# Connecting Visions

## Paving the Road to Exascale

December 2019

# HDR 200G InfiniBand Wins Next Generation HPC and AI Supercomputers (Examples)

TACC
TEXAS ADVANCED COMPUTING CENTER
NSF

23.5 Petaflops
8K HDR InfiniBand Nodes
Fat-Tree Topology

KMA Korea Meteorological Administration

50 Petaflops
7.2K HDR InfiniBand Nodes
Dragonfly+ Topology

Australian National University

3K HDR InfiniBand Nodes
Dragonfly+ Topology

MISSISSIPPI STATE UNIVERSITY
NOAA

3.1 Petaflops
1.8K HDR InfiniBand Nodes
Fat-Tree Topology

CSC
FINNISH METEOROLOGICAL INSTITUTE

1.7 Petaflops
2K HDR InfiniBand Nodes
Dragonfly+ Topology

Microsoft Azure

Highest Performance Cloud
HDR InfiniBand

SDSC
SAN DIEGO SUPERCOMPUTER CENTER
PITTSBURGH SUPERCOMPUTING CENTER
NSF

筑波大学
University of Tsukuba

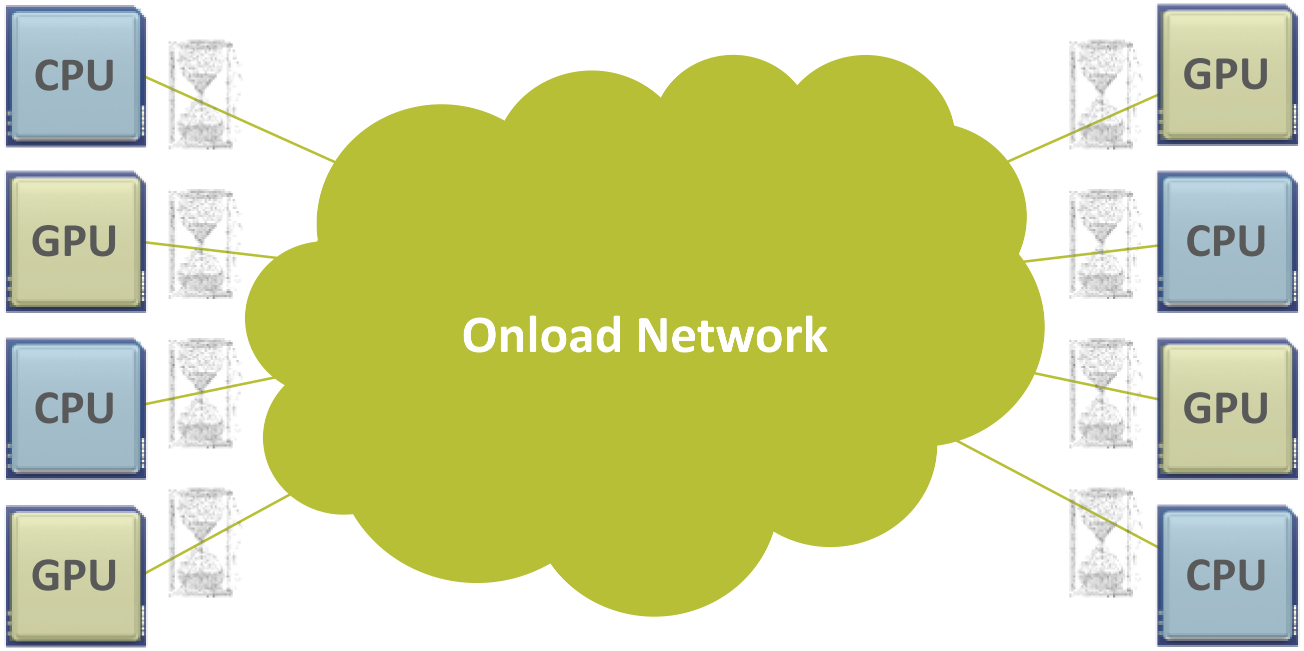
1.6 Petaflops
Hybrid CPU-GPU-FPGA
Fat-Tree Topology

InfiniBand H·D·R 200

# The Need for Intelligent and Faster Interconnect

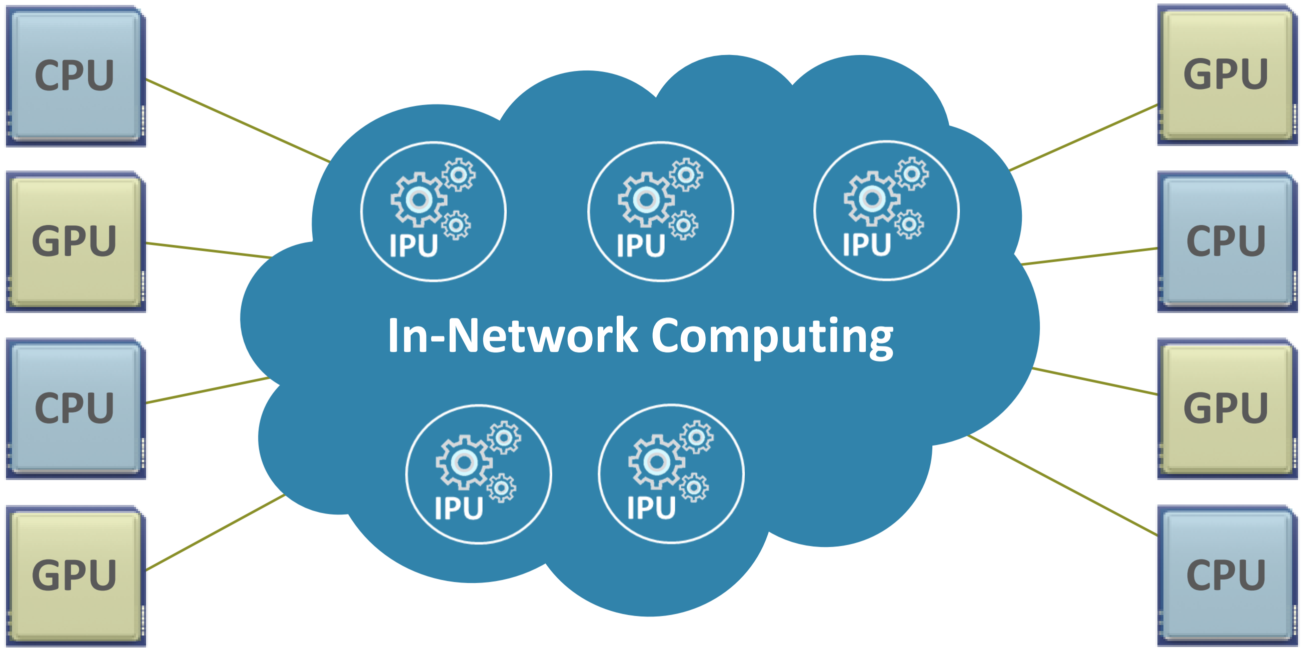
Faster Data Speeds and In-Network Computing
Enable Higher Performance and Scale

### CPU-Centric (Onload)



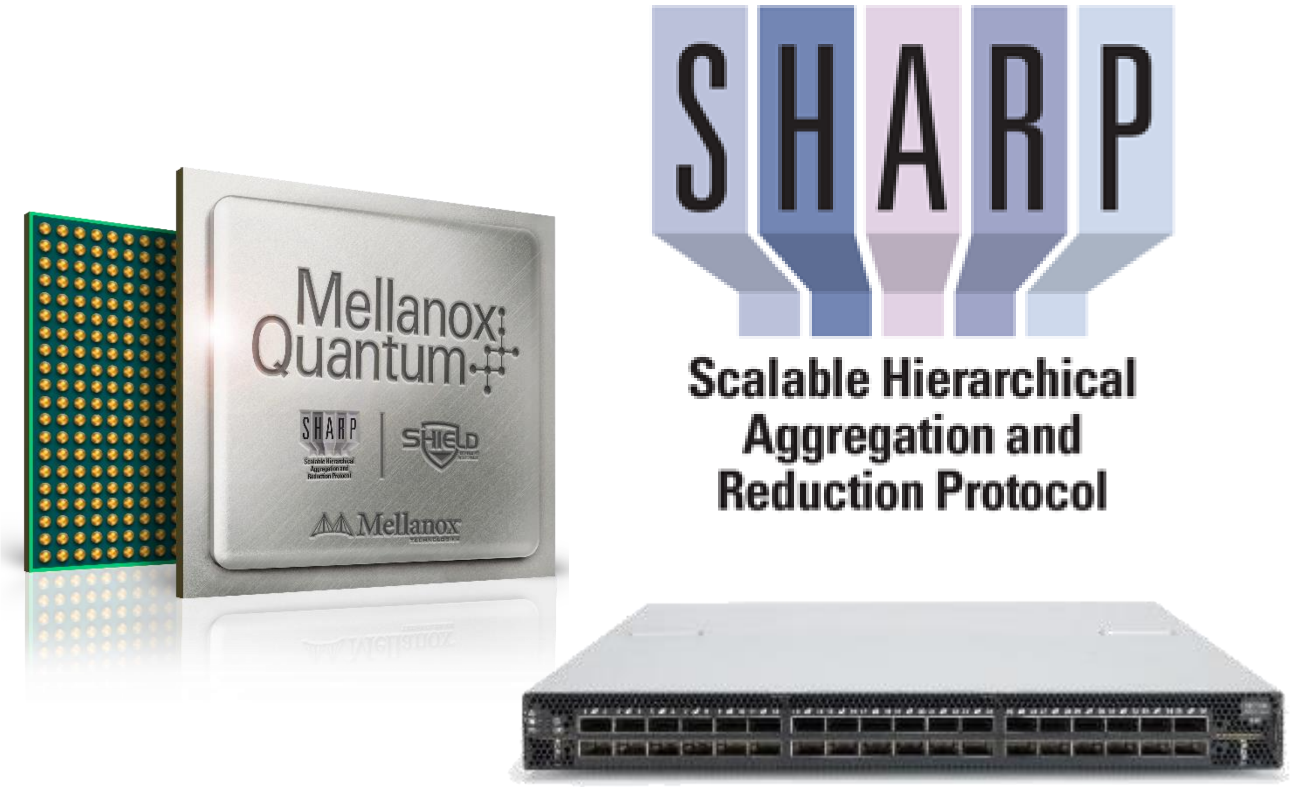
Must Wait for the Data
Creates Performance Bottlenecks

### Data-Centric (Offload)



Analyze Data as it Moves!
Higher Performance and Scale

# Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

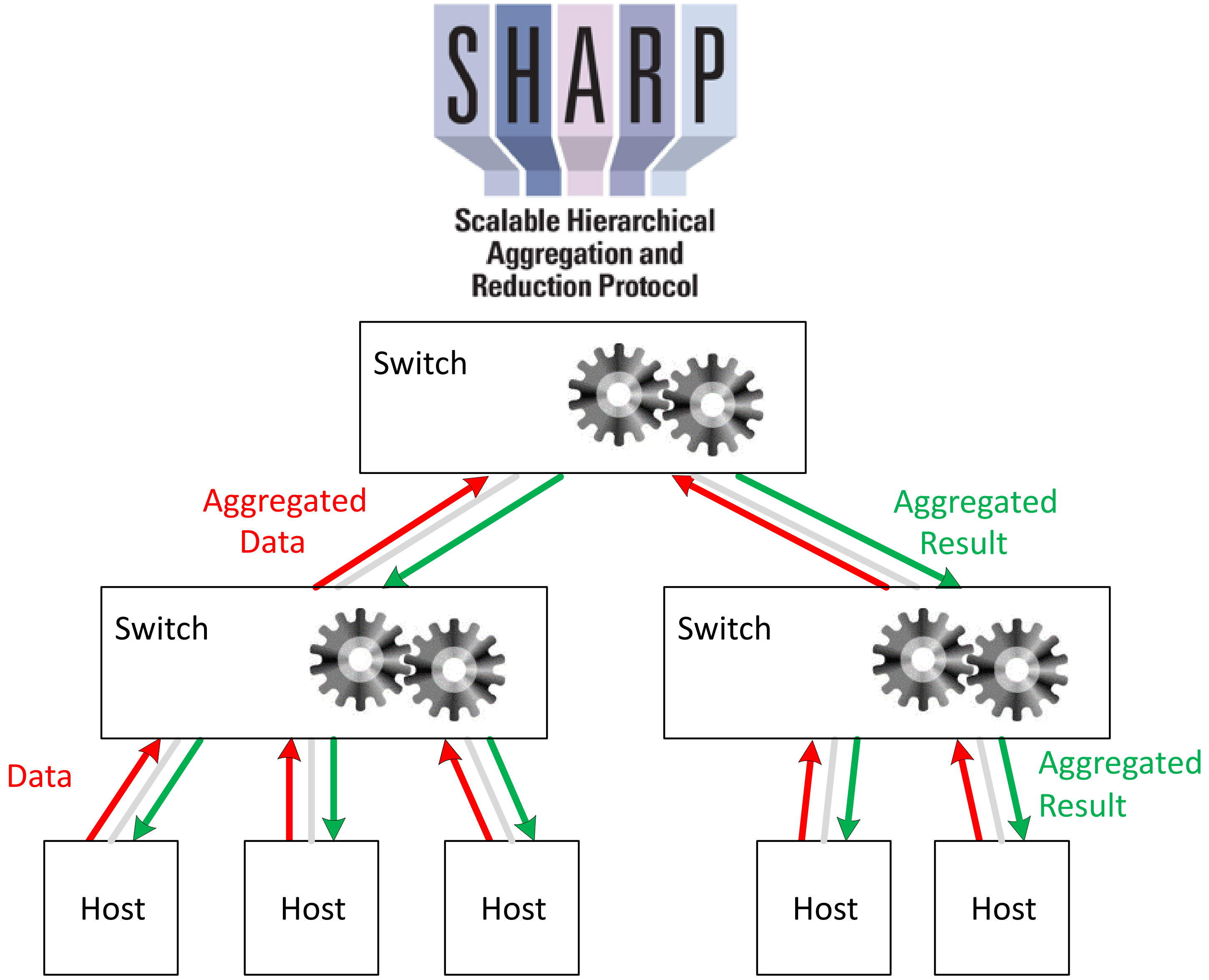


# Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

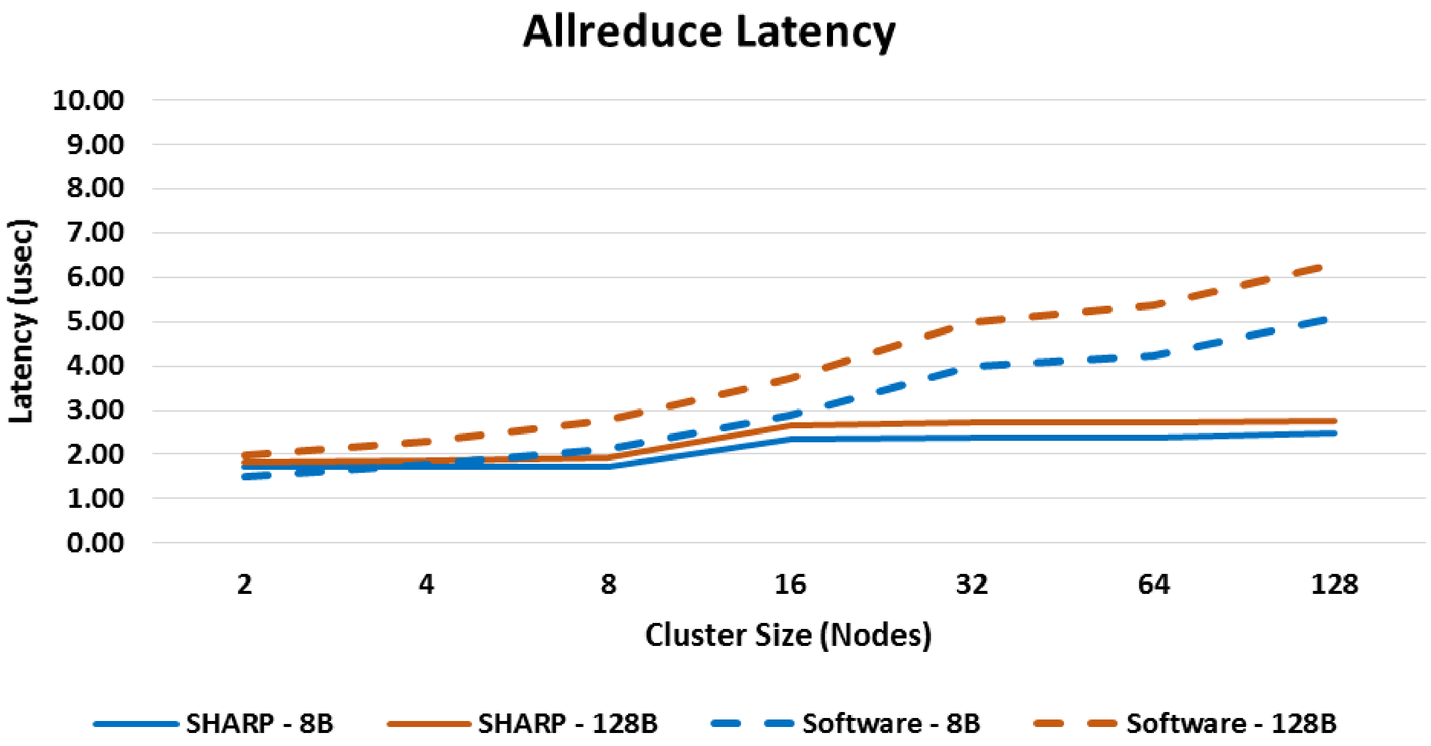
- Reliable Scalable General Purpose Primitive
  - In-network Tree based aggregation mechanism
  - Large number of groups
  - Multiple simultaneous outstanding operations

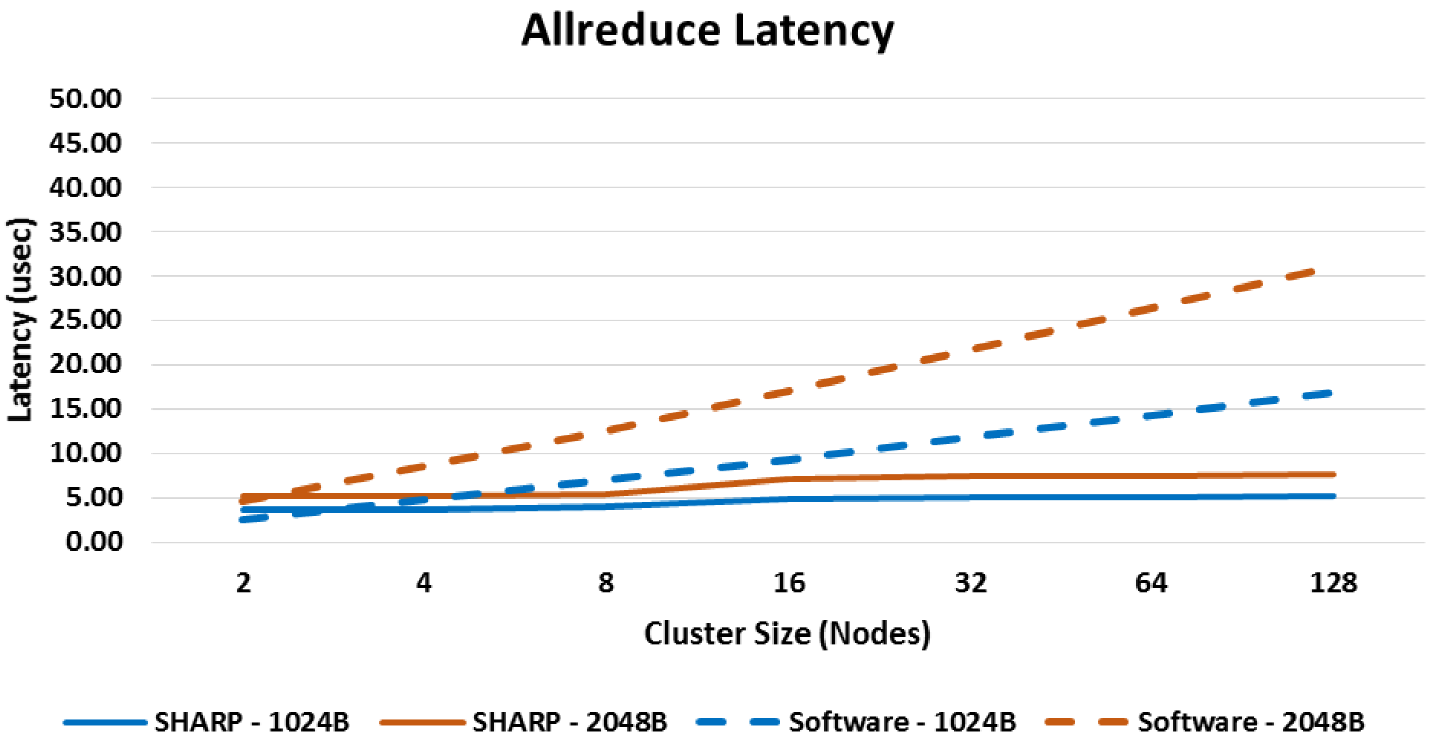
- Applicable to Multiple Use-cases
  - HPC Applications using MPI / SHMEM
  - Distributed Machine Learning applications

- Scalable High Performance Collective Offload
  - Barrier, Reduce, All-Reduce, Broadcast and more
  - Sum, Min, Max, Min-loc, max-loc, OR, XOR, AND
  - Integer and Floating-Point, 16/32/64 bits



# SHARP AllReduce Performance Advantages (128 Nodes)





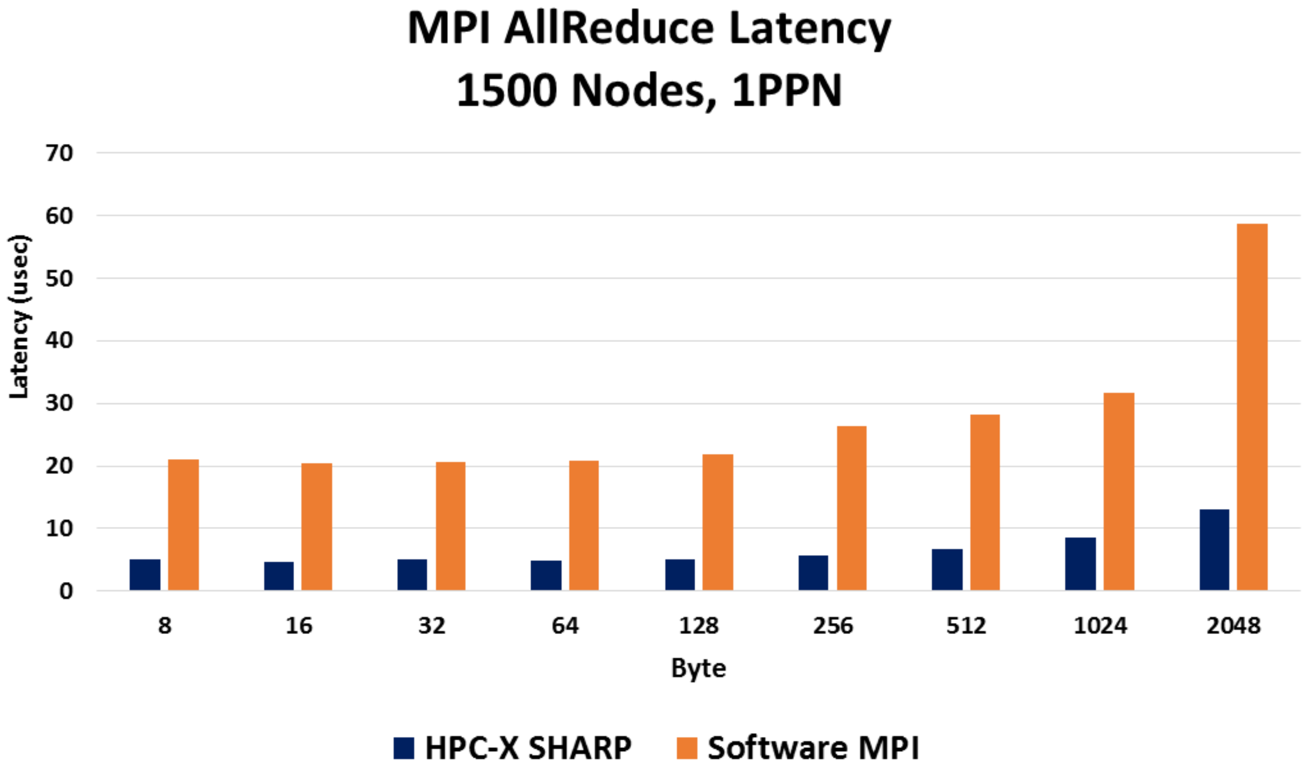
SHARP

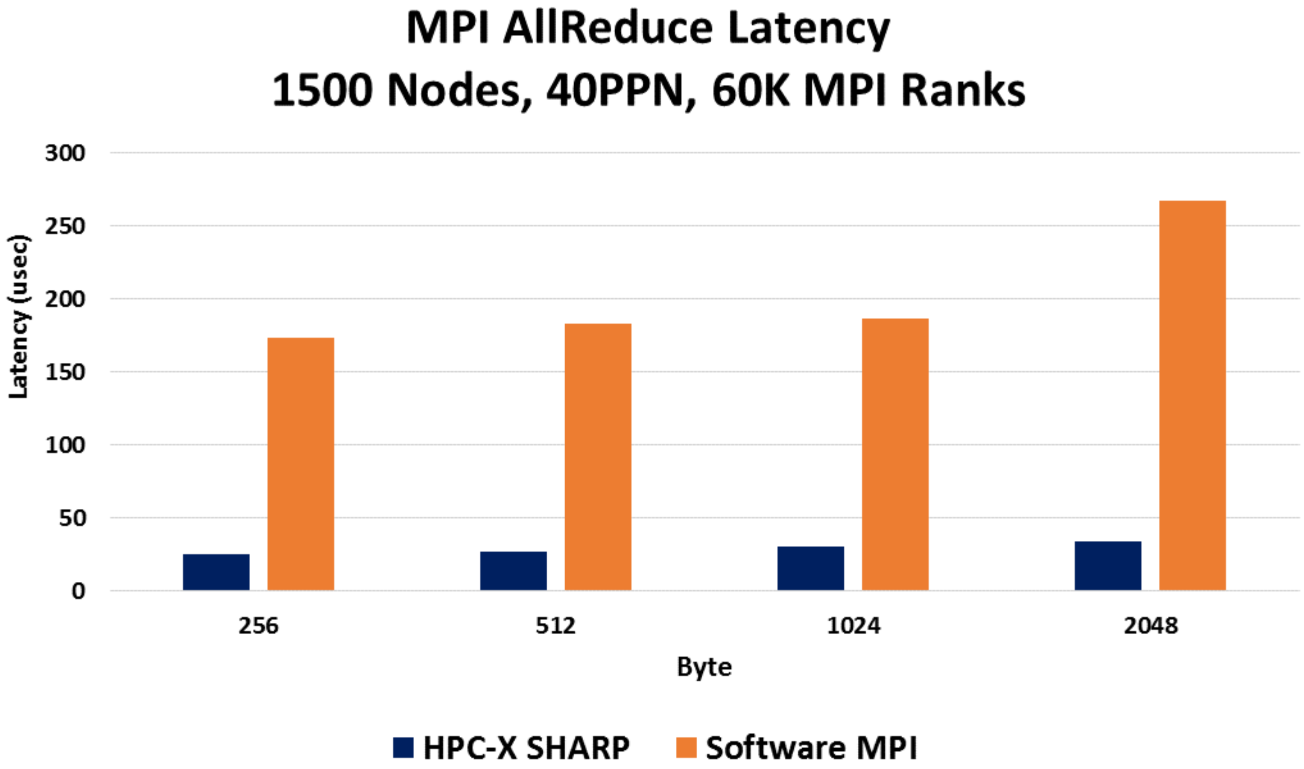
Scalable Hierarchical Aggregation and Reduction Protocol

SHARP enables 75% Reduction in Latency
Providing Scalable Flat Latency

# SHARP AllReduce Performance Advantages

## 1500 Nodes, 60K MPI Ranks, Dragonfly+ Topology



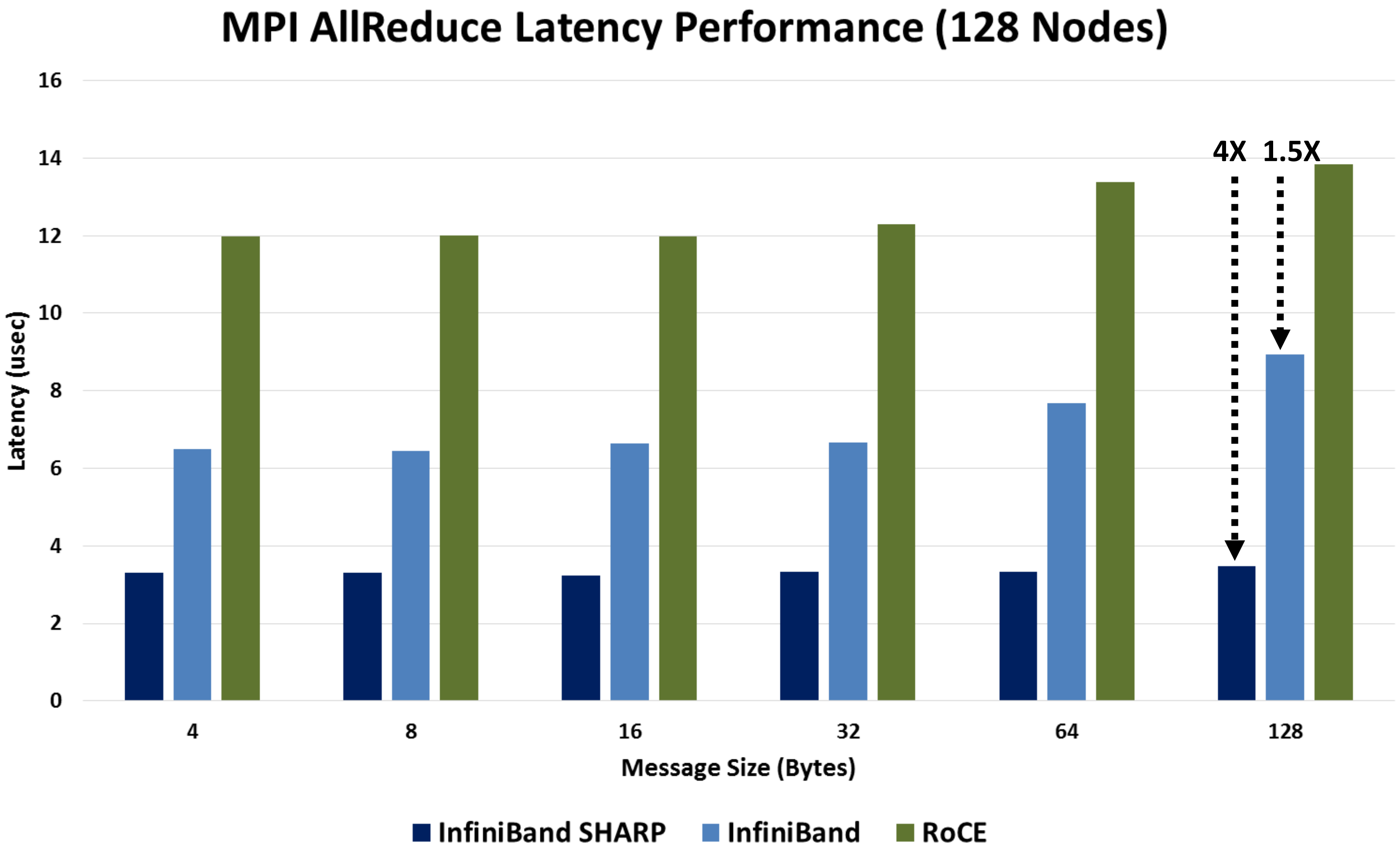


SHARP

Scalable Hierarchical Aggregation and Reduction Protocol

SHARP Enables Highest Performance

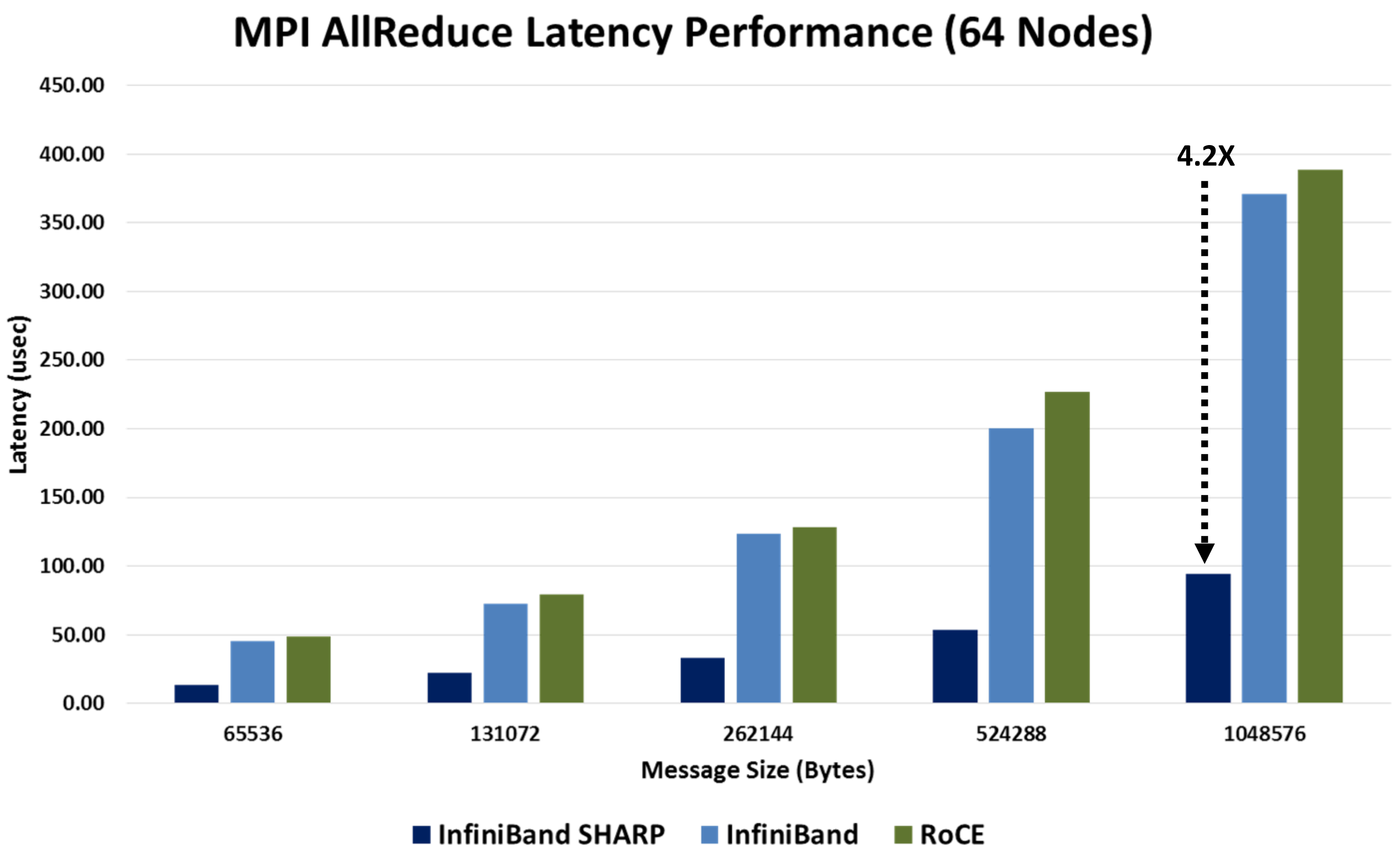
# SHARP Performance Advantage (Lower is Better)



Scalable Hierarchical Aggregation and Reduction Protocol

SHARP Enables 4X Higher Performance (Small Messages)

# SHARP Performance Advantage (Lower is Better)



SHARP

Scalable Hierarchical Aggregation and Reduction Protocol

SHARP Enables 4.2X Higher Performance (Large Messages)

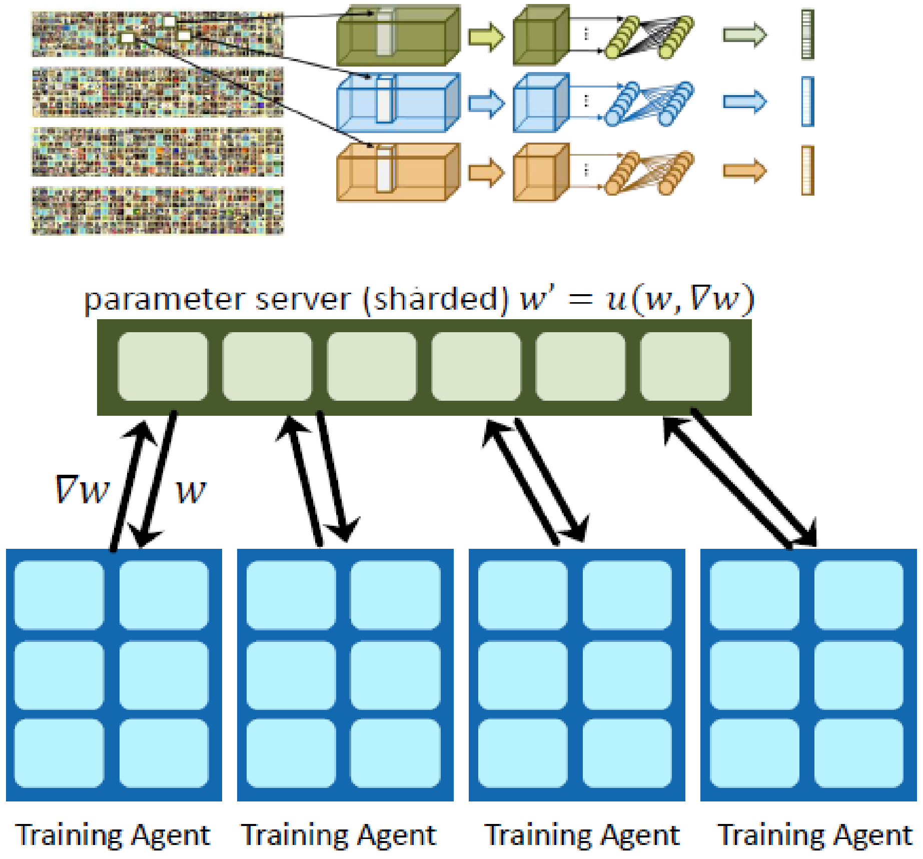
# SHARP Accelerates AI Performance

The CPU in a parameter server
becomes the bottleneck

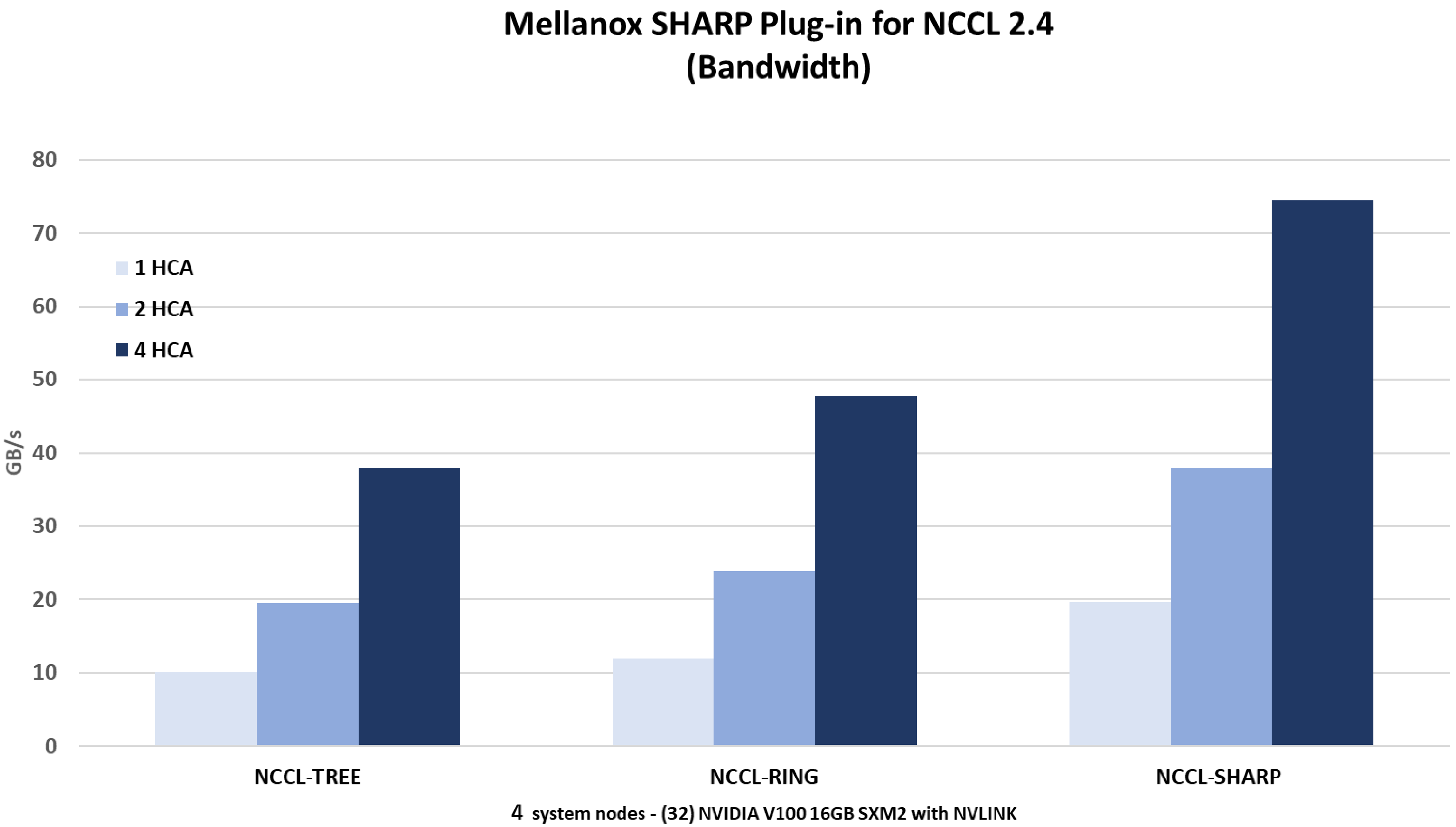
SHARP

Scalable Hierarchical
Aggregation and
Reduction Protocol

- Performs the Gradient Averaging
- Replaces all physical parameter servers
- Accelerate AI Performance

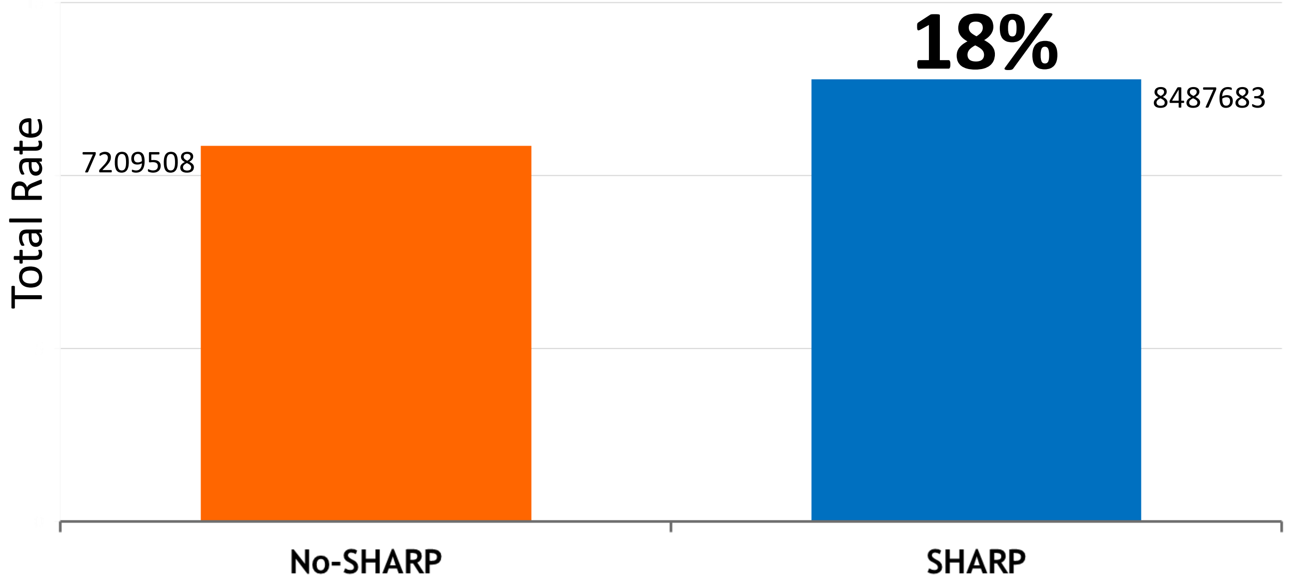


# SHARP Delivers Highest Performance for AI



# SHARP Delivers Highest Performance for AI

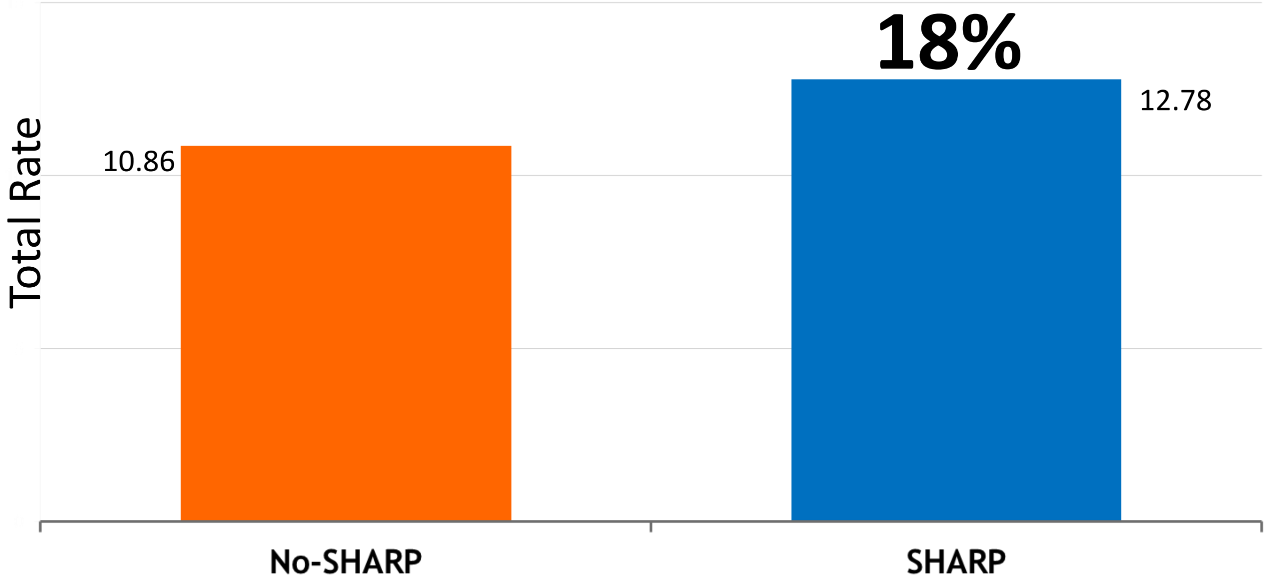
## GNMT MLPerf Benchmark
## Neural Machine Translation



24xDGX1V + 4xMellanox ConnectX-6

GNMT MLPerf 0.6 benchmark: Batch Size=32, Overlap=0.15

## VAE Benchmark
## Variable Auto-Encoder



32xDGX1V + 4xMellanox ConnectX-6

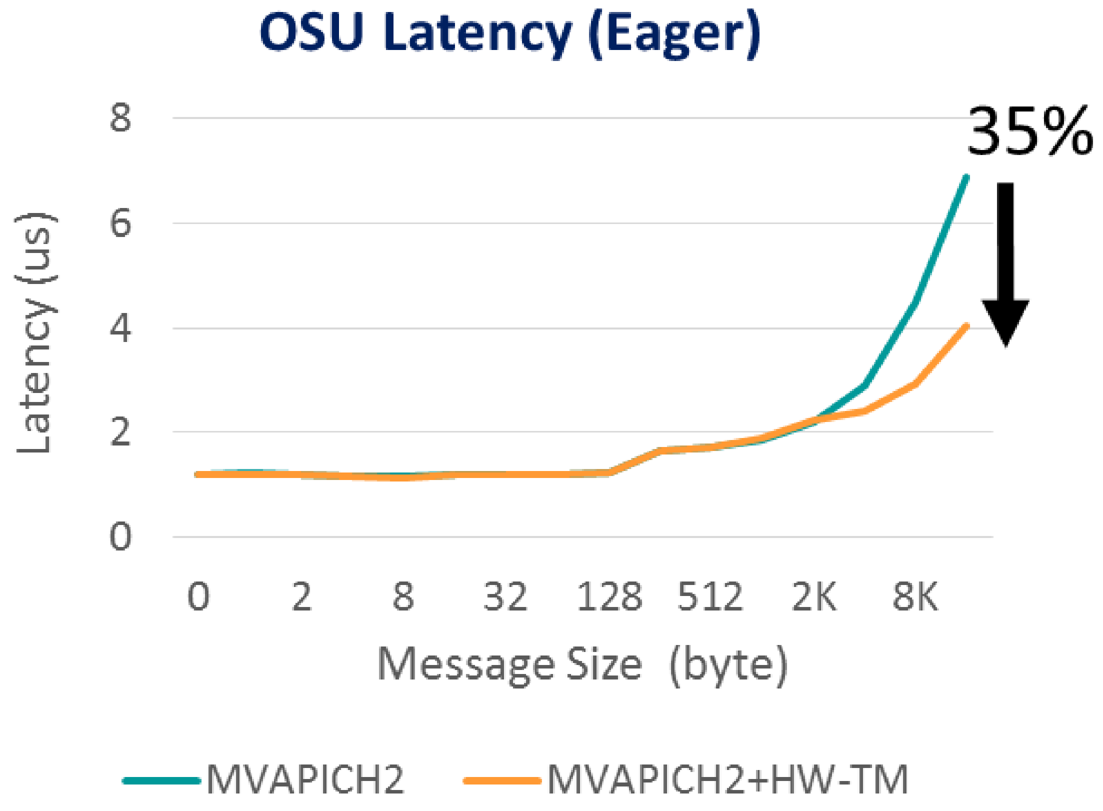
VAE benchmark: Model=3, BS=512

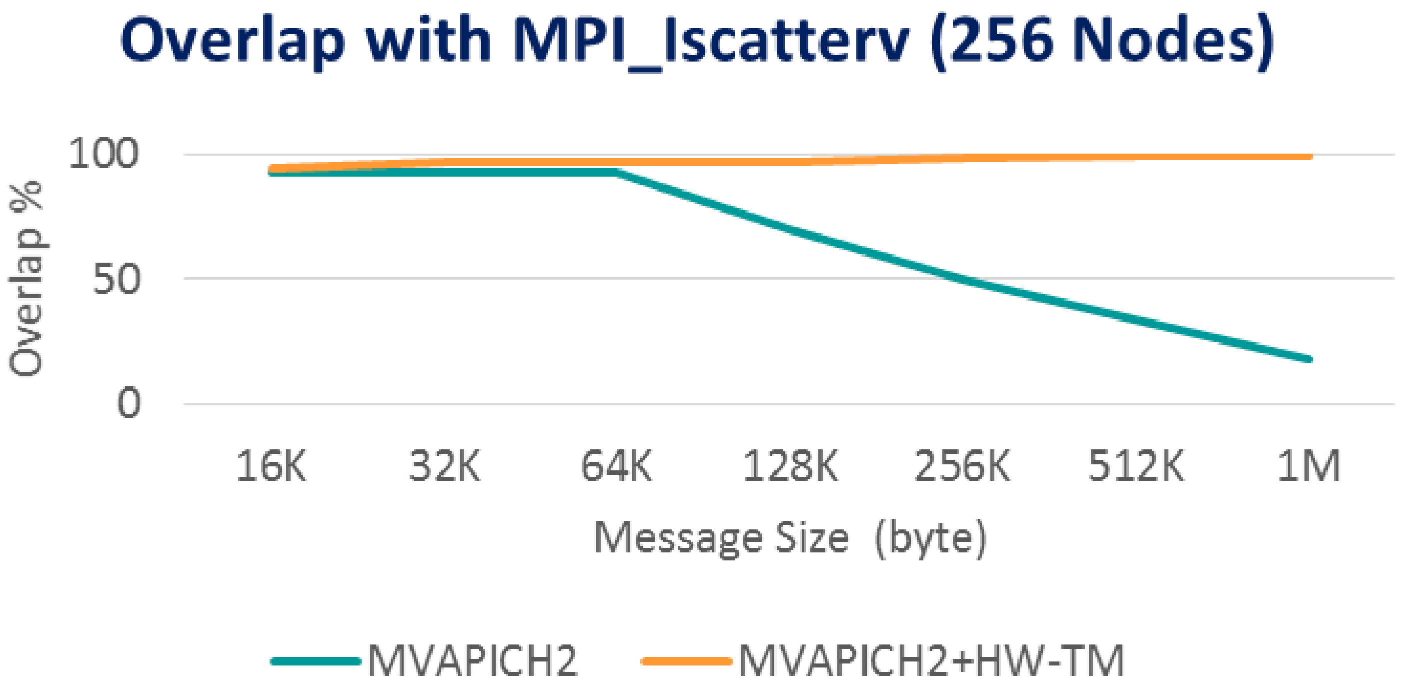
Scalable Hierarchical Aggregation and Reduction Protocol

**SHARP Delivers Highest Performance**

# MPI Tag Matching Hardware Engine

# Tag Matching Hardware Engine Performance Advantage

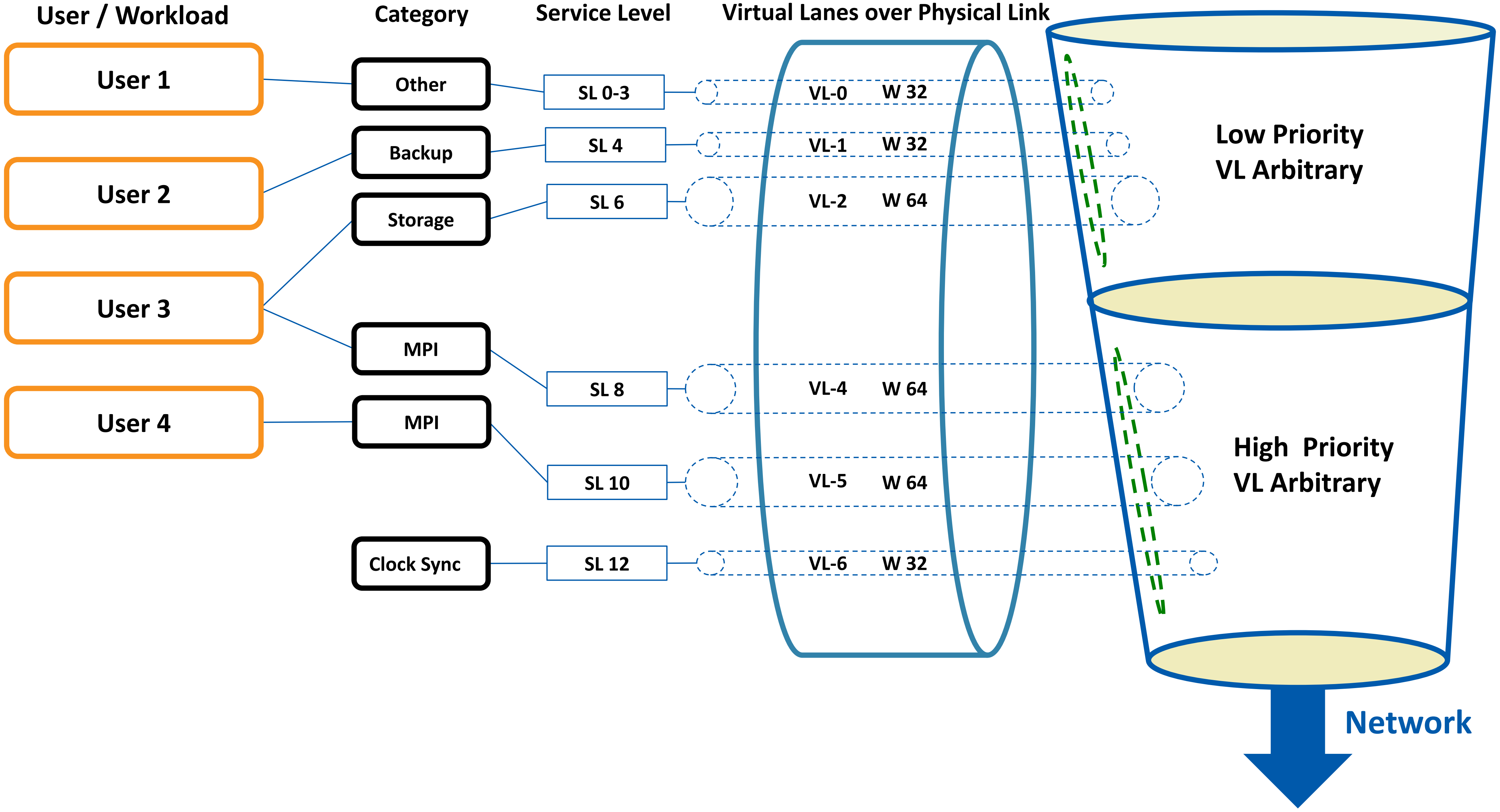




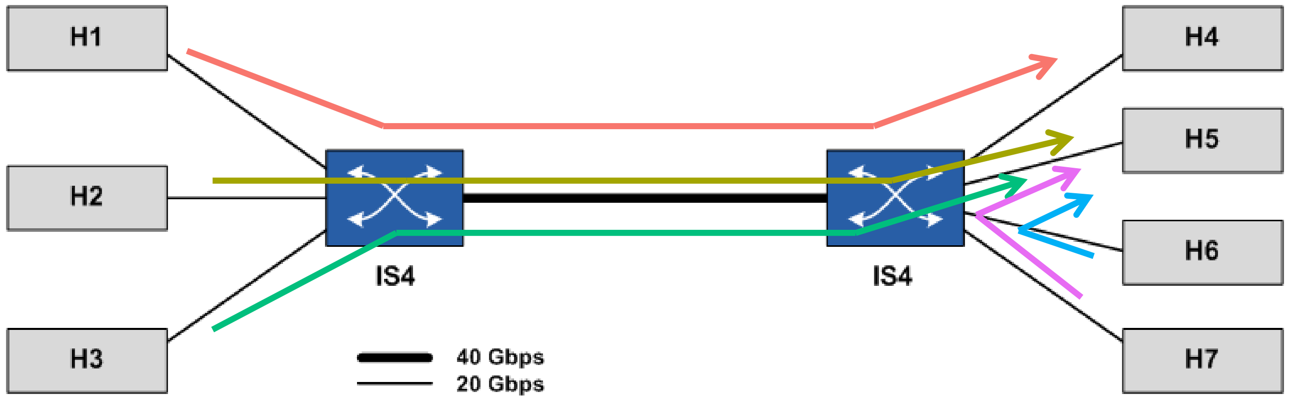
**Courtesy of Dhabaleswar K. (DK) Panda**
**Ohio State University**

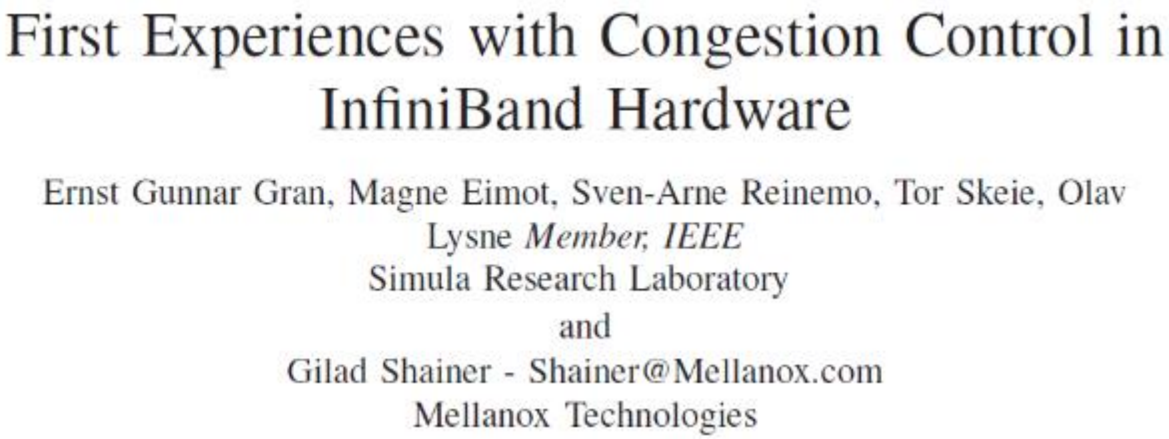
# Quality of Service

# InfiniBand Quality of Service



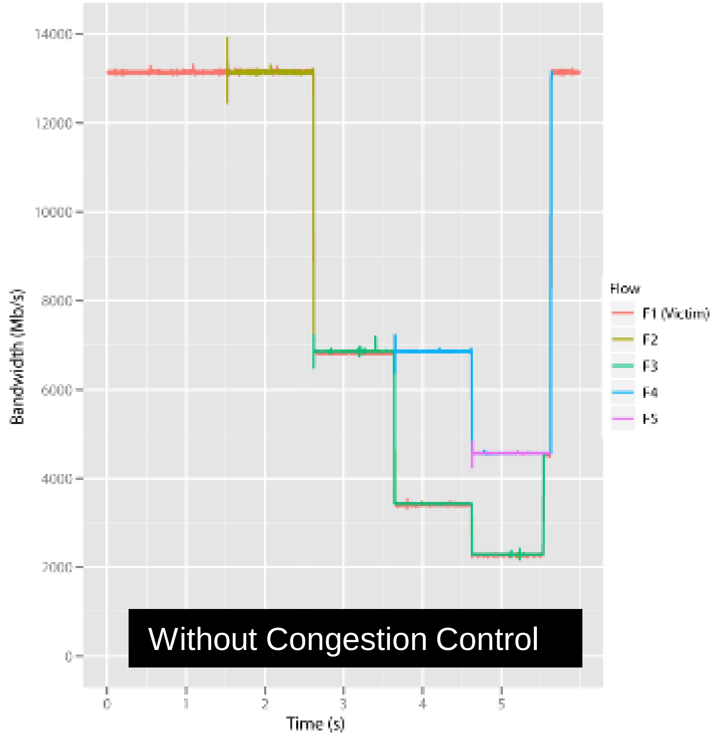
# InfiniBand Congestion Control



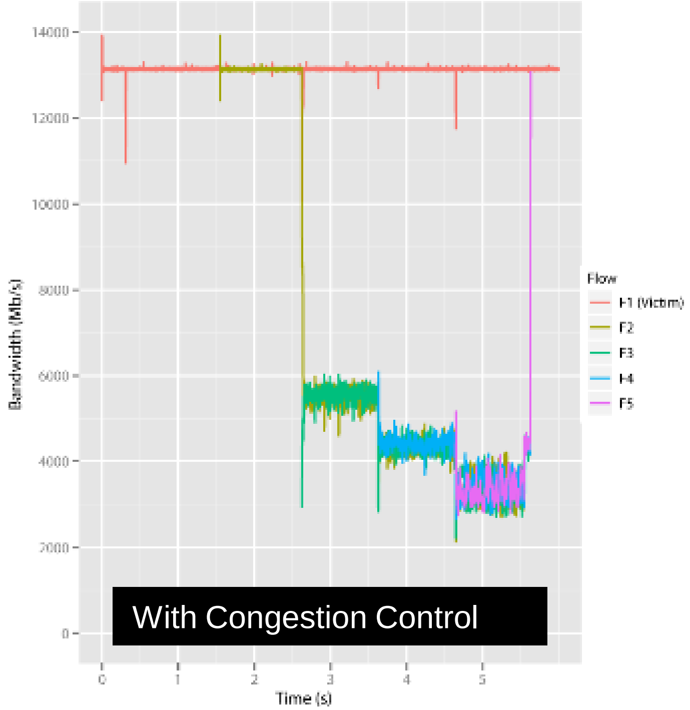

First Experiences with Congestion Control in InfiniBand Hardware

Ernst Gunnar Gran, Magne Eimot, Sven-Arne Reinemo, Tor Skeie, Olav Lysne *Member, IEEE*
Simula Research Laboratory
and
Gilad Shainer - Shainer@Mellanox.com
Mellanox Technologies

**Congestion – Throughput loss**



**No congestion – highest throughput!**



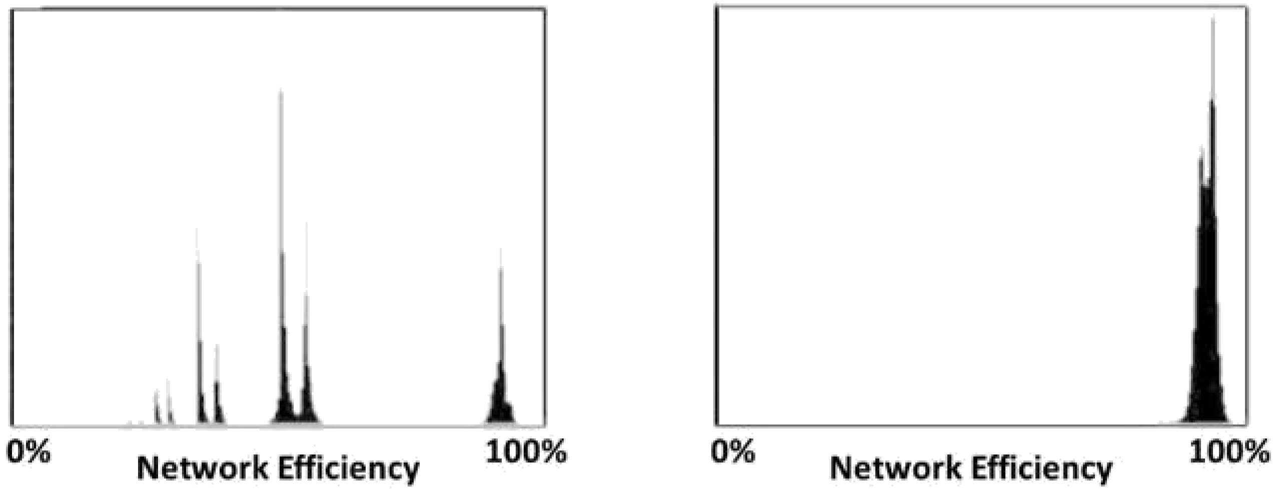
# Adaptive Routing

# 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

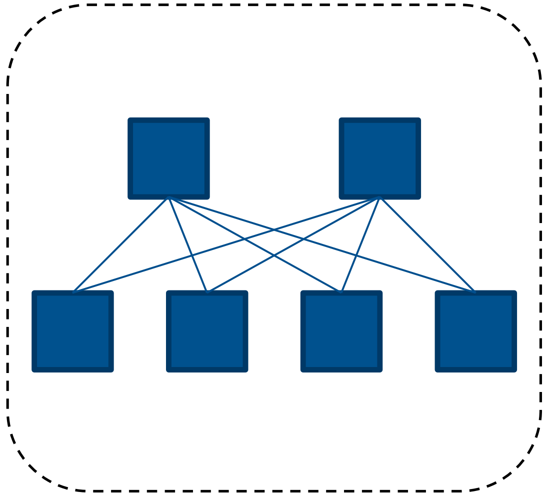
Adaptive Routing

## Oak Ridge National Lab Summit Supercomputer

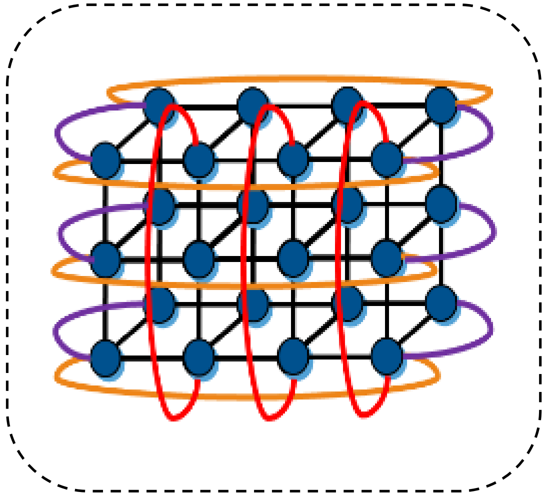
*"The Design, Deployment, and Evaluation of the CORAL Pre-Exascale Systems", Sudharshan S. Vazhkudai, Arthur S. Bland, Al Geist, Christopher J. Zimmer, Scott Atchley, Sarp Oral, Don E. Maxwell, Veronica G. Vergara Larrea, Wayne Joubert, Matthew A. Ezell, Dustin Leverman, James H. Rogers, Drew Schmidt, Mallikarjun Shankar, Feiyi Wang, Junqi Yin (Oak Ridge National Laboratory) and Bronis R. de Supinski, Adam Bertsch, Robin Goldstone, Chris Chambreau, 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 Rosenburg, Bob Walkup (IBM)*

# Network Topologies

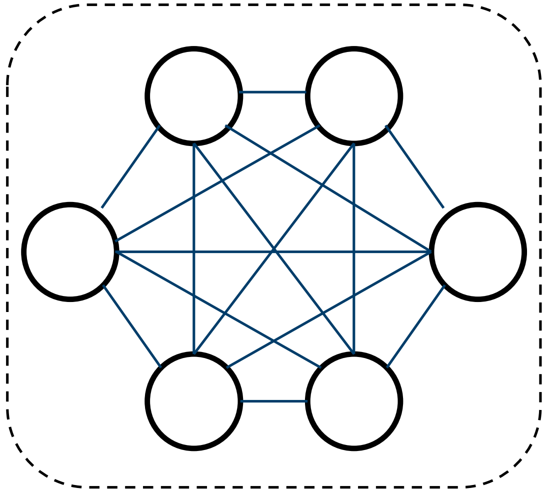
# Supporting Variety of Topologies



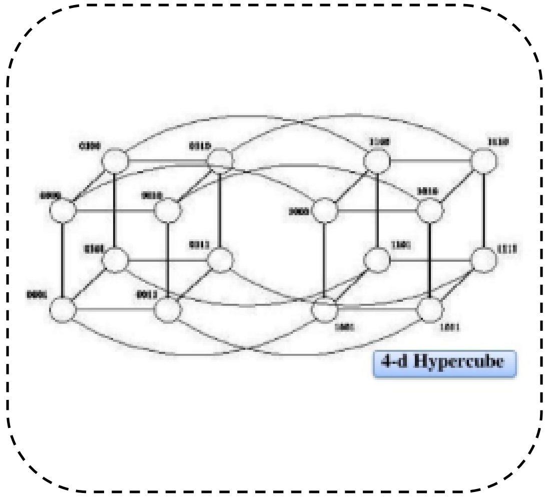
Fat Tree



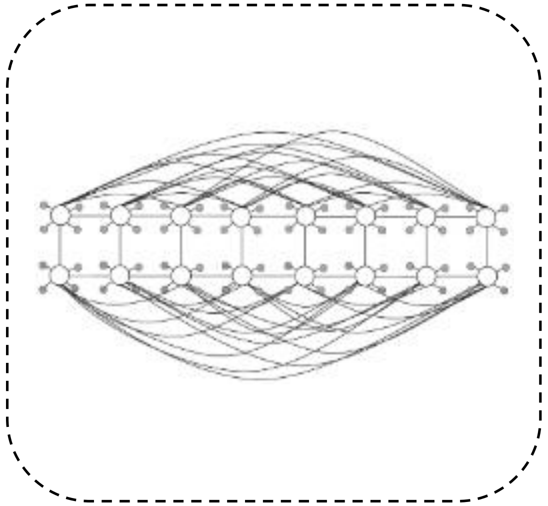
Torus



Dragonfly



Hypercube



HyperX

# HDR InfiniBand

# Highest-Performance 200Gb/s InfiniBand Solutions

Adapters

ConnectX-6

200Gb/s Adapter
215 million messages per second
(10 / 25 / 40 / 50 / 56 / 100 / 200Gb/s)

Switch

Mellanox Quantum

40 HDR (200Gb/s) InfiniBand Ports
80 HDR100 InfiniBand Ports
Throughput of 16Tb/s, 130ns Latency

SoC

BlueField-2

System on Chip and SmartNIC
Programmable adapter
Smart Offloads

Interconnect

LinkX™

Transceivers
Active Optical and Copper Cables
(10 / 25 / 40 / 50 / 56 / 100 / 200Gb/s)

Software

HPC-X™

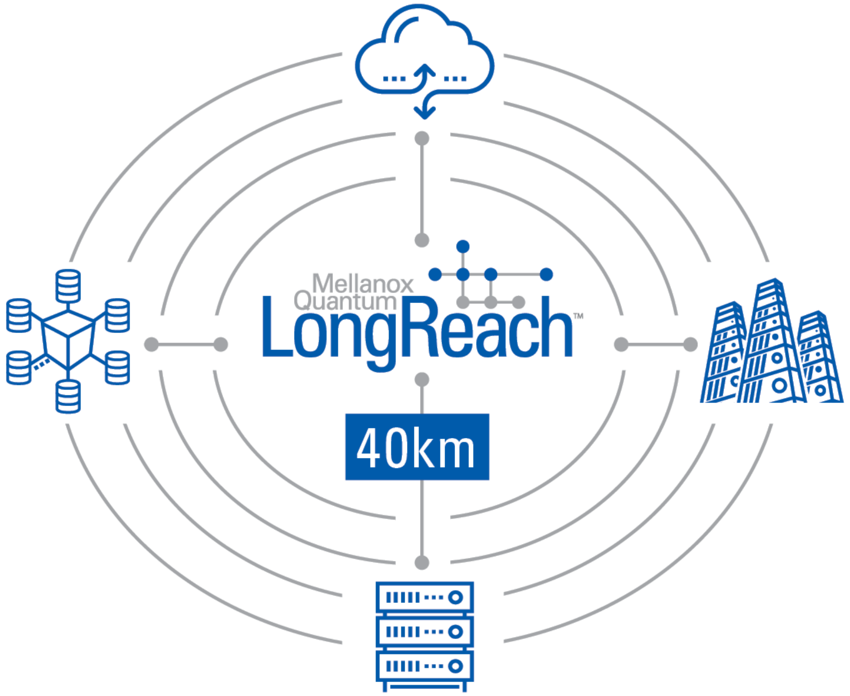
MPI, SHMEM/PGAS, UPC
For Commercial and Open Source Applications
Leverages Hardware Accelerations

HPC-X™
Message Passing Interface (MPI)
HPC-X™
Toolkit
HPC-X™
Open Shared Memory (Open SHMEM)

# Mellanox Quantum LongReach™

## Extending InfiniBand to 40km Reach

- Seamlessly connects InfiniBand data-centers up to 40 kilometers-apart
- Scalability and load balancing across data-centers
- Continues compute service in case of data-center failures
- Standard HDR and EDR InfiniBand end-to-end
- Advanced In-Network Computing



IN-NETWORK COMPUTING

SHARP
Scalable Hierarchical Aggregation and Reduction Protocol

SHIELD
SELF-HEALING INTERCONNECT

RDMA GPUDirect

# Mellanox Skyway™ InfiniBand to Ethernet Gateway

NEW

- 100G EDR / 200G HDR InfiniBand to 100G and 200G Ethernet gateway
- 400G NDR / 800G XDR InfiniBand speeds ready
- Eight EDR/HDR100/HDR InfiniBand ports to eight 100/200G Ethernet
- Max throughput of 1.6 Terabit per second
- High availability and load balancing
- Mellanox Gateway operating system
- Scalable and efficient

Mellanox Skyway™

# Highest Performance and Scalability for Exascale Platforms

SHARP

Scalable Hierarchical Aggregation and Reduction Protocol

SHIELD

SELF-HEALING INTERCONNECT

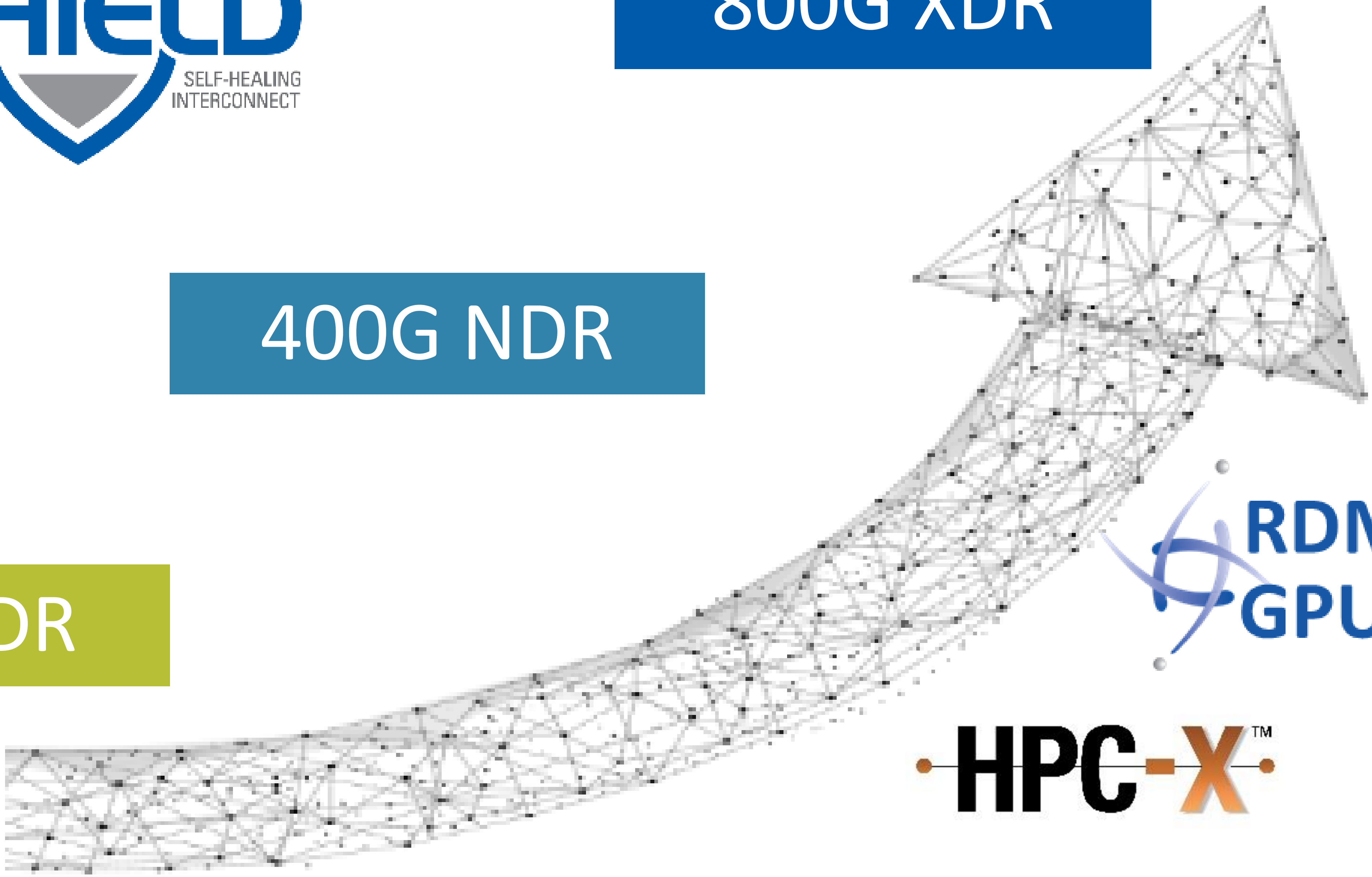
Mellanox

IN-NETWORK COMPUTING

800G XDR

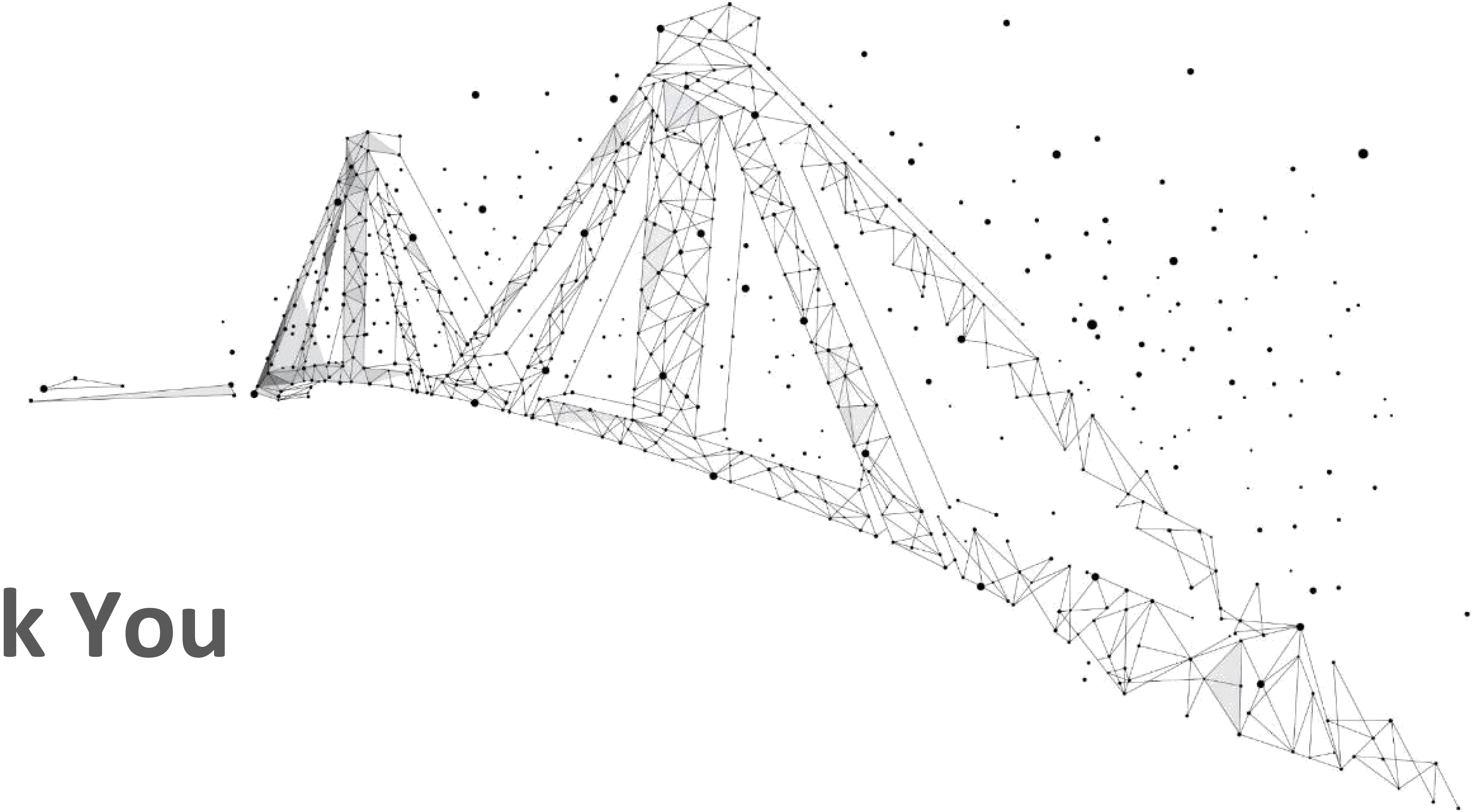
400G NDR

200G HDR



RDMA GPUDirect

HPC-X™



# Thank You