

Современные сети для HPC и ML/DL

Сентябрь 2019





SUPERCONNECTING the #1 Supercomputers



1 TOP 500
The List.



2 TOP 500
The List.



3 TOP 500
The List.



5 TOP 500
The List.



8 TOP 500
The List.



10 TOP 500
The List.

InfiniBand Accelerates 6 of Top 10 Supercomputers

HDR 200G InfiniBand Wins Next Generation Supercomputers



23.5 Petaflops
8K HDR InfiniBand Nodes
Fat-Tree Topology



50 Petaflops
7.2K HDR InfiniBand Nodes
Dragonfly+ Topology



Australian National University

3K HDR InfiniBand Nodes
Dragonfly+ Topology



3.1 Petaflops
1.8K HDR InfiniBand Nodes
Fat-Tree Topology



FINNISH METEOROLOGICAL INSTITUTE



1.7 Petaflops
2K HDR InfiniBand Nodes
Dragonfly+ Topology

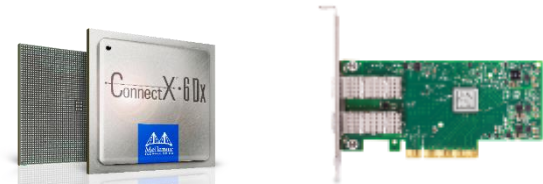


Highest Performance Cloud

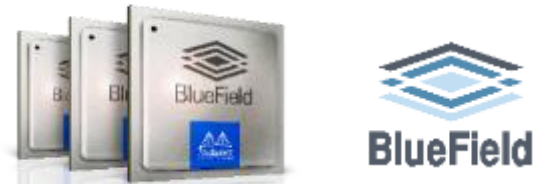


1.6 Petaflops
Hybrid CPU-GPU-FPGA
Fat-Tree Topology

HPC and AI Needs the Most Intelligent Interconnect



SmartNIC



System on a Chip

Higher

Data Speeds

Faster

Data Processing

Better

Data Security



Adapters



Switches



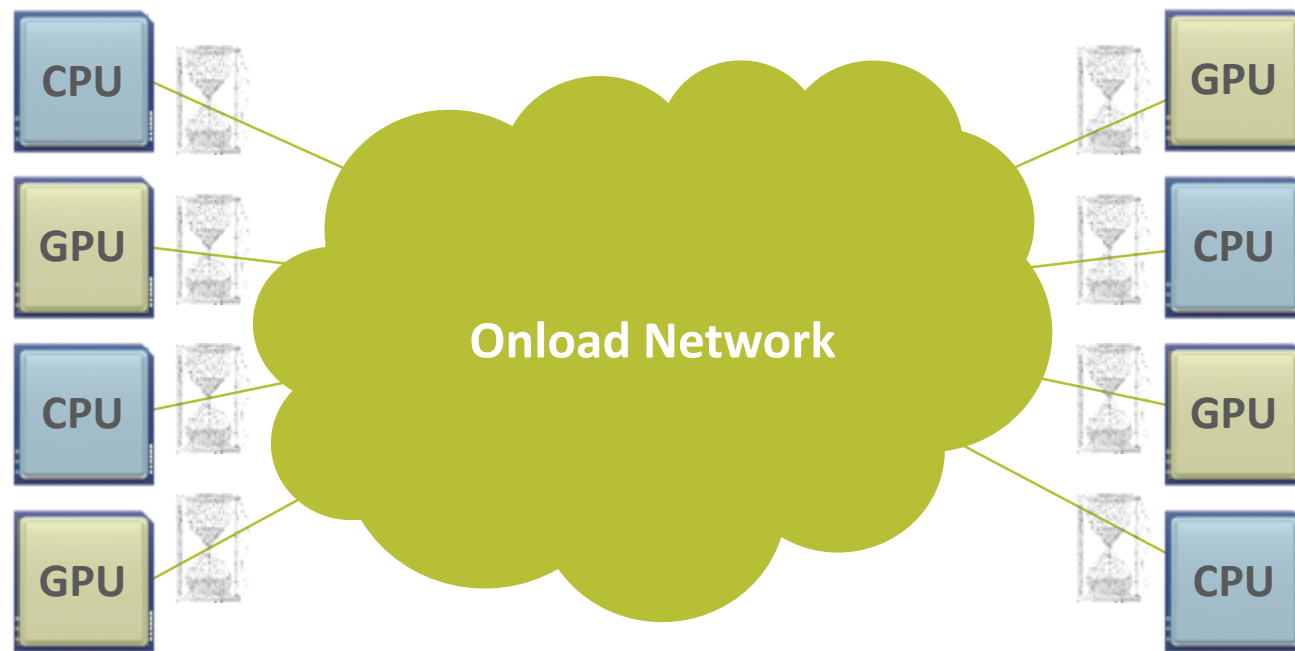
Cables & Transceivers



The Need for Intelligent and Faster Interconnect

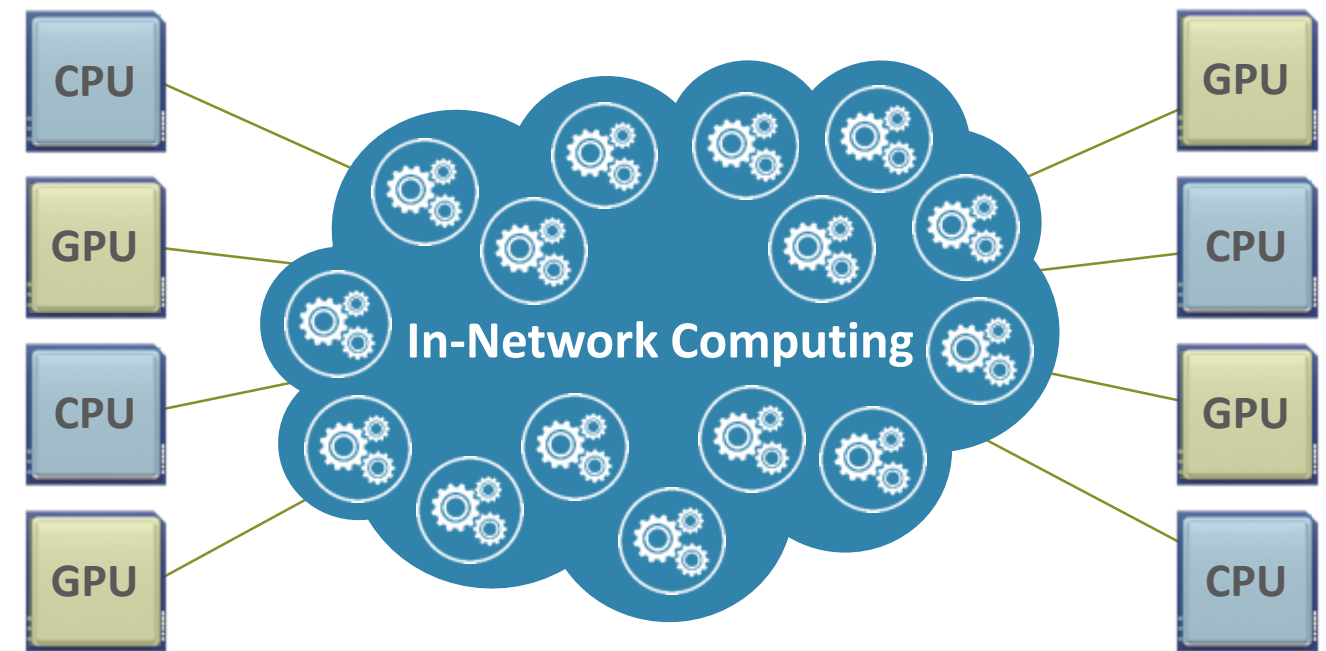
Faster Data Speeds and In-Network Computing
Enable Higher Performance and Scale

CPU-Centric (Onload)

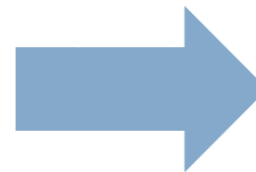


Must Wait for the Data
Creates Performance Bottlenecks

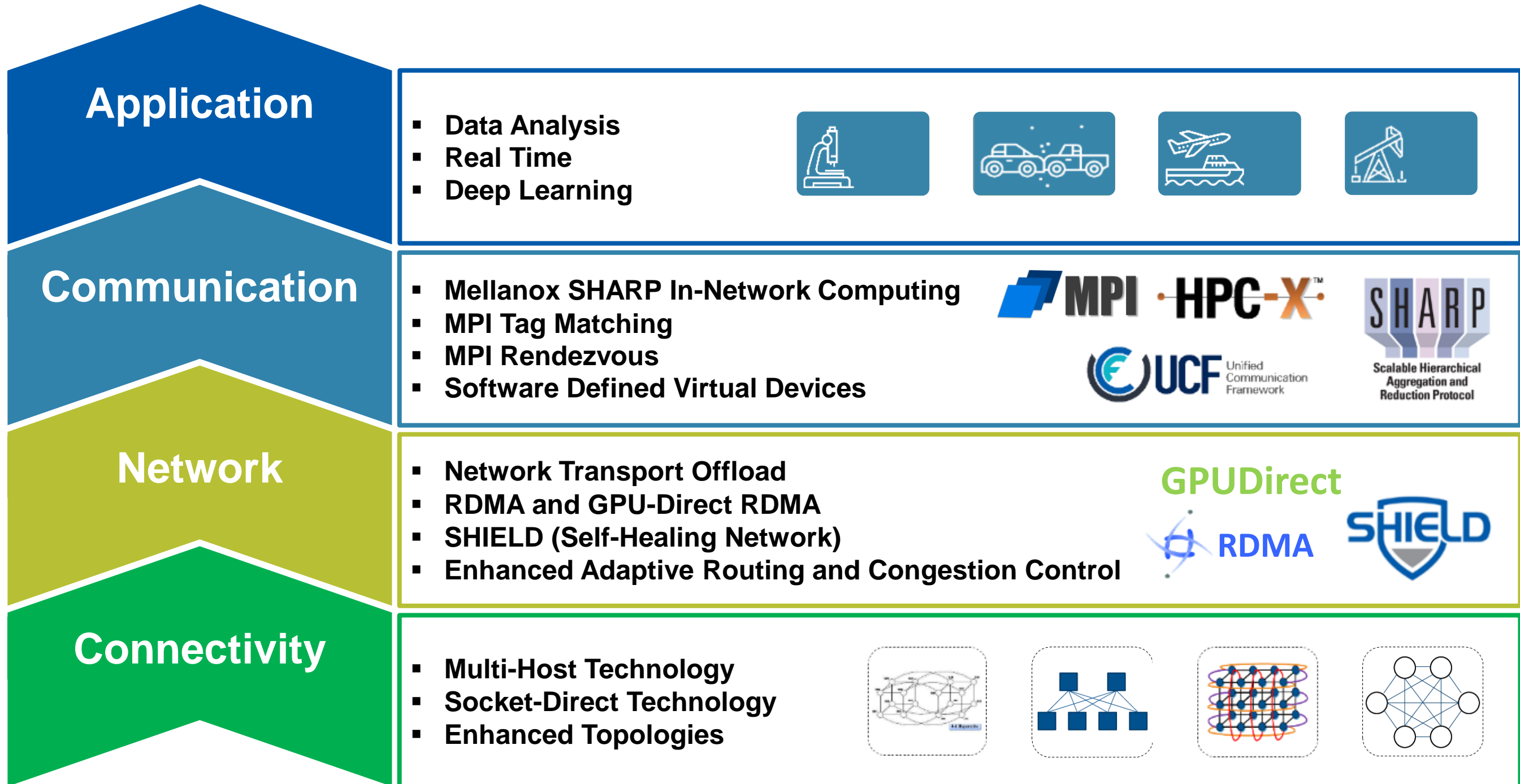
Data-Centric (Offload)



Analyze Data as it Moves!
Higher Performance and Scale

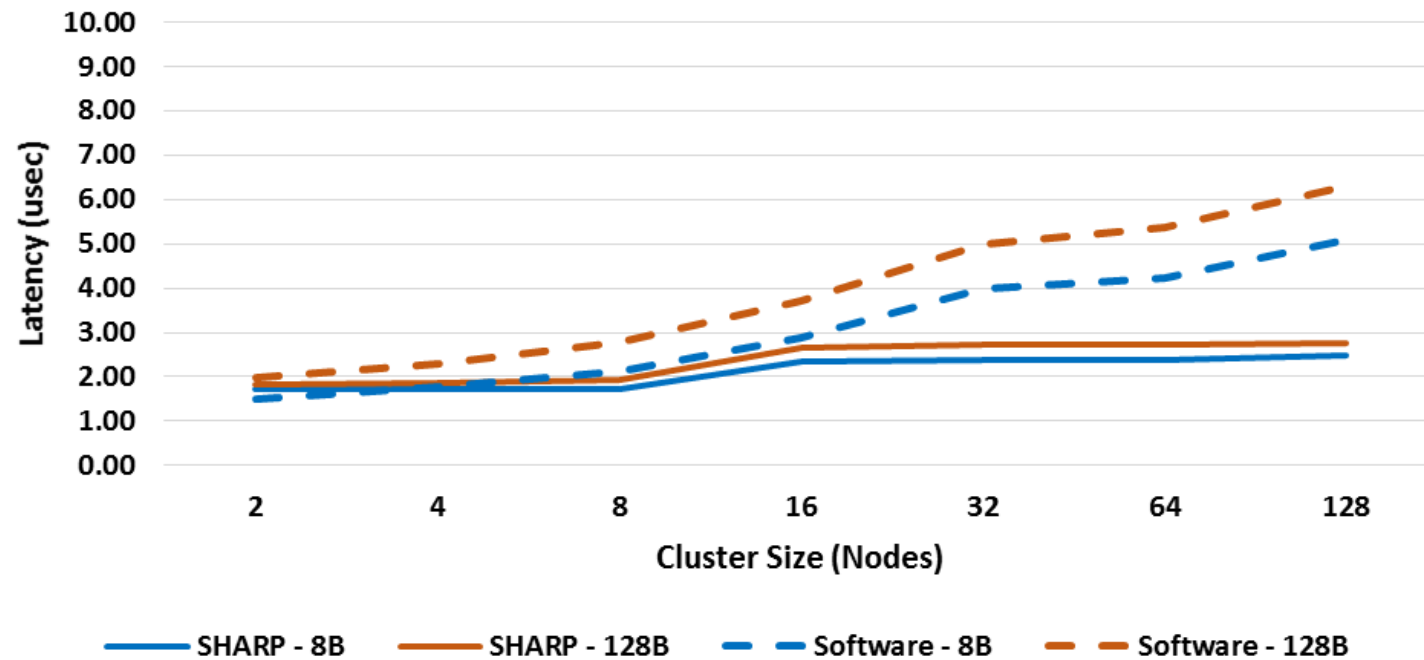


Accelerating All Levels of HPC / AI Frameworks

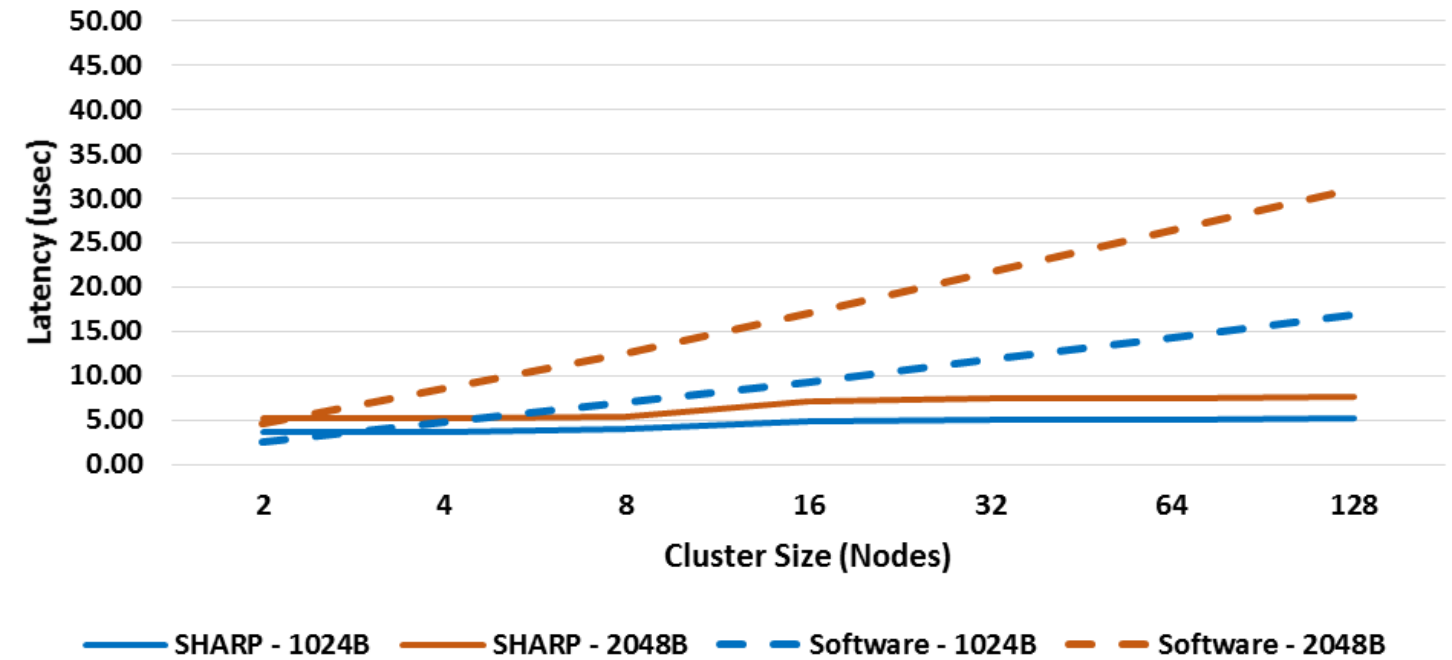


SHARP AllReduce Performance Advantages (128 Nodes)

Allreduce Latency



Allreduce Latency



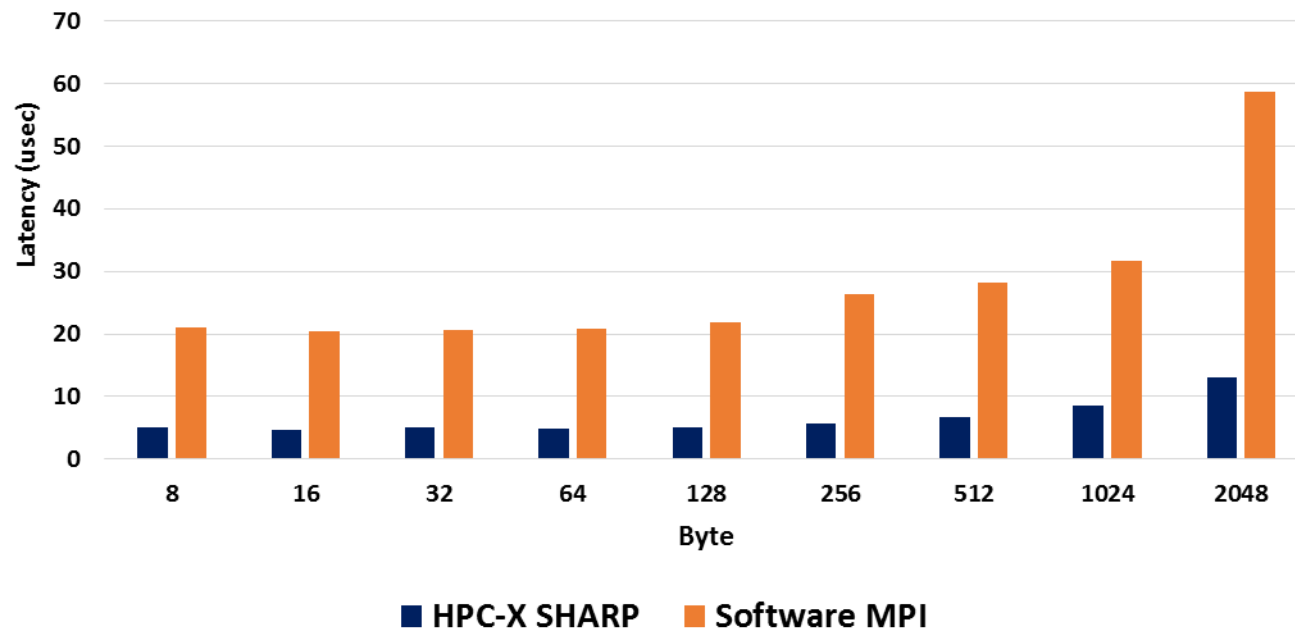
Scalable Hierarchical
Aggregation and
Reduction Protocol

SHARP enables 75% Reduction in Latency
Providing Scalable Flat Latency

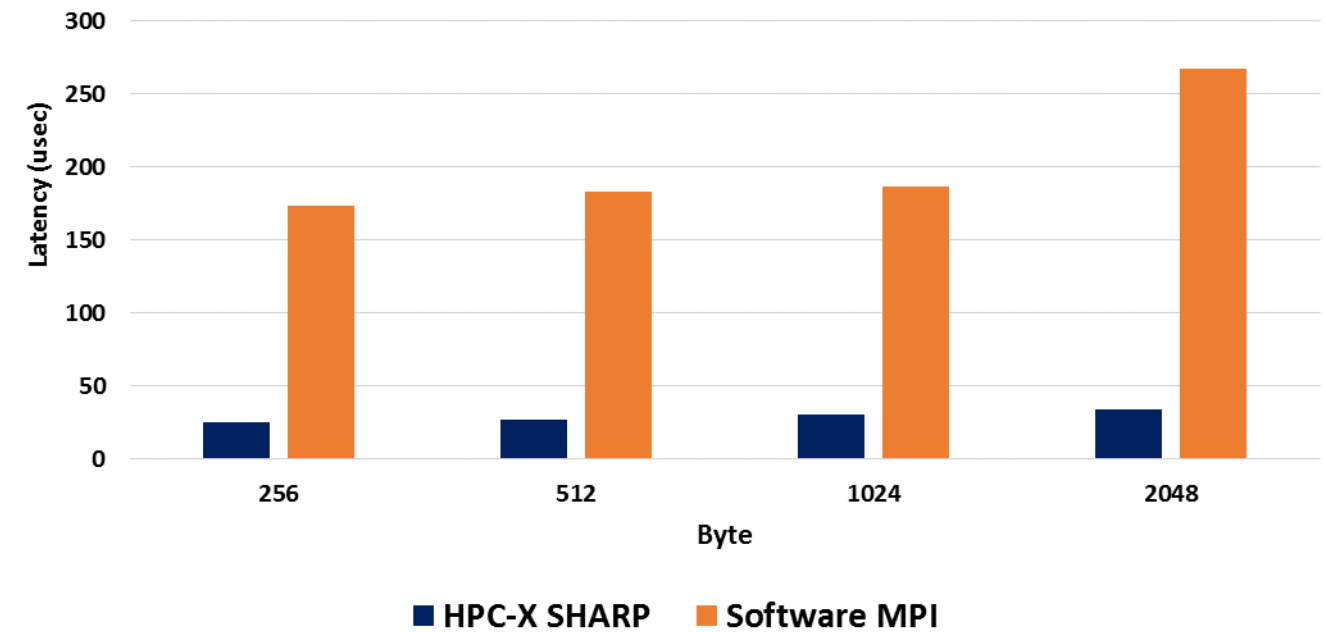
SHARP AllReduce Performance Advantages

1500 Nodes, 60K MPI Ranks, Dragonfly+ Topology

MPI AllReduce Latency
1500 Nodes, 1PPN



MPI AllReduce Latency
1500 Nodes, 40PPN, 60K MPI Ranks

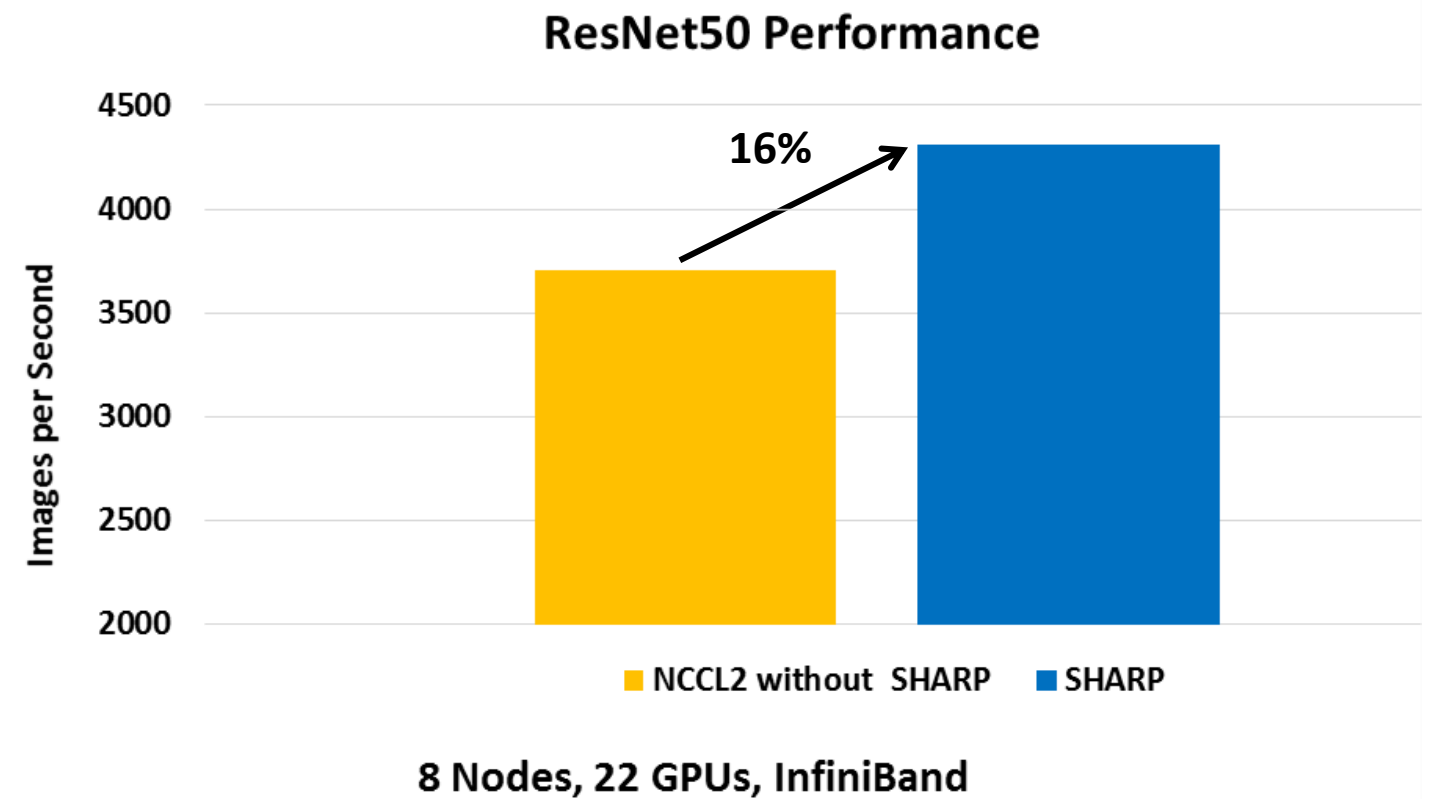
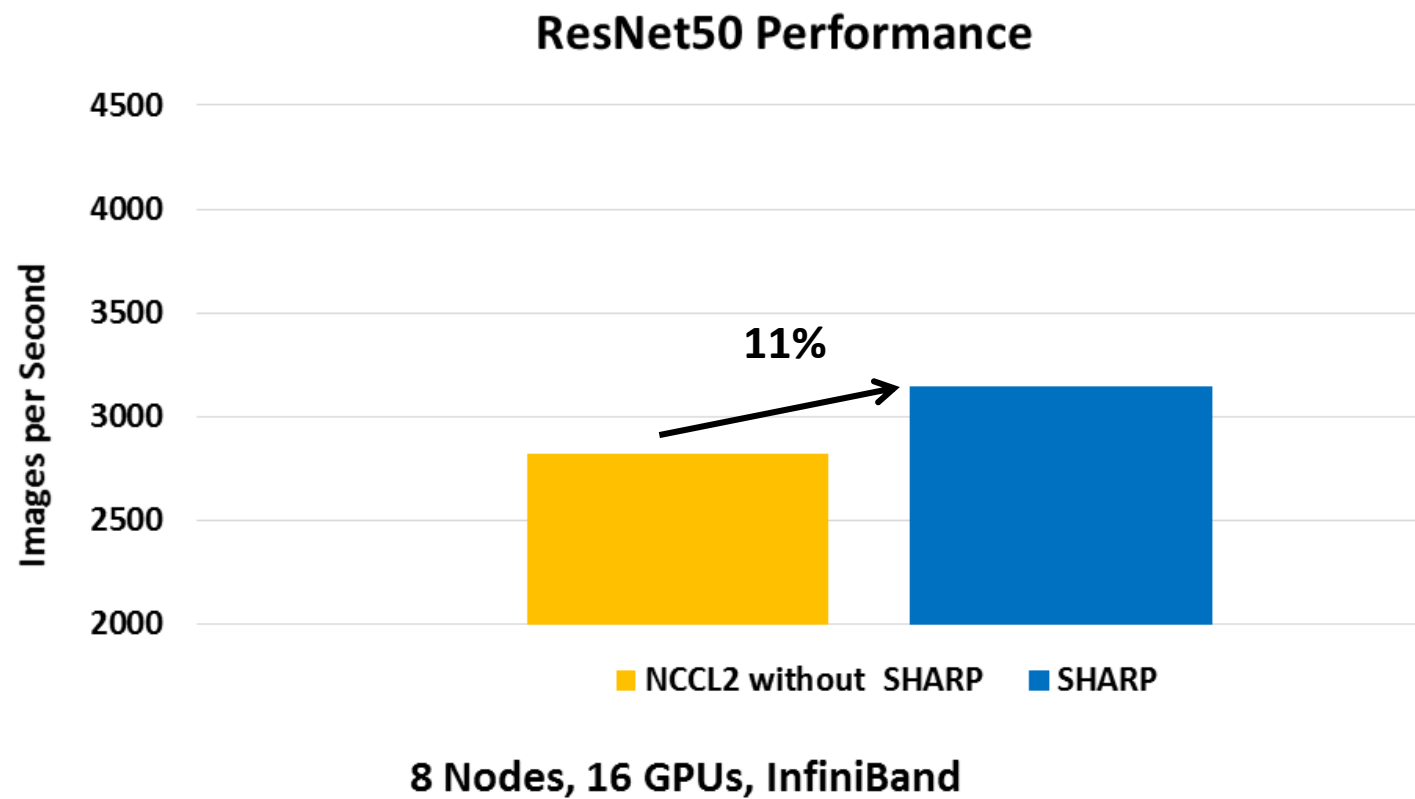


Scalable Hierarchical
Aggregation and
Reduction Protocol

SHARP Enables Highest Performance

SHARP Performance Advantage for AI

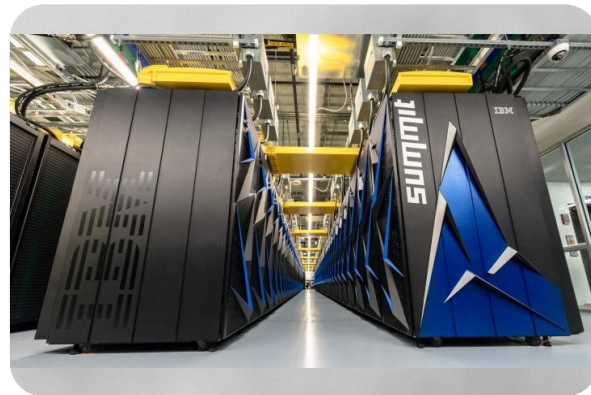
- SHARP provides 16% Performance Increase for deep learning, initial results
- TensorFlow with Horovod running ResNet50 benchmark, HDR InfiniBand (ConnectX-6, Quantum)



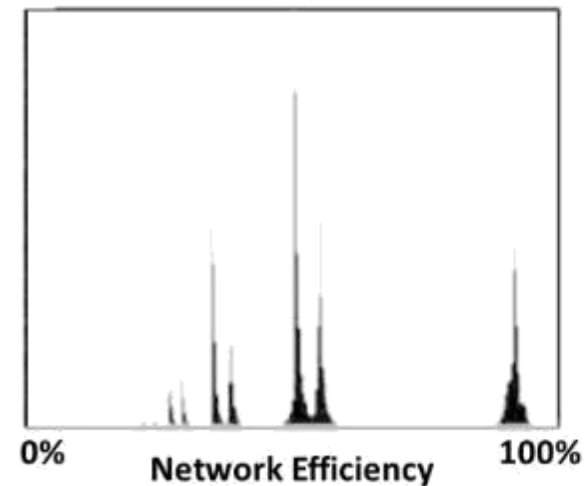
InfiniBand Proven Adaptive Routing Performance

- Oak Ridge National Laboratory – Coral Summit supercomputer
- Bisection bandwidth benchmark, based on mpiGraph
 - Explores the bandwidth between possible MPI process pairs
- AR results demonstrate an average performance of 96% of the maximum bandwidth measured

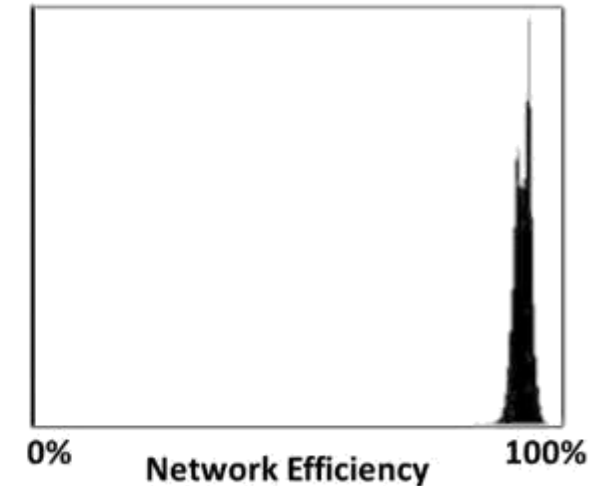
mpiGraph explores the bandwidth between possible MPI process pairs. In the histograms, the single cluster with AR indicates that all pairs achieve nearly maximum bandwidth while single-path static routing has nine clusters as congestion limits bandwidth, negatively impacting overall application performance.



InfiniBand High Network Efficiency - mpiGraph



Static Routing

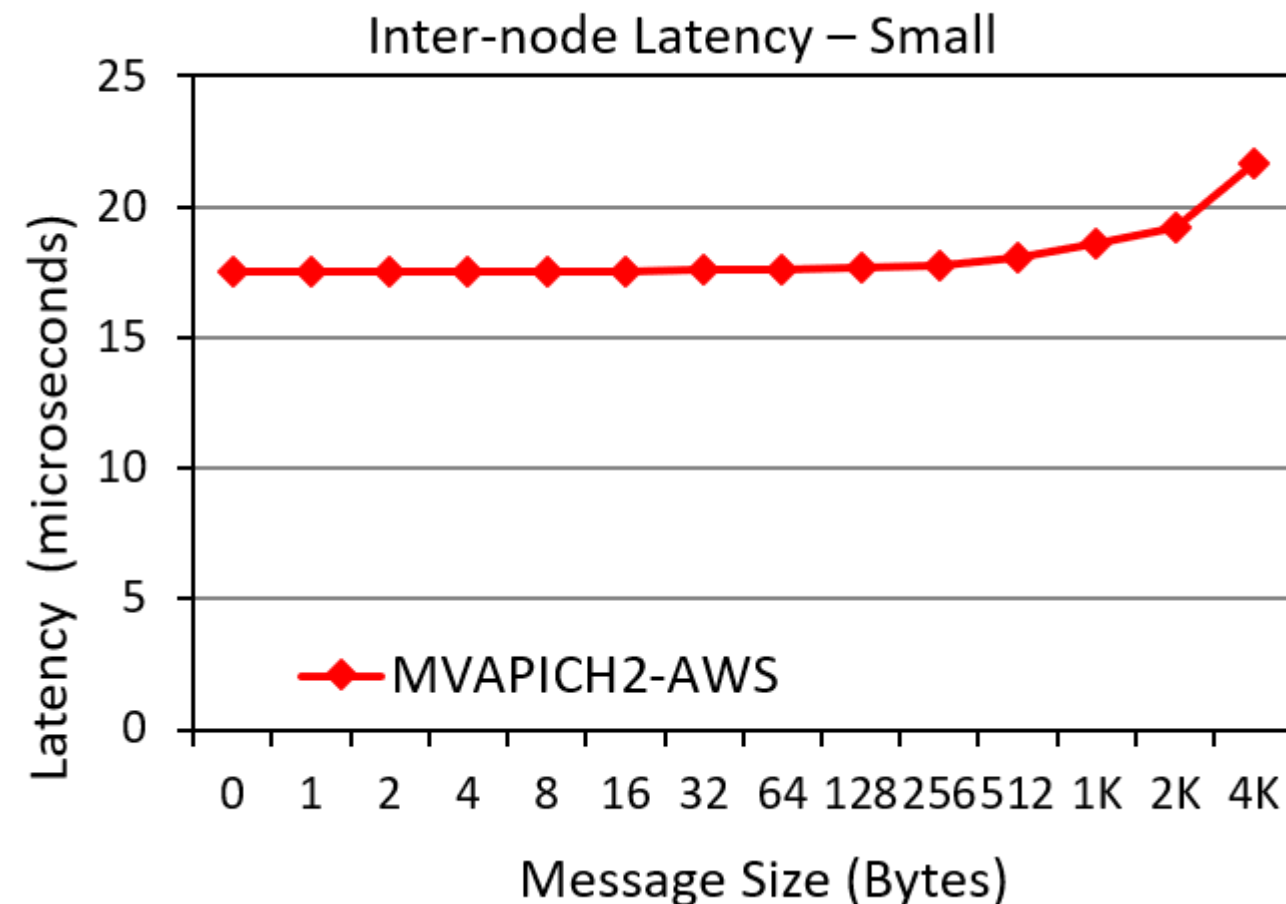
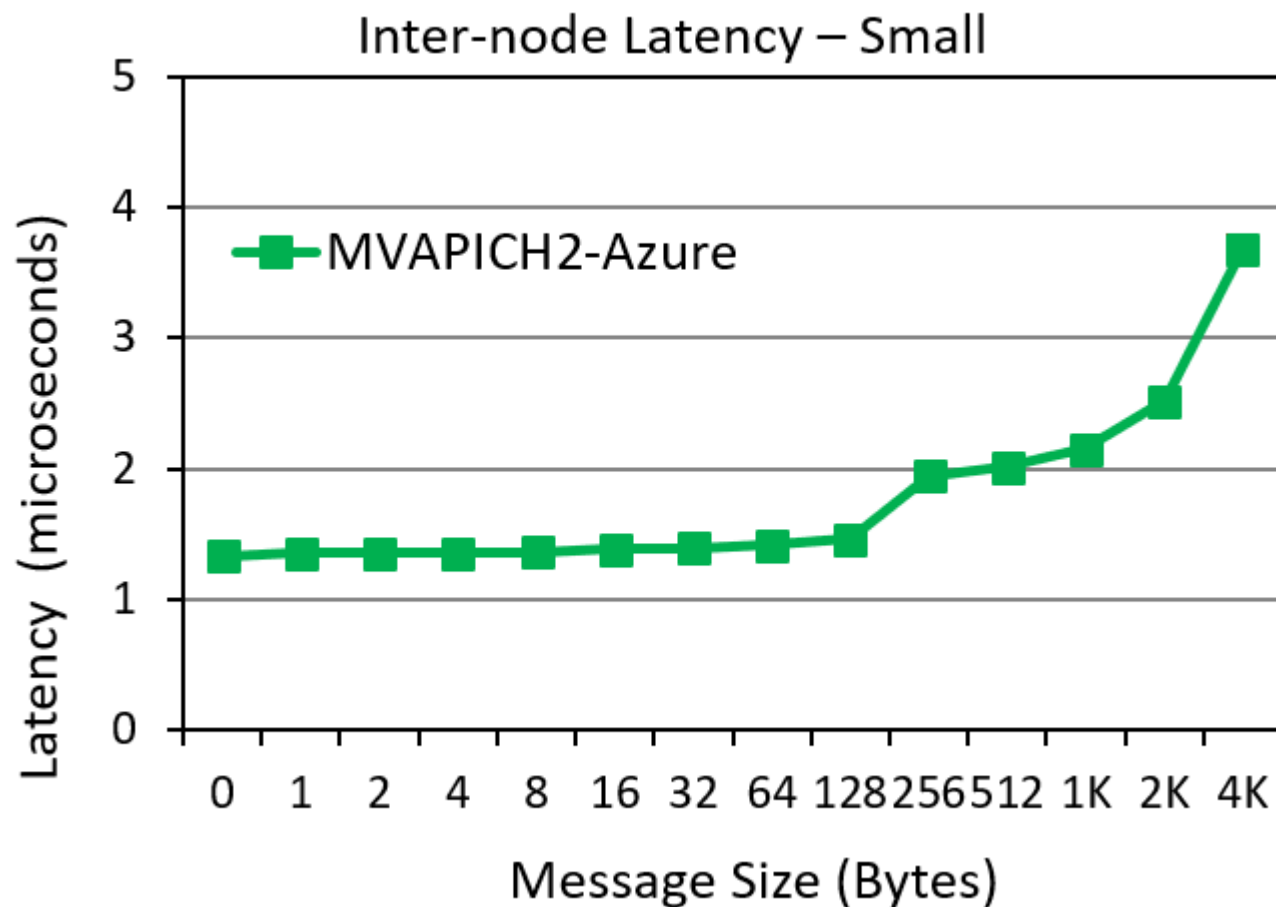


Adaptive Routing

Oak Ridge National Lab Summit Supercomputer

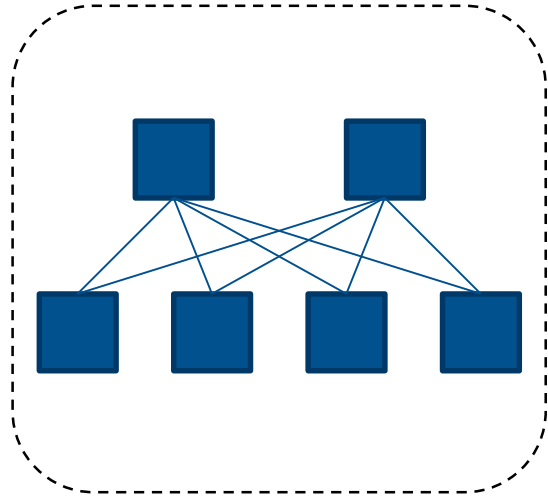
*“The Design, Deployment, and Evaluation of the CORAL Pre-Exascale Systems”,
Sudharshan S. Vazhkudai, Arthur S. Bland, Al Geist, Christopher J. Zimmer, Scott Atchley, Sarp Oral, Don E. Maxwell, Veronica G. Vergara Larrea, Wayne Joubert, Matthew A. Ezell, Dustin Leverman, James H. Rogers, Drew Schmidt, Mallikarjun Shankar, Feiyi Wang, Junqi Yin (Oak Ridge National Laboratory) and Bronis R. de Supinski, Adam Bertsch, Robin Goldstone, Chris Chembreau, Ben Casses, Elsa Gonsiorowski, Ian Karlin, Matthew L. Leininger, Adam Moody, Martin Ohmacht, Ramesh Pankajakshan, Fernando Pizzano, Py Watson, Lance D. Weems (Lawrence Livermore National Laboratory) and James Sexton, Jim Kahle, David Appelhans, Robert Blackmore, George Chochia, Gene Davison, Tom Gooding, Leopold Grinberg, Bill Hanson, Bill Hartner, Chris Marroquin, Bryan Rosenberg, Bob Walkup (IBM)*

Azure (100G InfiniBand) vs AWS (100G) - MPI Performance

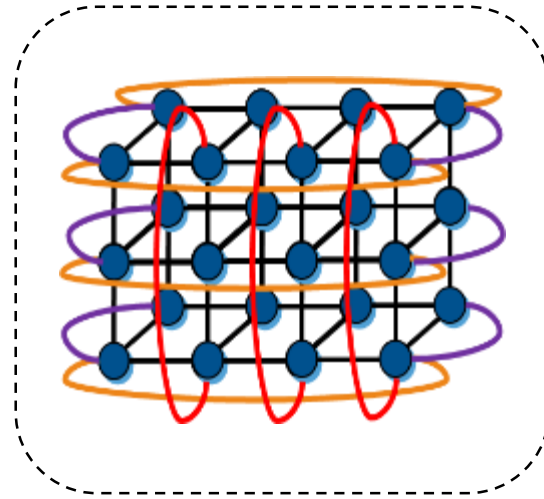


InfiniBand Delivers **13.5X** Higher Performance for Small Message Latency

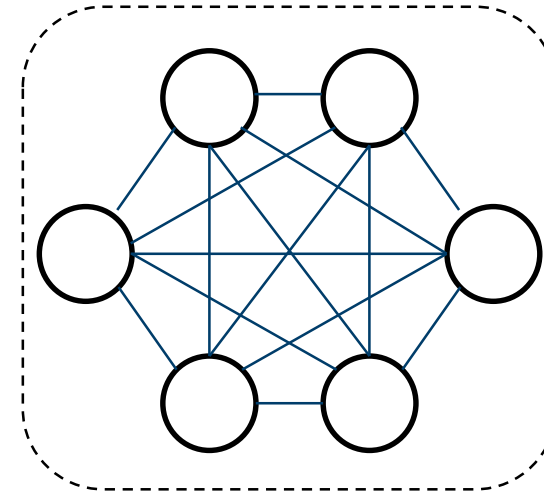
Supporting Variety of Topologies



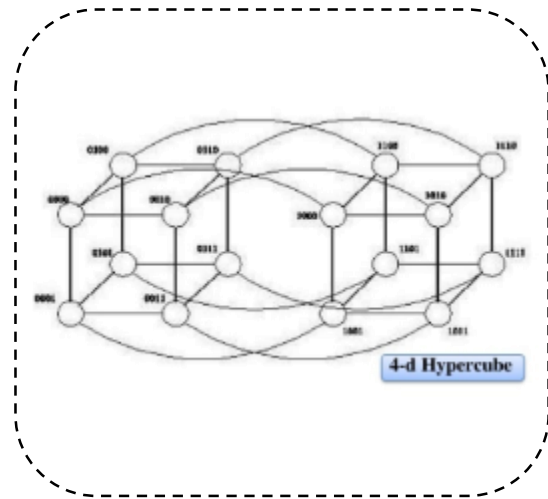
Fat Tree



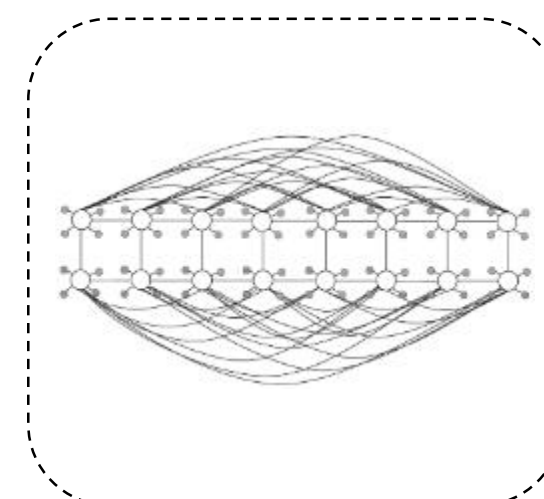
Torus



Dragonfly













Hypercube



HyperX

Highest-Performance 200Gb/s InfiniBand Solutions

Adapters		<p>200Gb/s Adapter, 0.6us latency 215 million messages per second (10 / 25 / 40 / 50 / 56 / 100 / 200Gb/s)</p>	
Switch		<p>40 HDR (200Gb/s) InfiniBand Ports 80 HDR100 InfiniBand Ports Throughput of 16Tb/s, <90ns Latency</p>	
SoC		<p>System on Chip and SmartNIC Programmable adapter Smart Offloads</p>	
Interconnect		<p>Transceivers Active Optical and Copper Cables (10 / 25 / 40 / 50 / 56 / 100 / 200Gb/s)</p>	
Software		<p>MPI, SHMEM/PGAS, UPC For Commercial and Open Source Applications Leverages Hardware Accelerations</p>	

ConnectX-6 HDR InfiniBand Adapter

Leading Connectivity

- 200Gb/s InfiniBand and Ethernet
 - HDR, HDR100, EDR (100Gb/s) and lower speeds
 - 200GbE, 100GbE and lower speeds
- Single and dual ports

Leading Performance

- 200Gb/s throughput, 215 million message per second
- PCIe Gen3 / Gen4, 32 lanes
- Integrated PCIe switch
- Multi-Host

Leading Features

- In-network computing and memory for HPC collective offloads
- Security – Block-level encryption to storage, key management, FIPS
- Storage – NVMe Emulation, NVMe-oF target, Erasure coding, T10/DIF

ConnectX[®]·6



HDR InfiniBand Switches

40 QSFP56 ports

- 40 ports of HDR, 200G
- 80 ports of HDR100, 100G

800 QSFP56 ports

- 800 ports of HDR, 200G
- 1600 ports of HDR100, 100G



Real Time Network Visibility

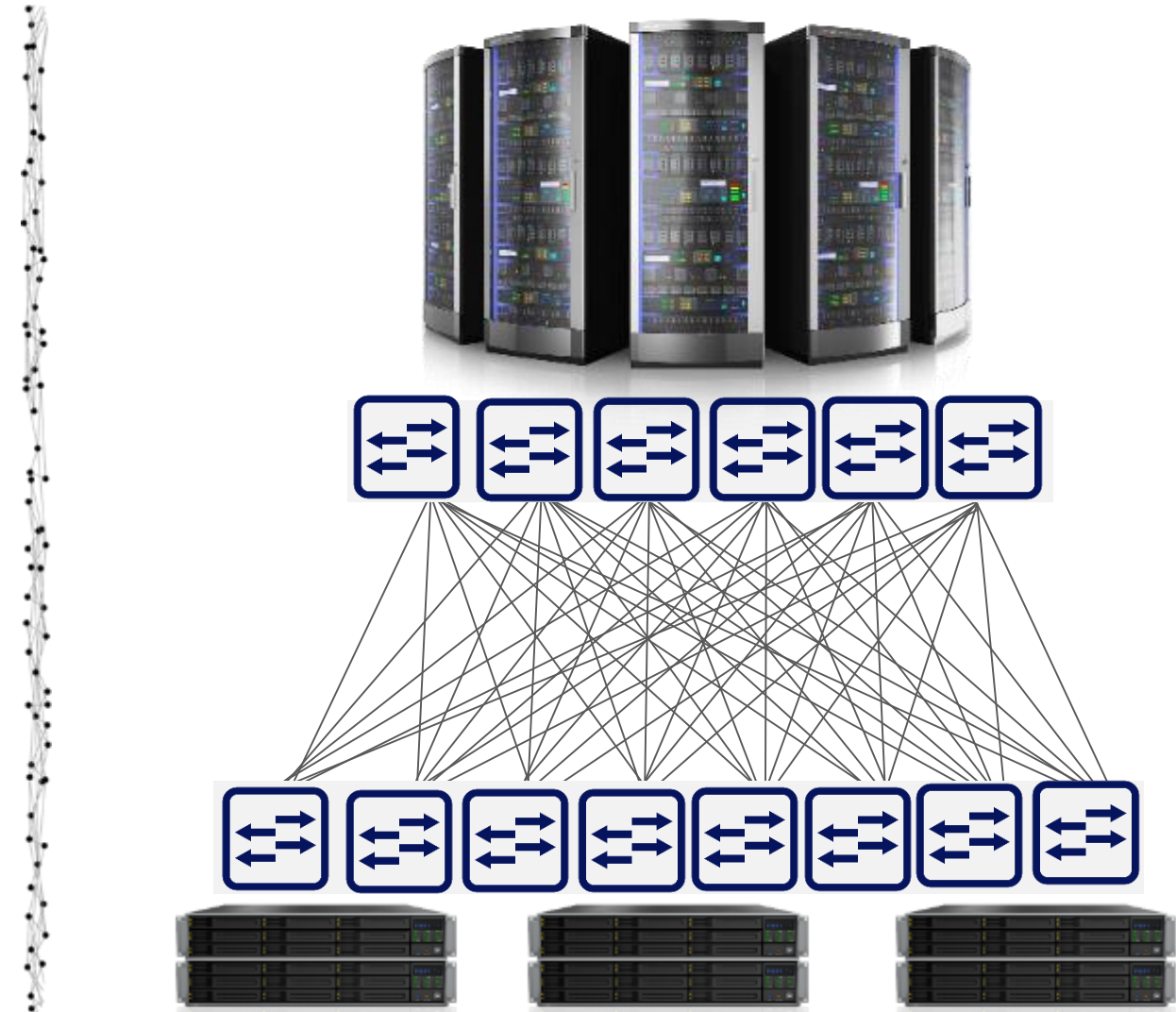
Built-in Hardware Sensors for Rich Traffic Telemetry and Data Collection

Advanced monitoring for troubleshooting

- 8 mirror agents triggered by congestion, buffer usage and latency
- Measure queue depth using histograms (64ns granularity)

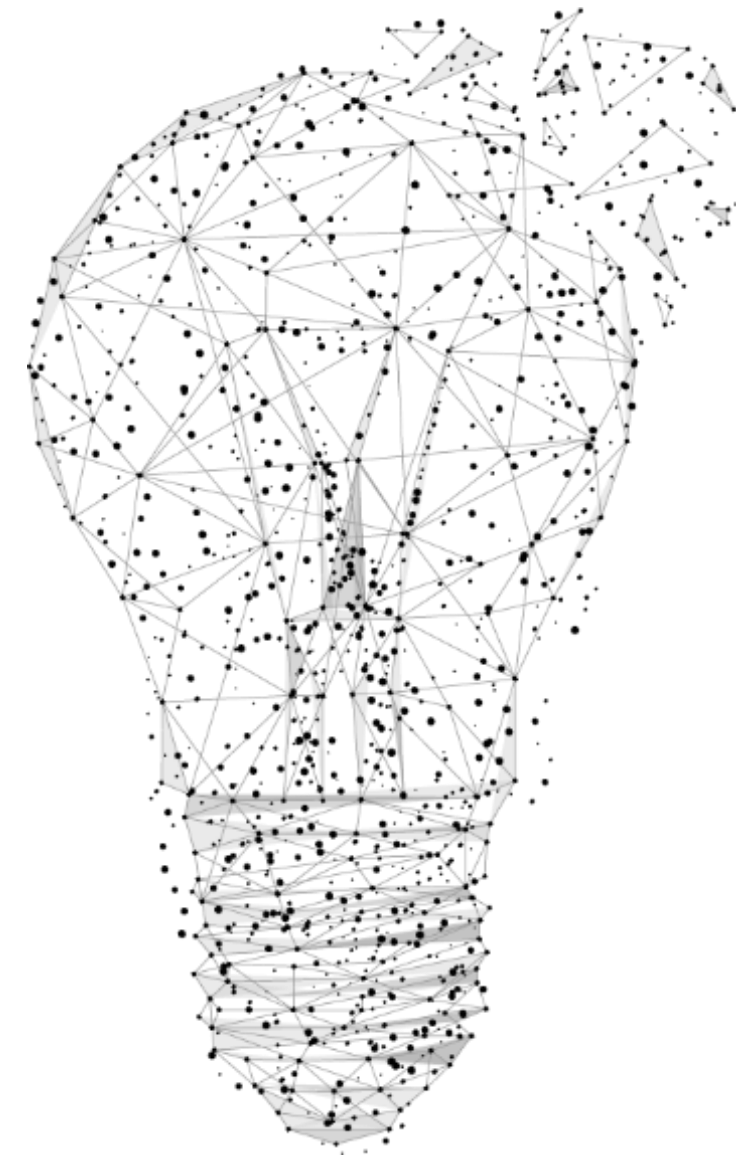
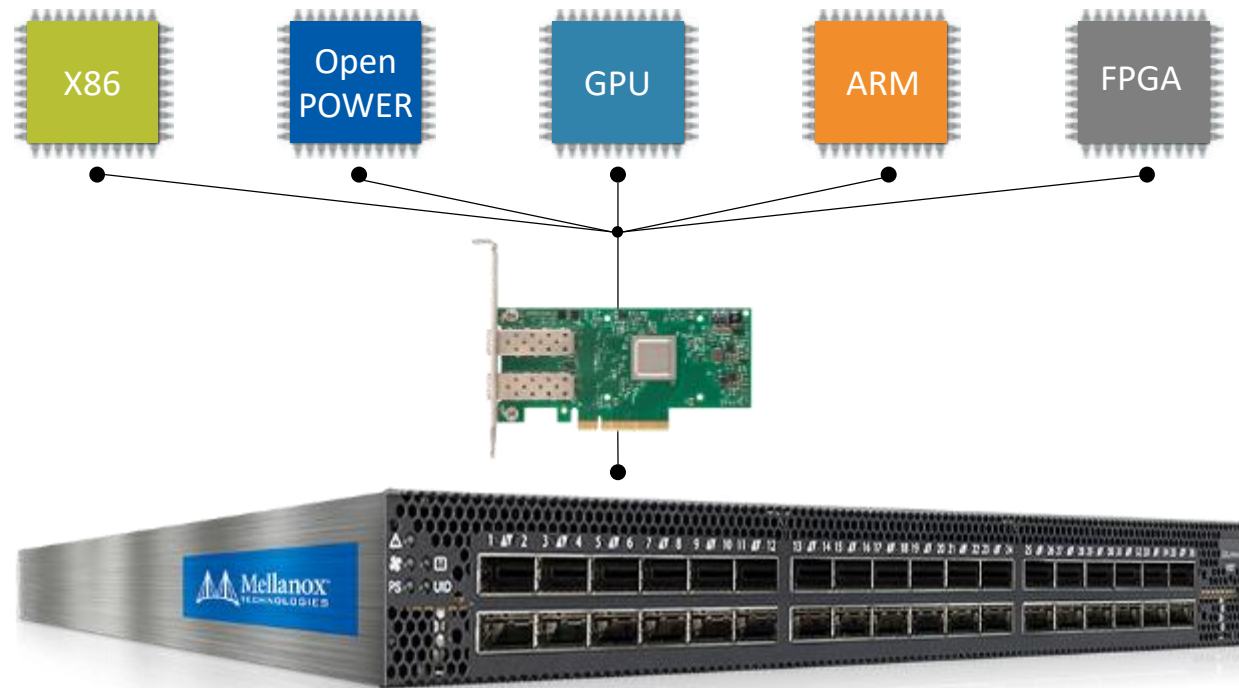
Network status/health in real time

- Buffer snapshots
- Congestion notifications and buffers status



End-to-End Solutions for All Platforms

Unleashing the Power of all Compute Architectures



Highest Performance and Scalability for Intel, AMD, IBM Power, NVIDIA, Arm and FPGA-based Compute and Storage Platforms at 10, 20, 25, 40, 50, 100, 200 and 400Gb/s Speeds

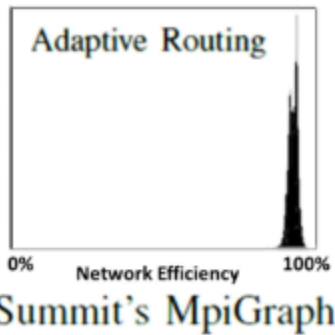
Highest Performance and Scalability for Exascale Platforms

OAK RIDGE
National Laboratory

SHARP SHIELD
SELF-HEALING



96%
Network
Utilization

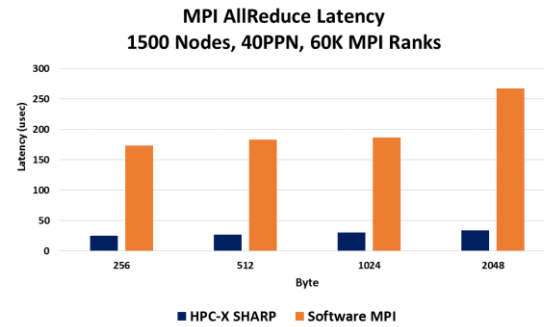


UNIVERSITY OF
TORONTO
SciNet

SHARP

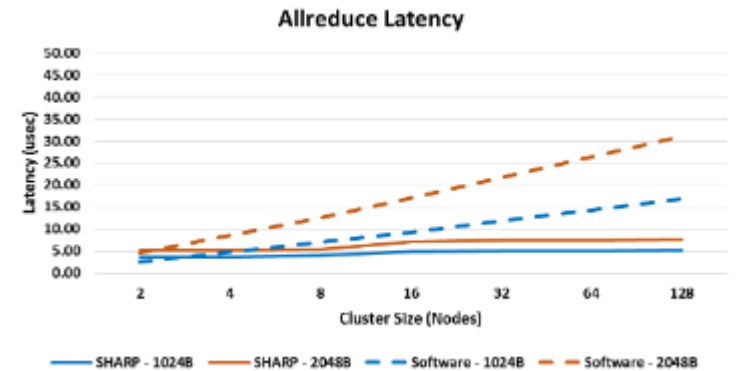


7X
Higher
Performance

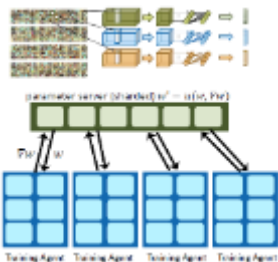


SHARP
Scalable Hierarchical
Aggregation and
Reduction Protocol

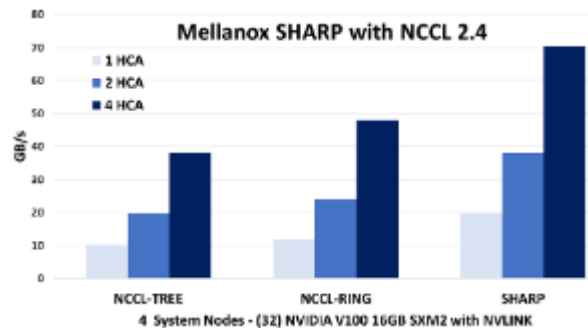
**Flat
Latency**



**Deep
Learning**

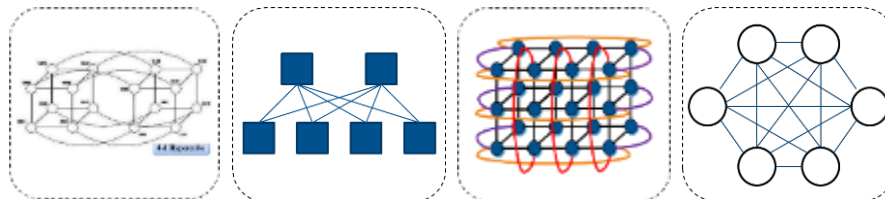


2X
Higher
Performance



SHIELD
SELF-HEALING

5000X
Higher
Resiliency



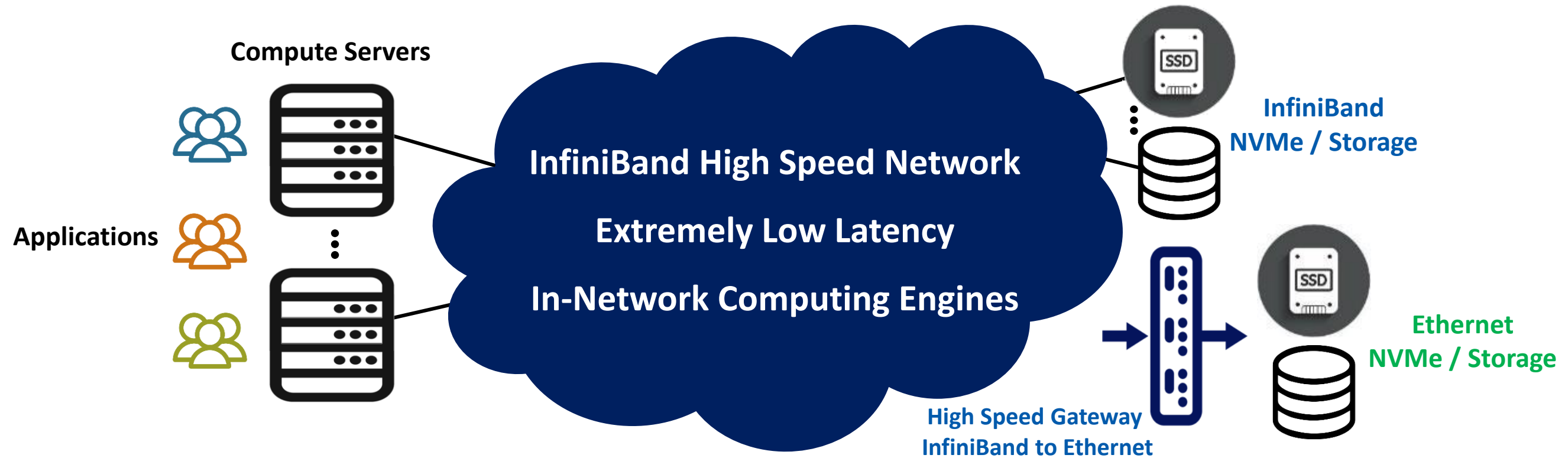
XDR 1000G

NDR 400G

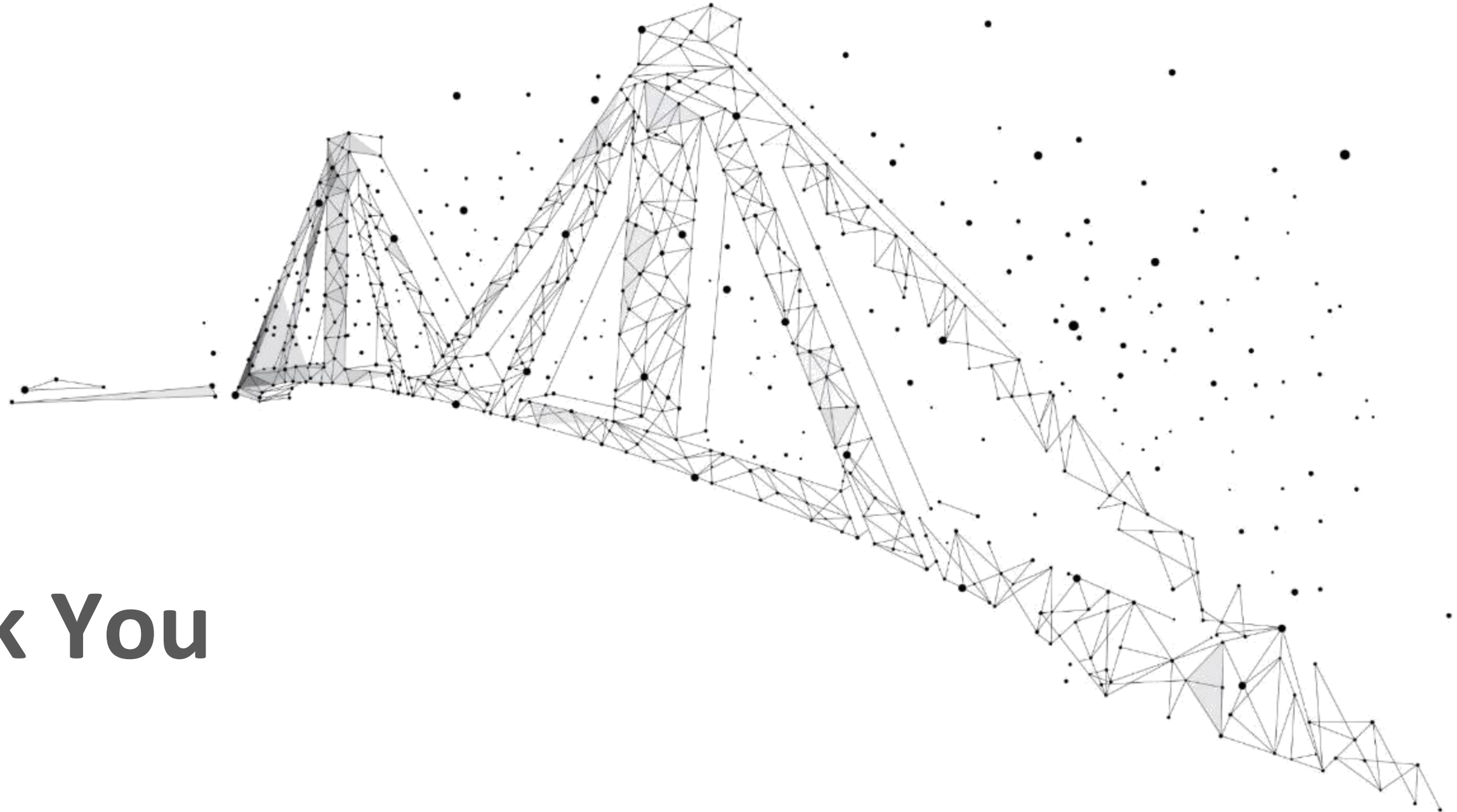
HDR 200G



InfiniBand Delivers Highest Performance and ROI



- High data throughput, extremely low latency, high message rate, RDMA and GPUDirect
- In-Network Computing – SHARP™, MPI acceleration engines
- Self Healing Network with SHIELD for highest network resiliency
- End to end adaptive routing and Quality of Service
- InfiniBand to Ethernet gateway for Ethernet storage or other Ethernet connectivity



Thank You

