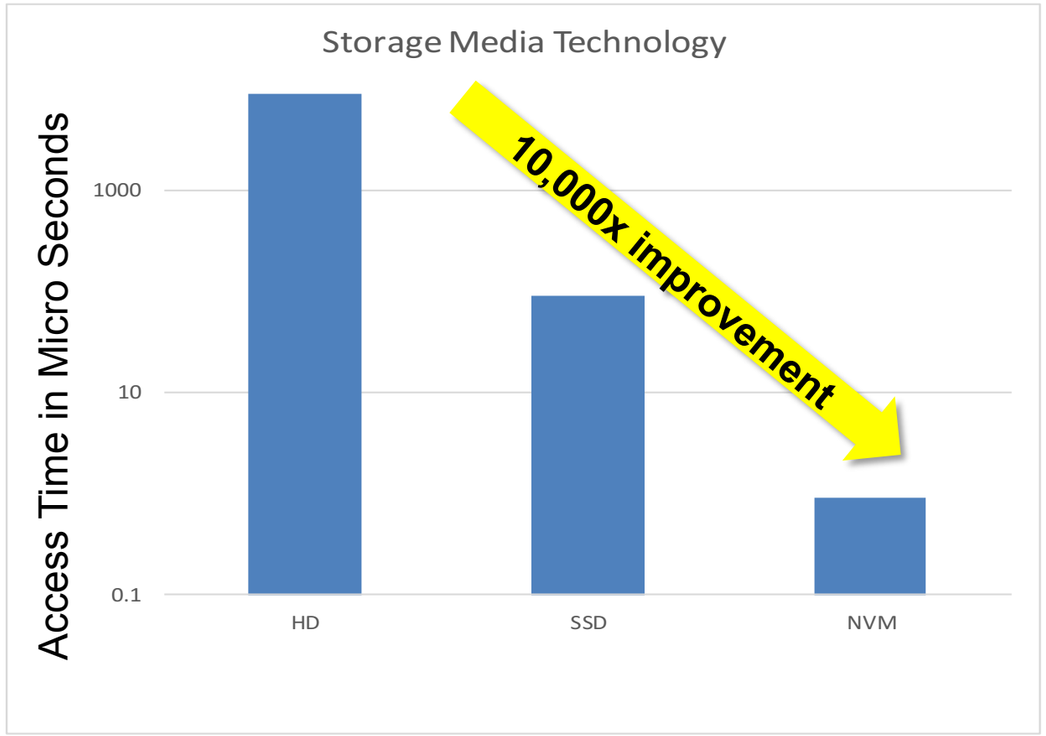
Mellanox PeerDirect™ with NVIDIA GPUDirect RDMA

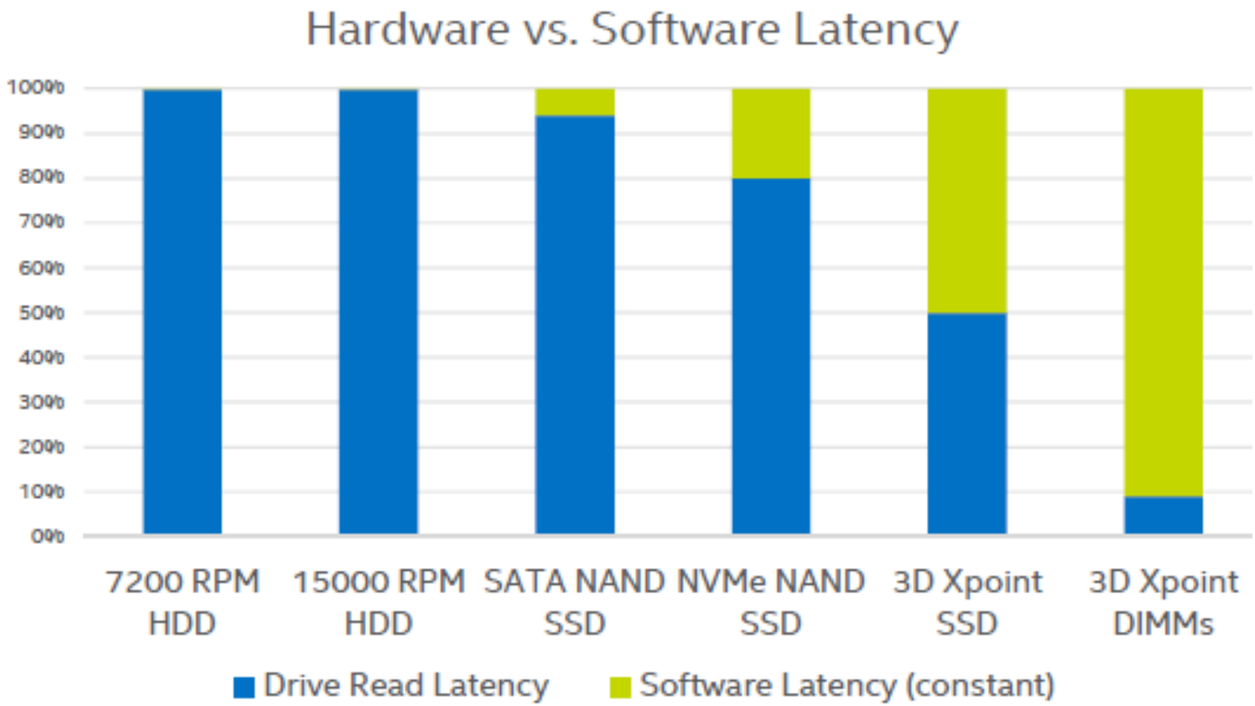




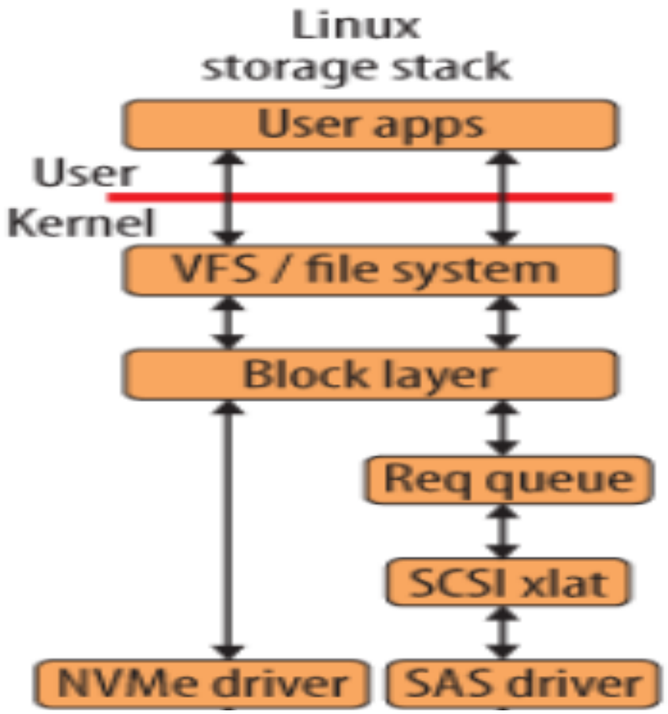
Storage and Data Access

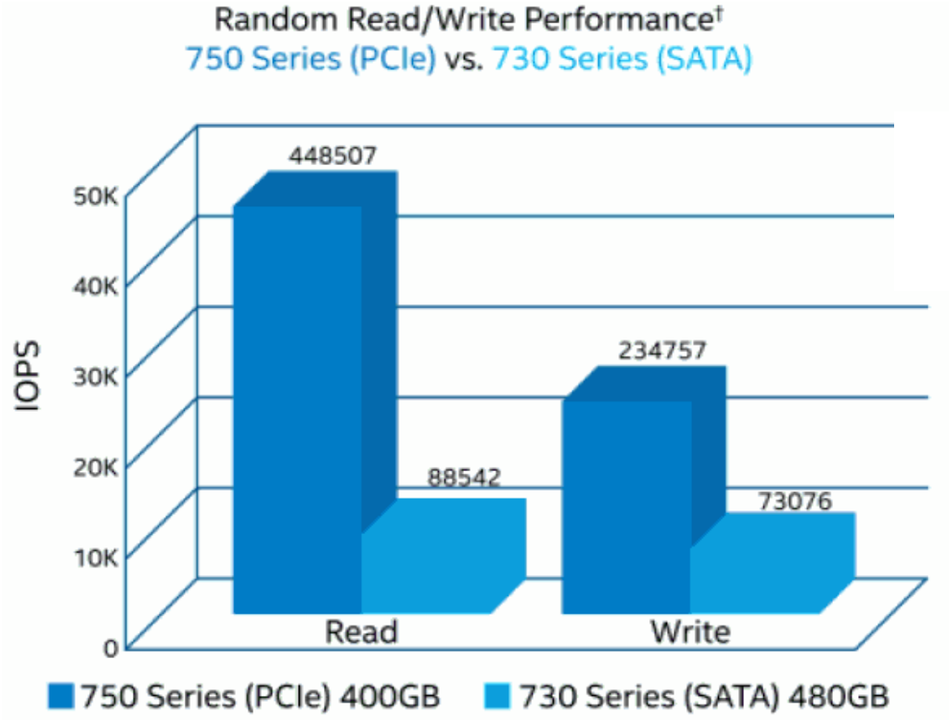
Storage Technology Evolution





NVMe Technology



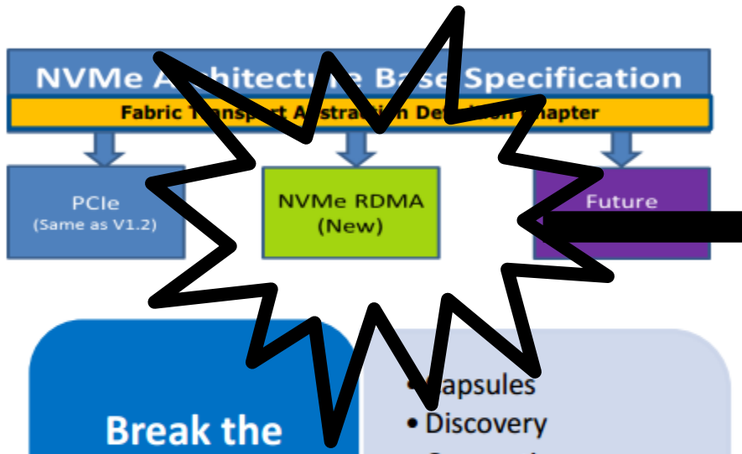


NVME over Fabrics – RDMA-based networking storage

Specification Strategy and Breakdown of Work

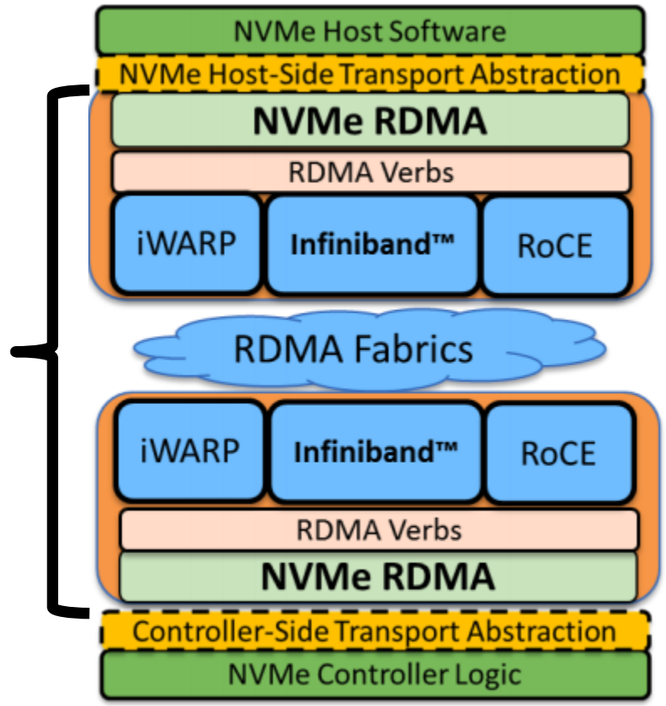
Do not create a standalone specification

- Initial goal is to minimize changes to existing specification
- Cleanly separate out the non-PCIe NVMe Transport layers through separate chapters/sections
  - Fabrics Core (concepts and RDMA binding)
  - Fabrics Base Differences (SGL changes, etc.)
- Long-term goal is to create a Transport agnostic base spec

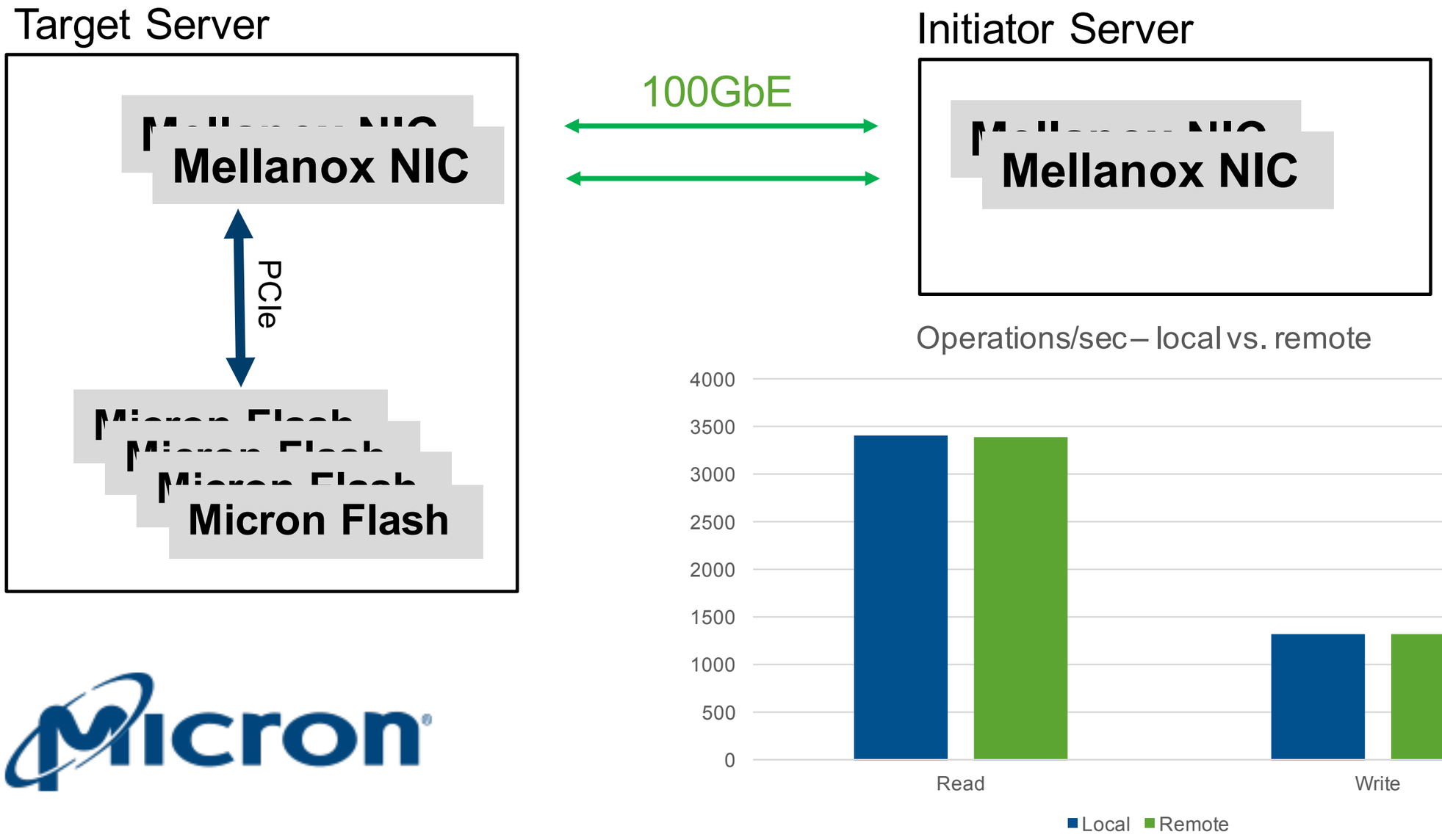


Break the work into functional sub-sections

- Capsules
- Discovery
- Connections
- Flow Control
- Naming
- Binding
- Error Handling



RDMA-based Remote NVME Access (NVME over Fabrics)



Offload versus Onload

Claims Used to “Market” Onload Architecture – “Too Many Cores”

- Claim: There are many CPU cores, and the applications cannot use them all, so one can dedicate some cores to manage the interconnect operations

- Reality: False claim
  - CPU vendors increase the CPU core count due to applications requirements!
  - In cases where applications require less core, data center owners can buy the needed core count
  - CPUs with less core are dramatically cheaper! Why would one spend more $ if not needed?

Intel Haswell CPU 10-Cores	Intel Haswell CPU 12-Cores	Intel Haswell CPU 14-Cores	Conclusions
CPU cost: $1502	CPU cost: $2170	CPU cost: $3003	**CPU cores cost more than the interconnect!**
	12-Core to 10-Core Difference: $668	14-Core to 12-Core Difference: $833	
	Dual Socket Server Difference: $1336	Dual Socket Server Difference: $1666	**CPU cores are not free!**

Sandia National Laboratory Paper – Offloading versus Onloading

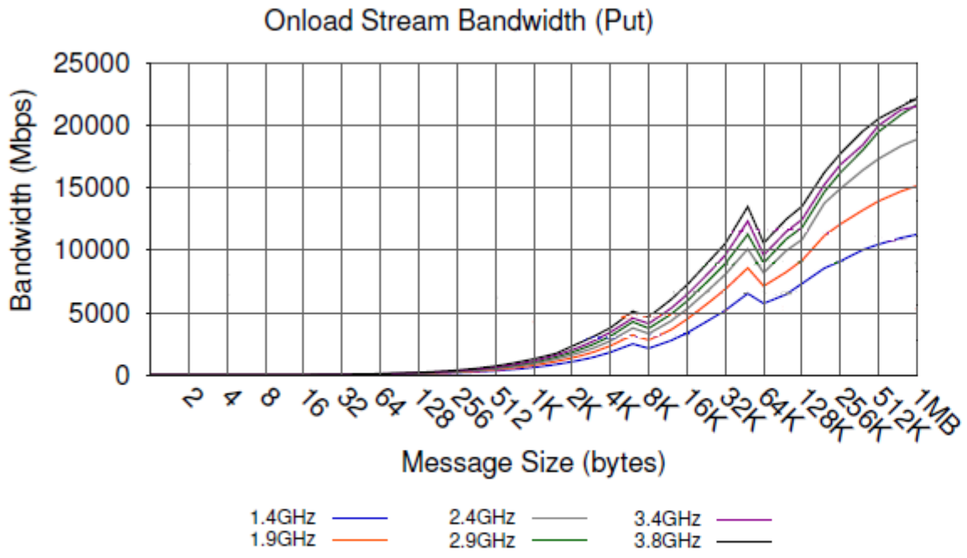
2015 IEEE International Conference on Cluster Computing

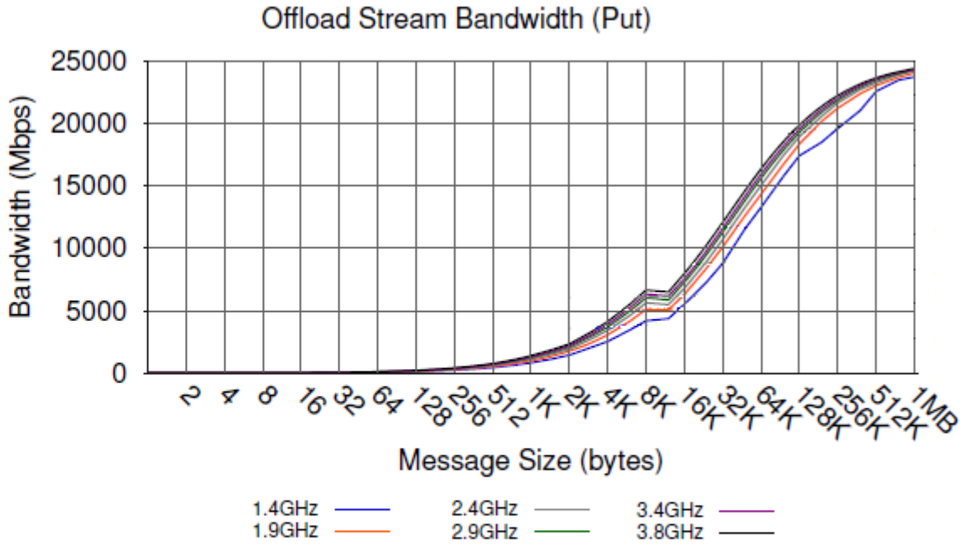
Re-evaluating Network Onload vs. Offload for the Many-Core Era

Matthew G. F. Dosanjh*, Ryan E. Grant†, Patrick G. Bridges* and Ron Brightwell†

*Scalable Systems Laboratory
Department of Computer Science
University of New Mexico

†Center for Computing Research
Sandia National Laboratories*





Onload vs. offloaded with varying CPU frequencies

Interconnect Throughput – Offload versus Onload

Network Performance Dramatically Depends on CPU Frequency!

Common Xeon Frequency 2.6GHz

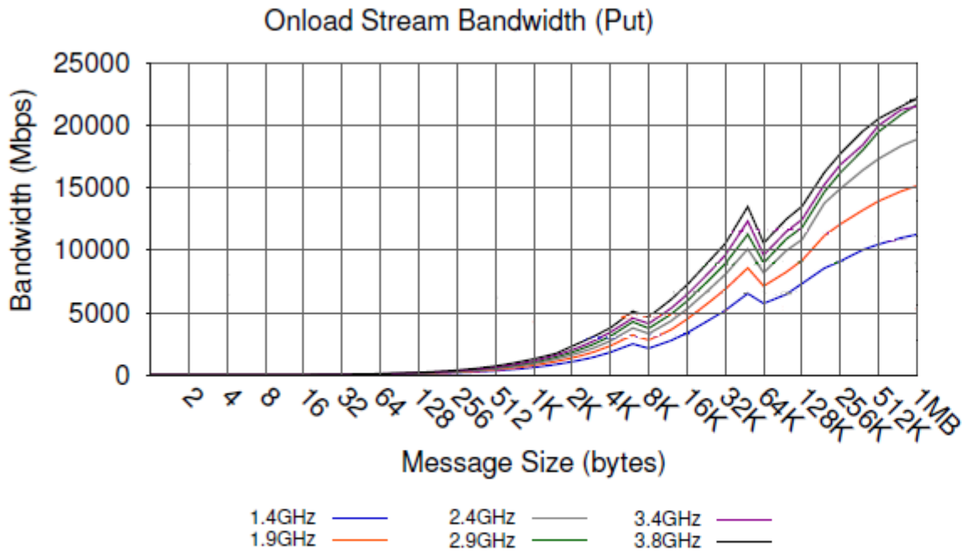
Common Xeon Phi Frequency ~1Ghz

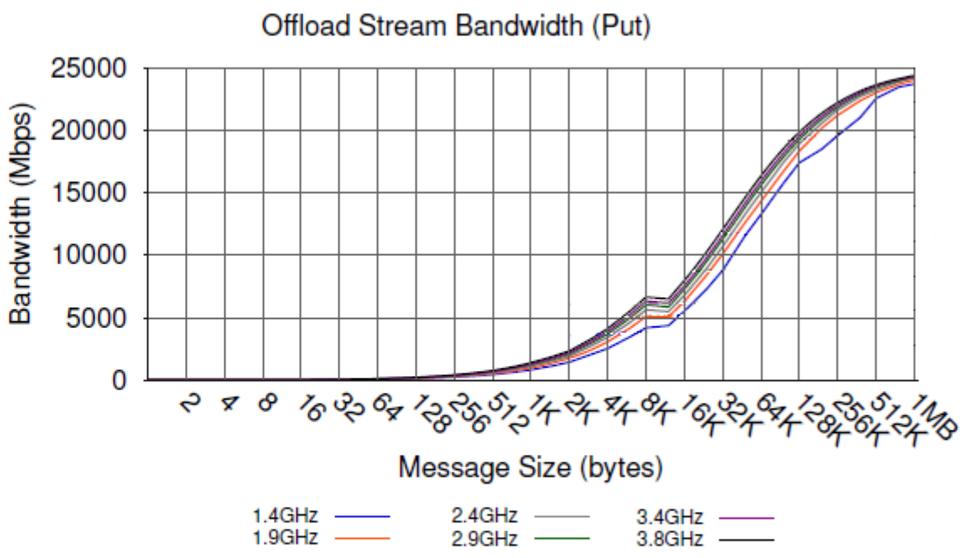
The Offloading Advantage!

Data Throughput:

20% Higher at common Xeon Frequency

250% Higher at common Xeon Phi Frequency

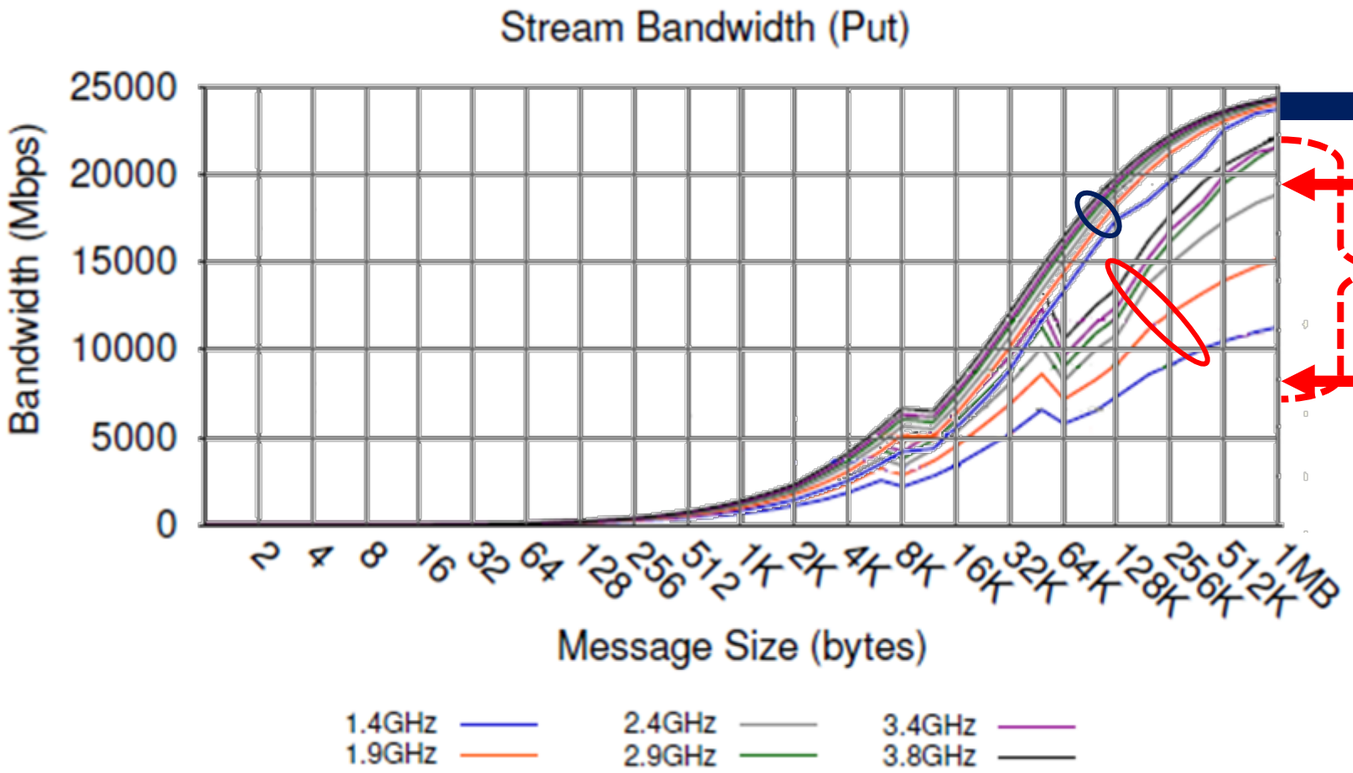




Onload vs. offloaded with varying CPU frequencies

Only Offload Architecture Can Enable Co-Processors

Onloading Technology Not Suitable for Co-Processors!



Offloading (Highest Performance for all Frequencies)

Common Xeon Frequency

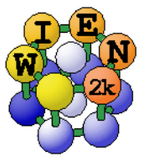
Onloading (performance loss with lower CPU frequency)

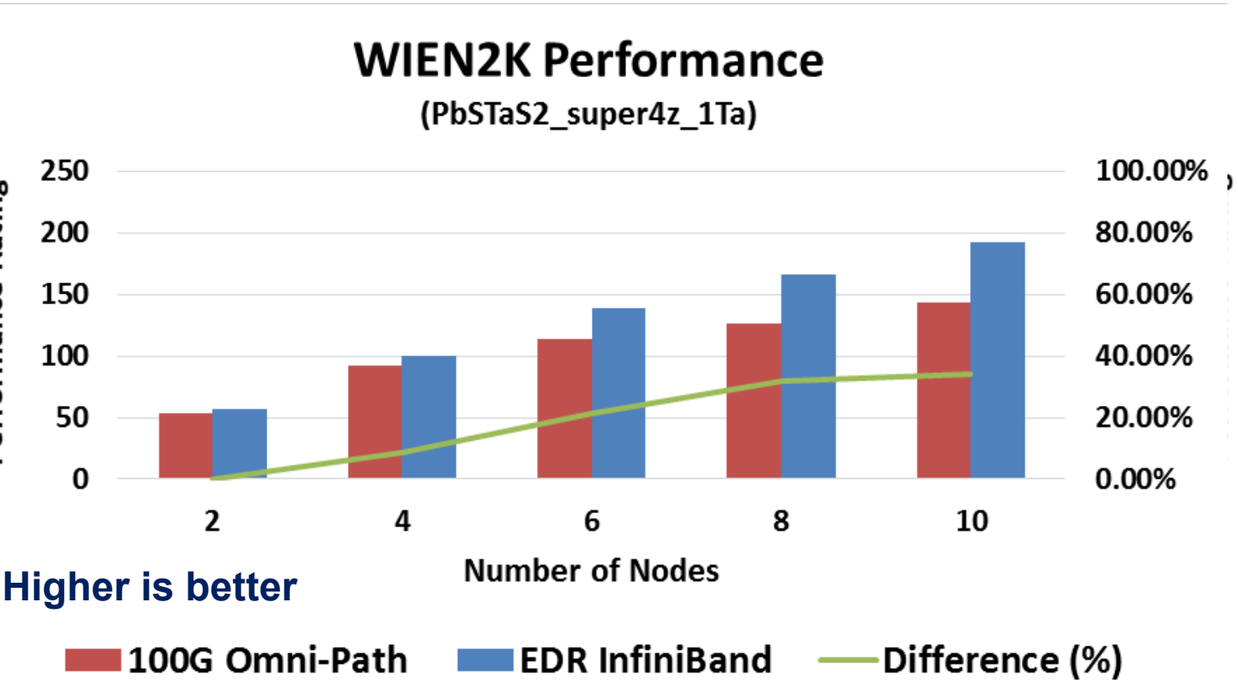
Common Xeon Phi Frequency

Onload vs. offloaded with varying CPU frequencies

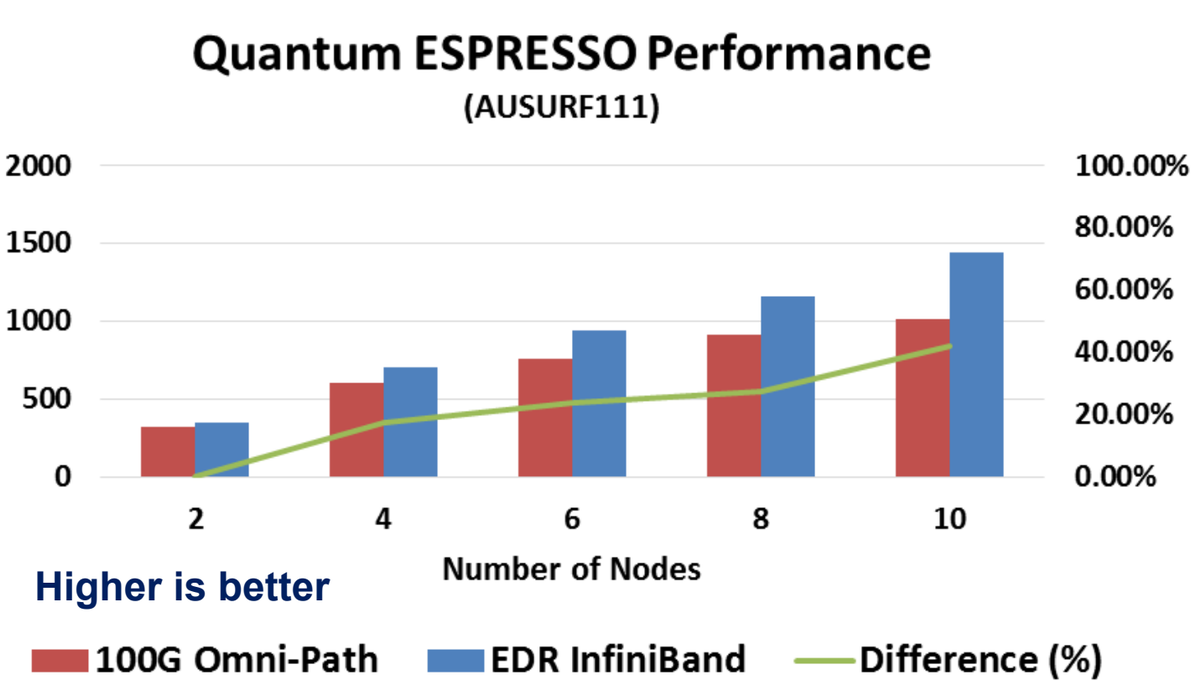
Application Performance Comparison – Quantum ESPRESSO

WIEN2k is a Quantum Mechanical Simulation





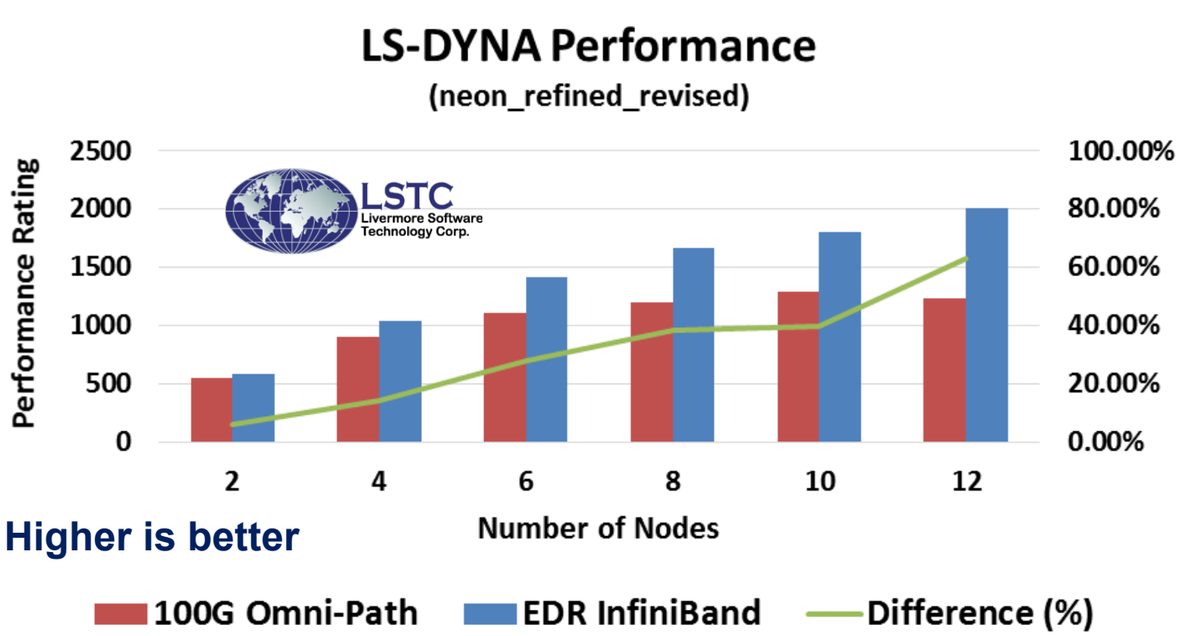
Quantum ESPRESSO is an electronic structure and materials modeling Simulation

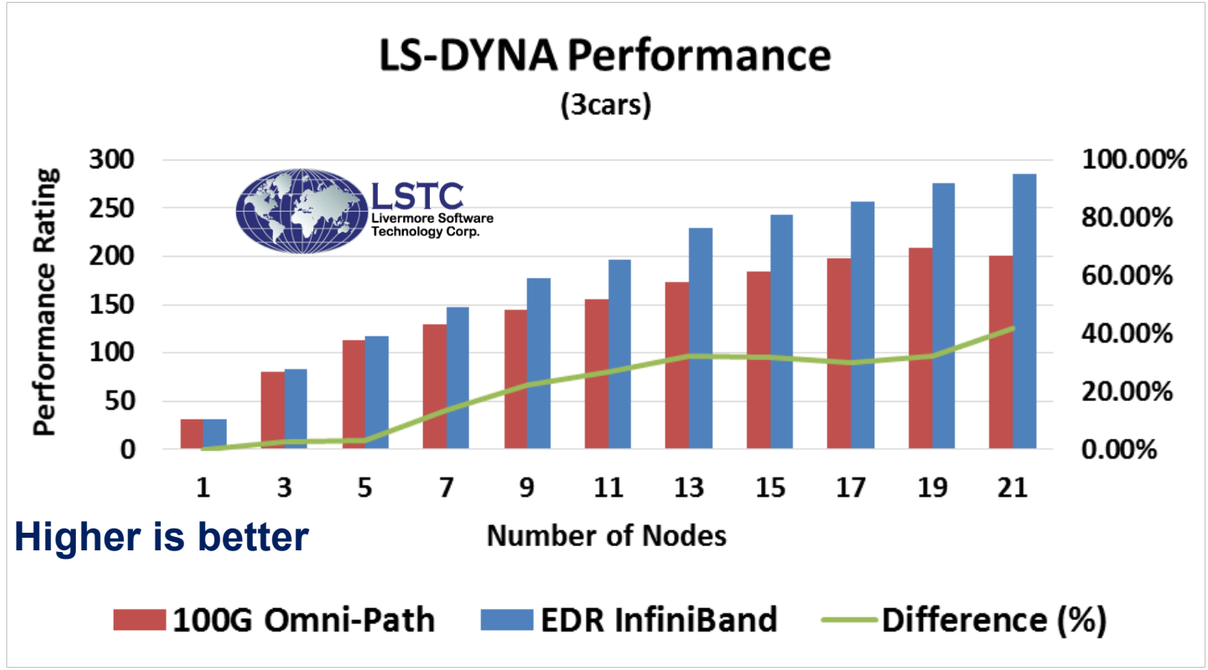


InfiniBand Delivers Higher Performance and Scaling

Application Performance Comparison – LS-DYNA

A structural and fluid analysis software, used for automotive, aerospace, manufacturing simulations and more



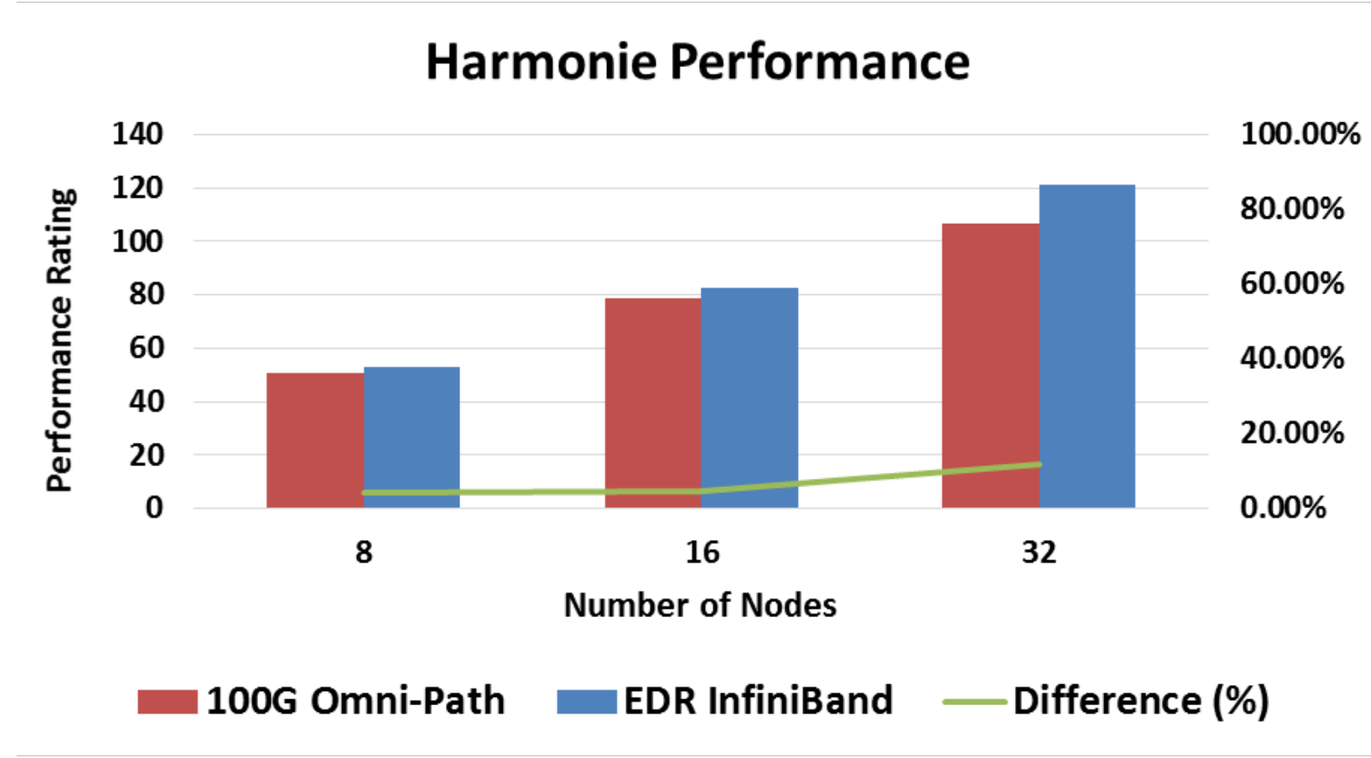


InfiniBand Delivers 42-63% Higher Performance With Only 12 Nodes

Omni-Path Does Not Scale Beyond 10 Nodes

Application Performance Comparison – HARMONIE

HARMONIE (HiRLAM Aladin Regional Mesoscale Operational NWP In Europe) is numerical weather prediction consortium which develops the HARMONIE application for short range weather forecasting

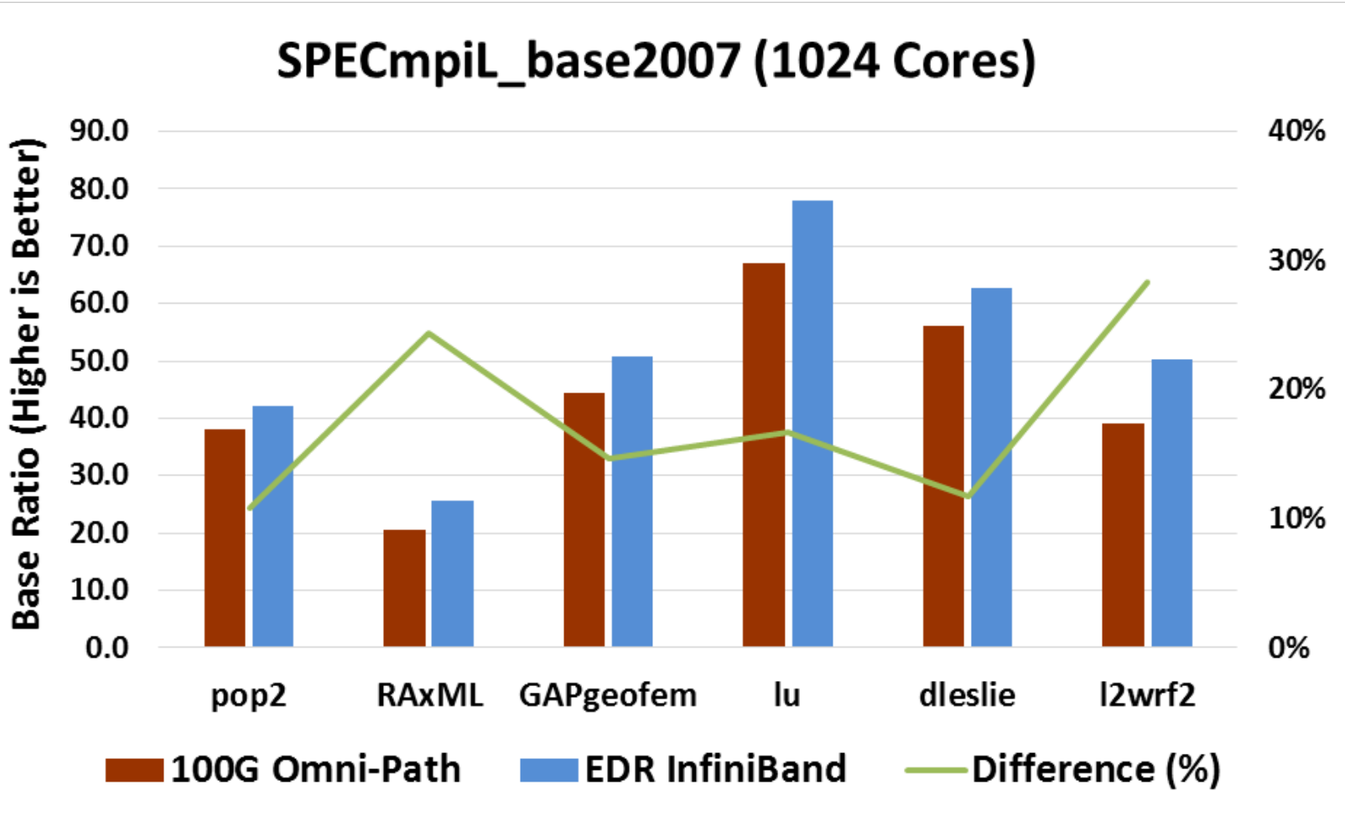


InfiniBand Delivers Higher Performance and Scaling

Application Performance Comparison – SPEC MPI Benchmark Suite

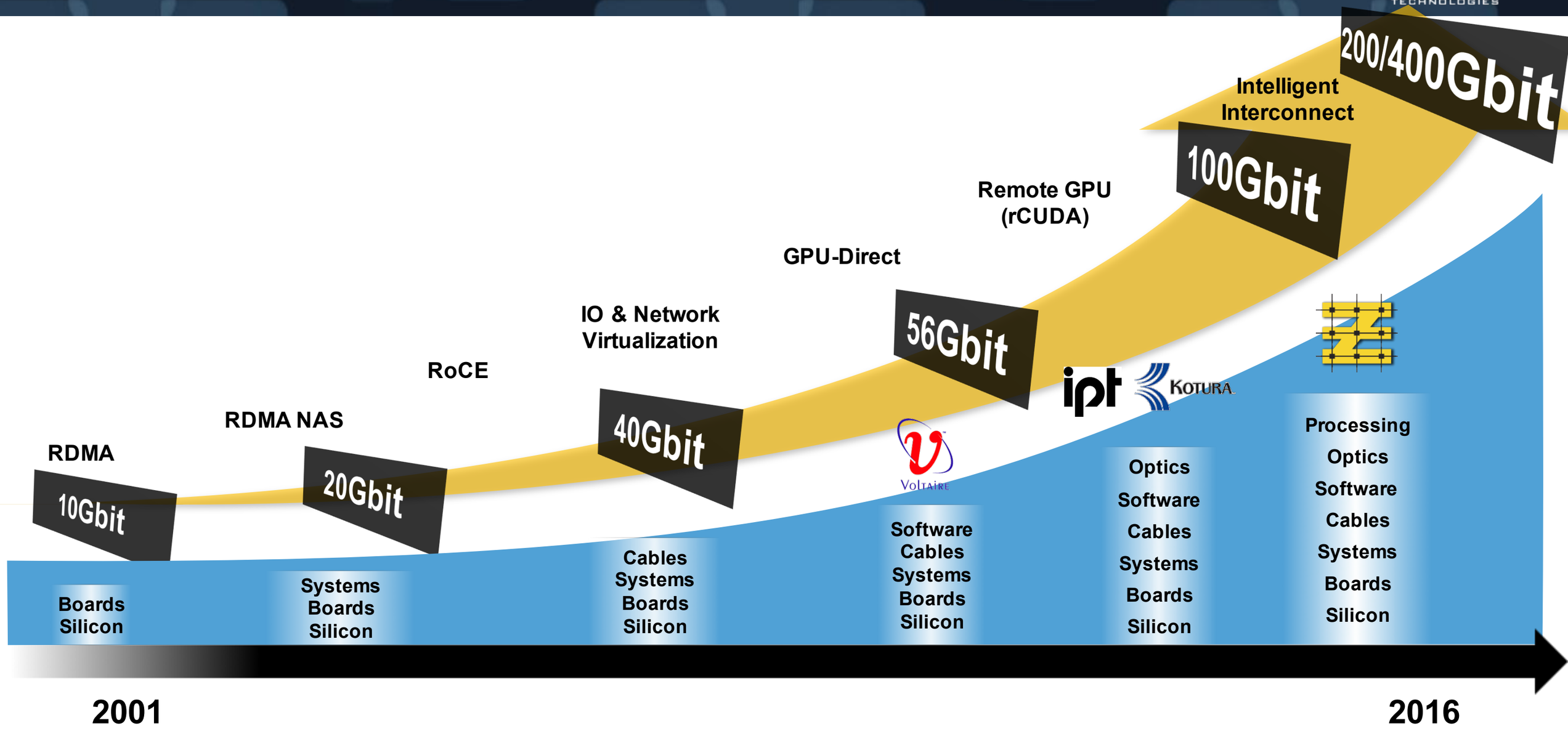
The SPEC MPI benchmark suite evaluates MPI-parallel, floating point, compute intensive performance, across a wide range of compute intensive applications using the Message-Passing Interface (MPI)

The Standard Performance Evaluation Corporation (SPEC) is a non-profit corporation formed to establish, maintain and endorse a standardized set of relevant benchmarks that can be applied to the newest generation of high-performance computers

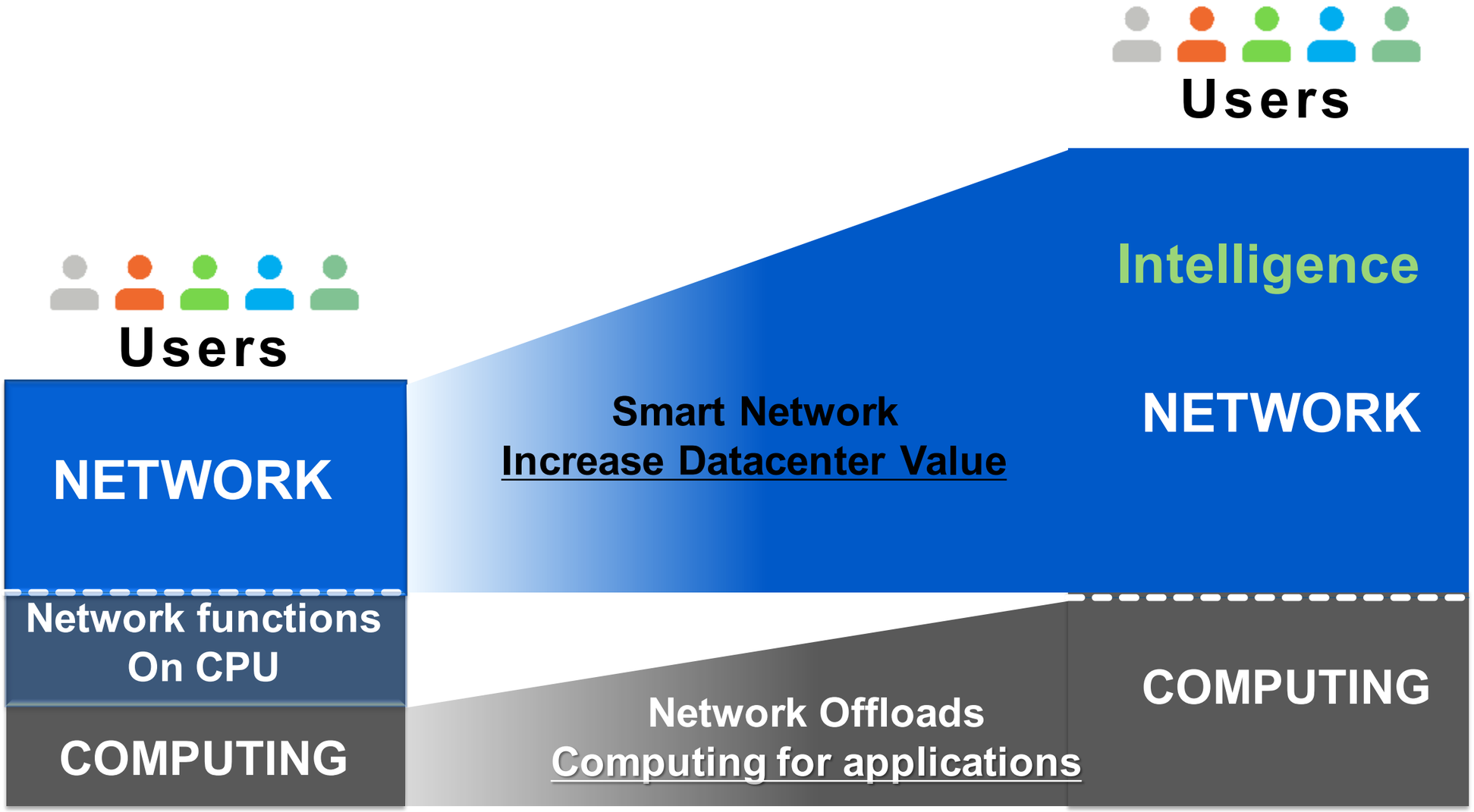


InfiniBand Delivers Superior Performance and Scaling

Technology Leadership



Strategy: More Value to Customer



Summary

- To overcome the performance limitations of today's HPC systems we need an intelligent interconnect
- The interconnect becomes a co-processor - delivering in-network computing
  - Enabling data analysis everywhere, offloading the CPU, increasing data center efficiency
- Mellanox InfiniBand delivers leading performance over Omni-Path promises
  - 68% higher message rate, 20% lower latency, 25% lower power consumption
- InfiniBand enables higher applications performance with Lower CPAR ($/performance)
  - On average 45% higher application performance, at 27% lower cost per application
  - Mellanox EDR solution is robust, and delivering scalable performance
- Other technologies lack RDMA or any offloading capabilities
  - The same technology from Pathscale and QLogic that failed

Protect Your Future with Mellanox InfiniBand

Thank You

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™