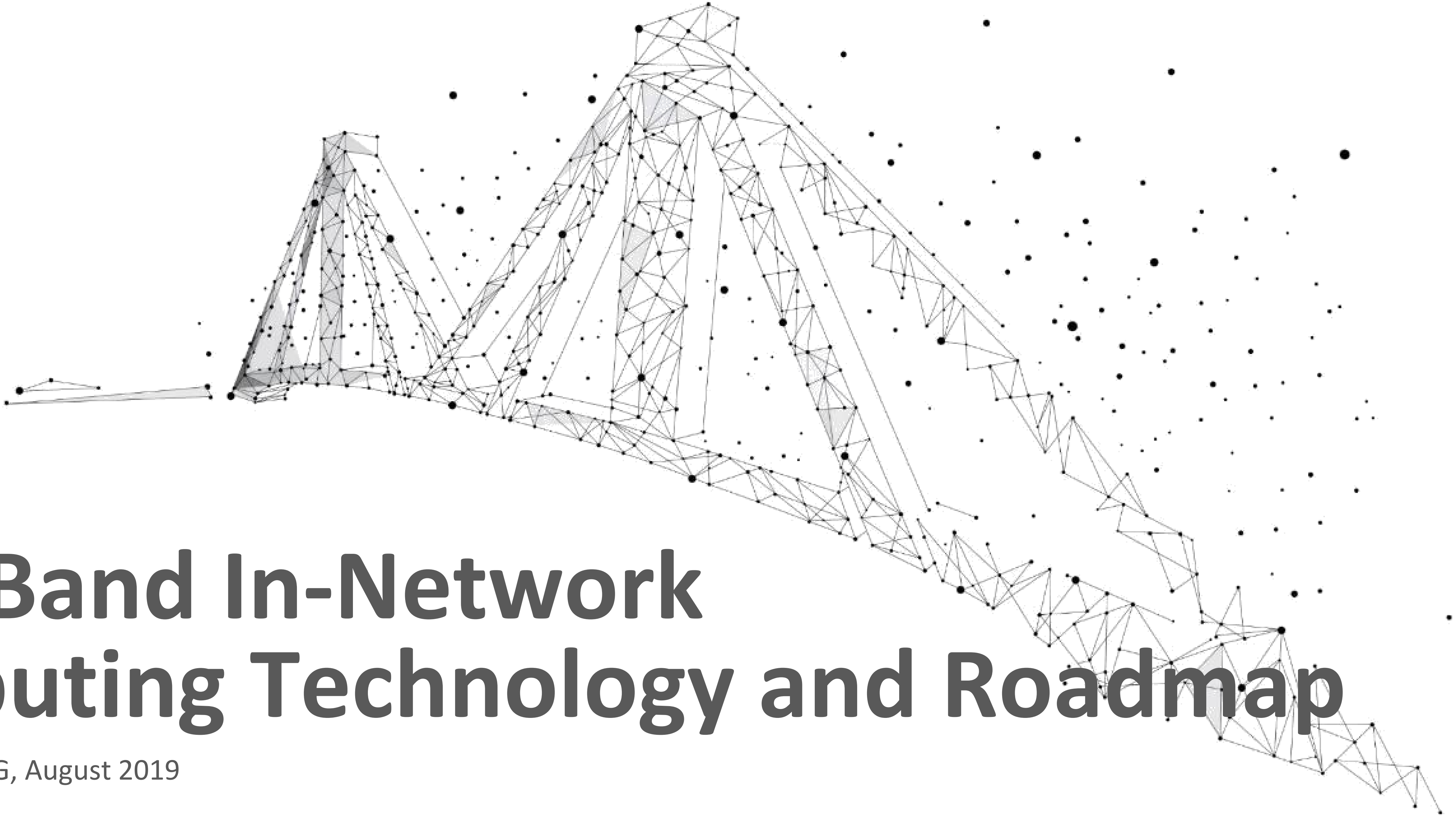
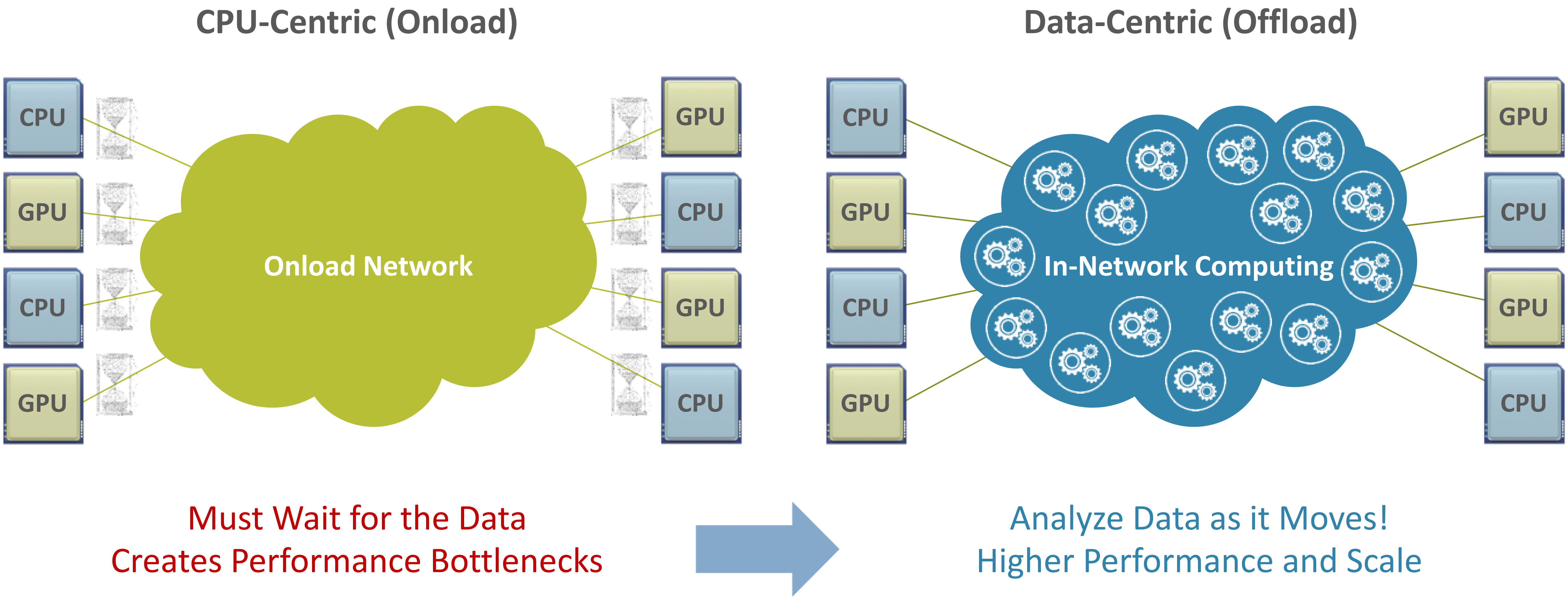


InfiniBand In-Network Computing Technology and Roadmap

Gilad Shainer, MUG, August 2019

The Need for Intelligent and Faster Interconnect

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



Must Wait for the Data
Creates Performance Bottlenecks

Analyze Data as it Moves!
Higher Performance and Scale

Accelerating All Levels of HPC / AI Frameworks

Application

- Data Analysis
- Real Time
- Deep Learning

Communication

- Mellanox SHARP In-Network Computing
- MPI Tag Matching
- MPI Rendezvous

UCF Unified Communication Framework

SHARP Scalable Hierarchical Aggregation and Reduction Protocol

Network

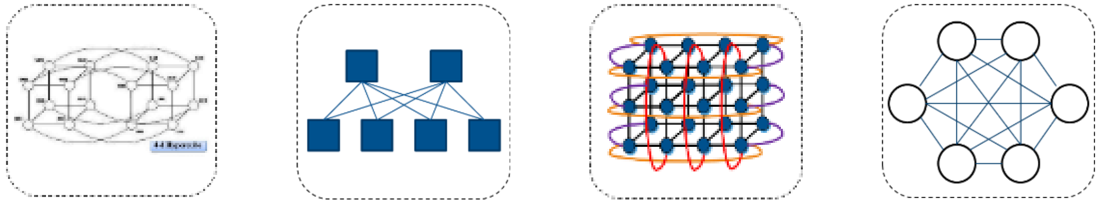
- Network Transport Offload
- RDMA and GPU-Direct RDMA
- SHIELD (Self-Healing Network)
- Enhanced Adaptive Routing and Congestion Control

GPUDirect RDMA

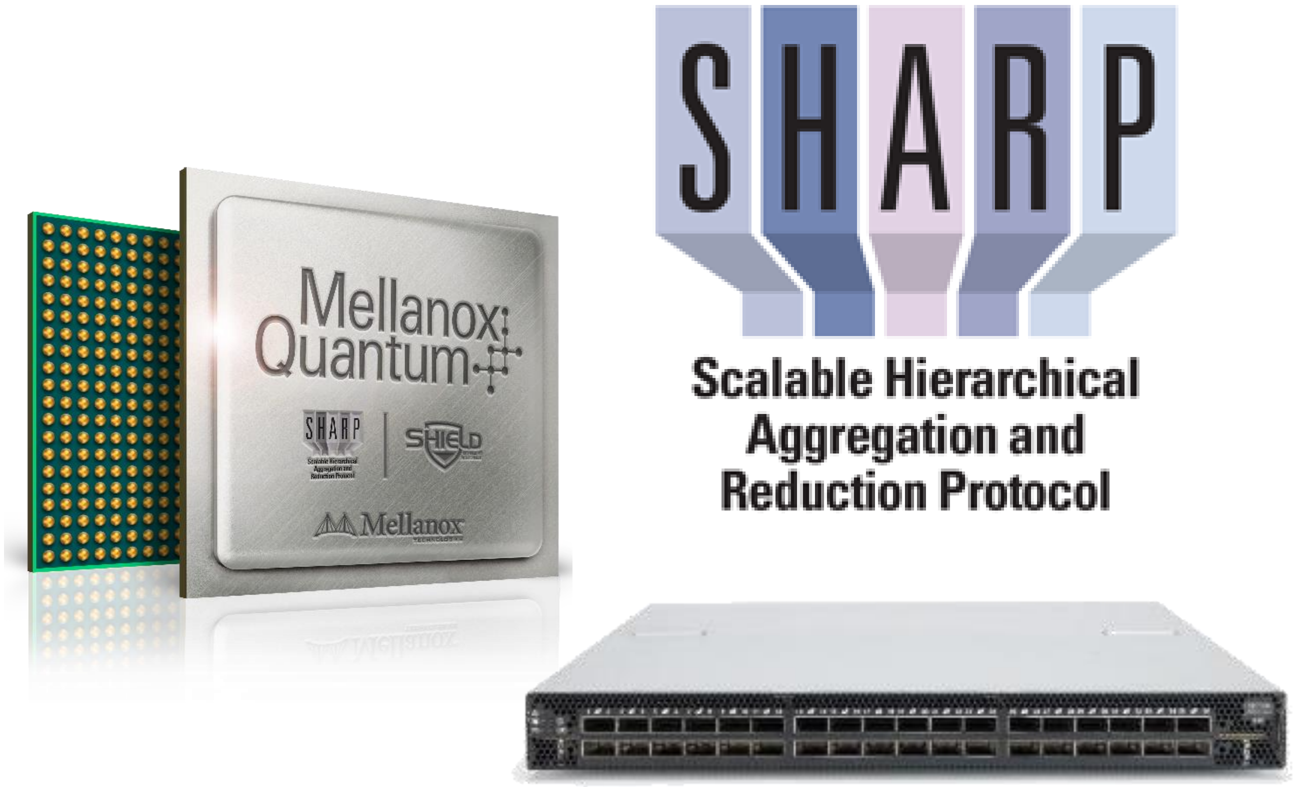
SHIELD

Connectivity

- Multi-Host Technology
- Socket-Direct Technology
- Enhanced Topologies



Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

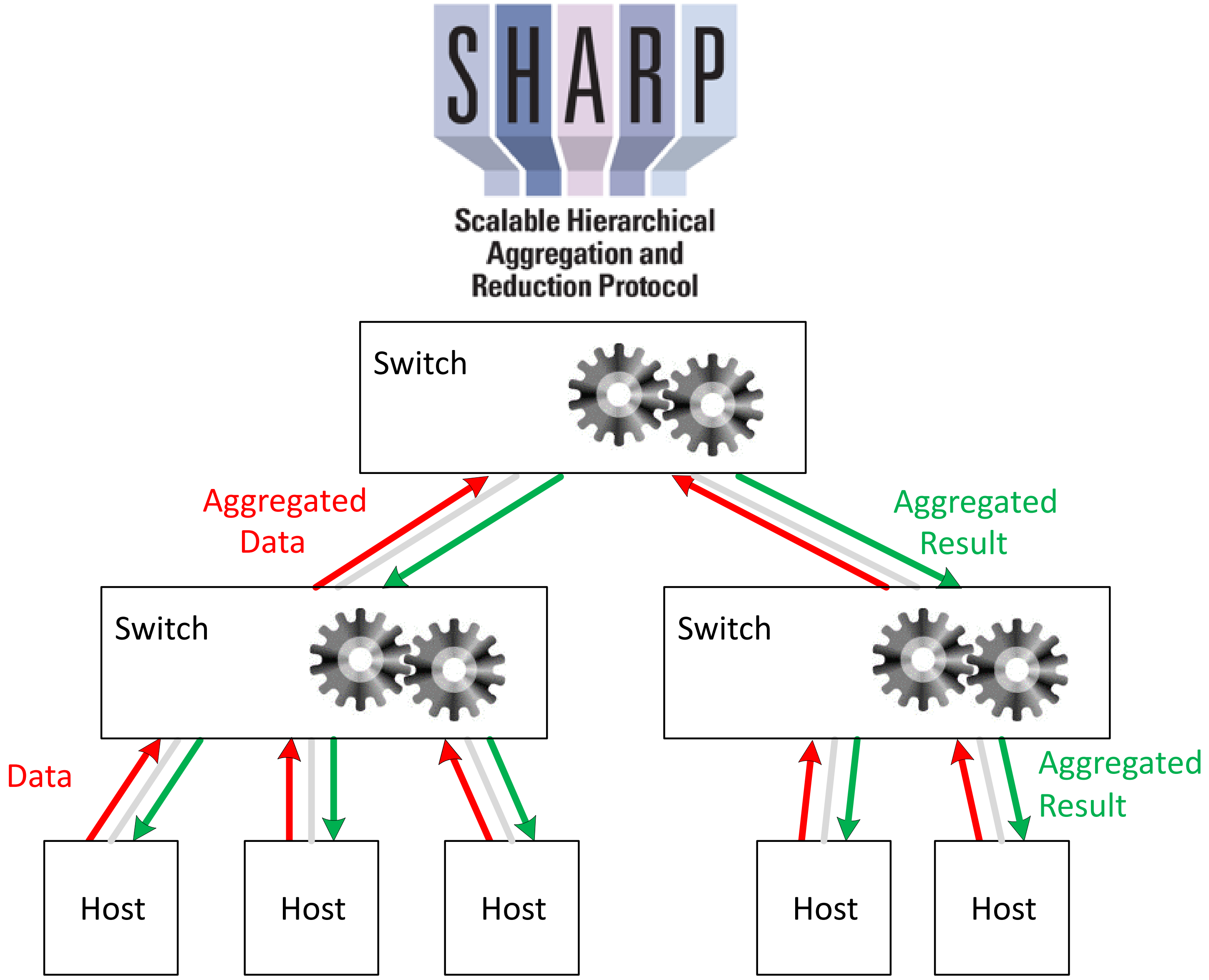


Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

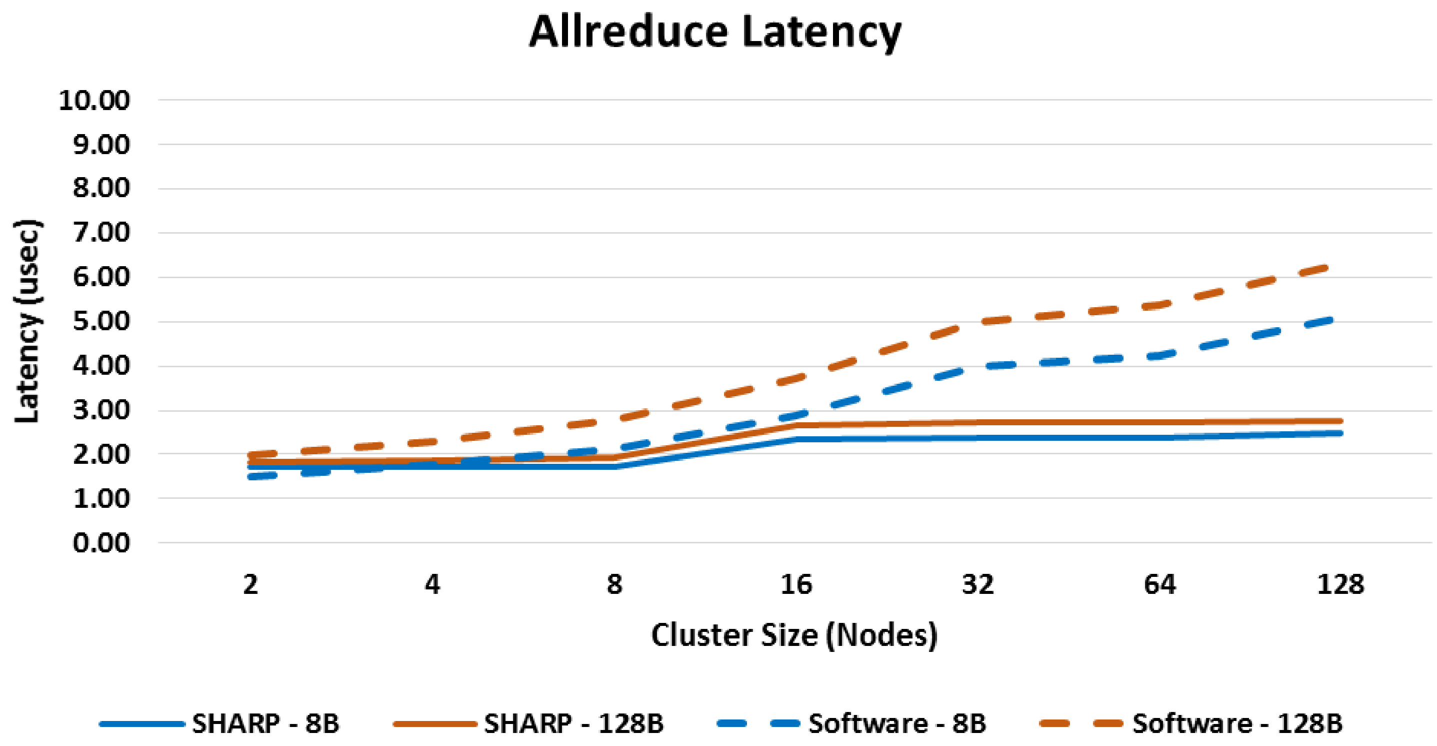
- Reliable Scalable General Purpose Primitive
  - In-network Tree based aggregation mechanism
  - Large number of groups
  - Multiple simultaneous outstanding operations

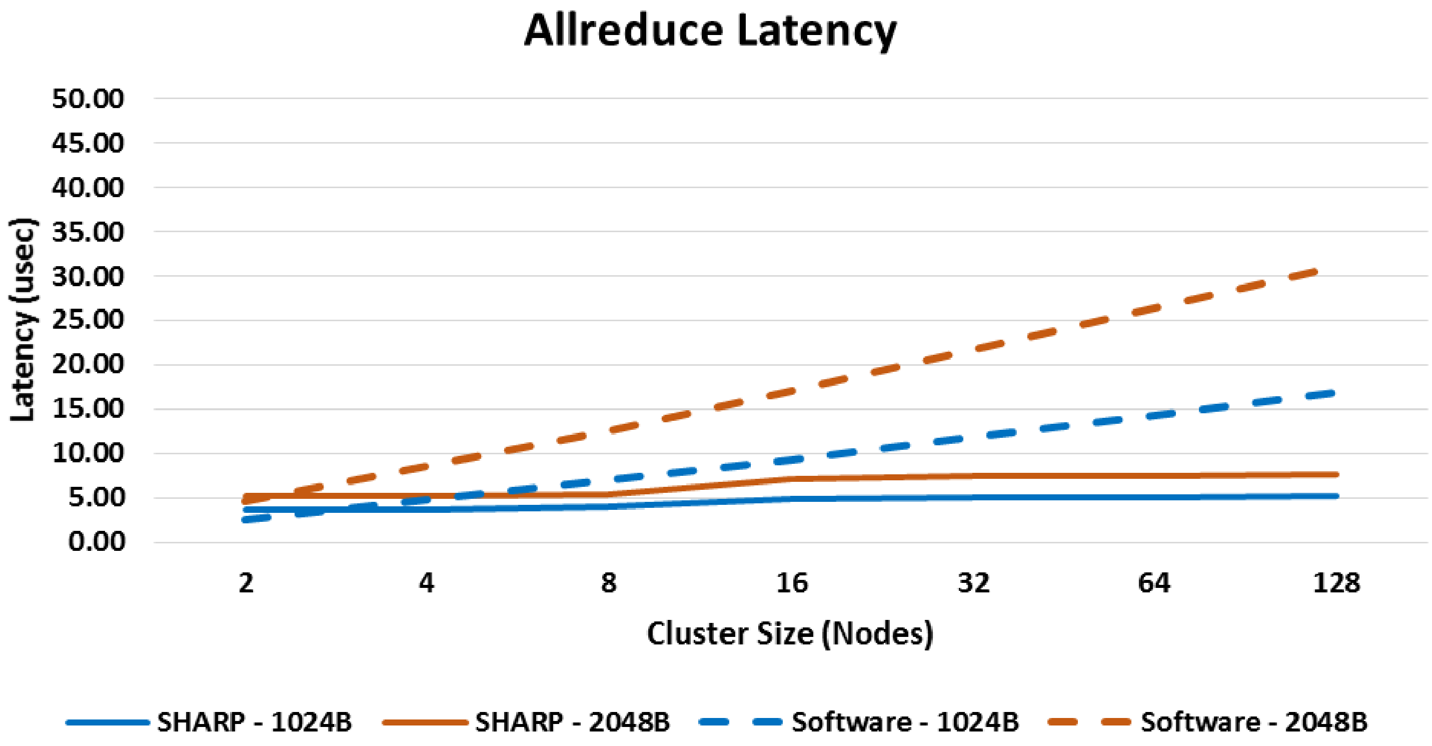
- Applicable to Multiple Use-cases
  - HPC Applications using MPI / SHMEM
  - Distributed Machine Learning applications

- 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)



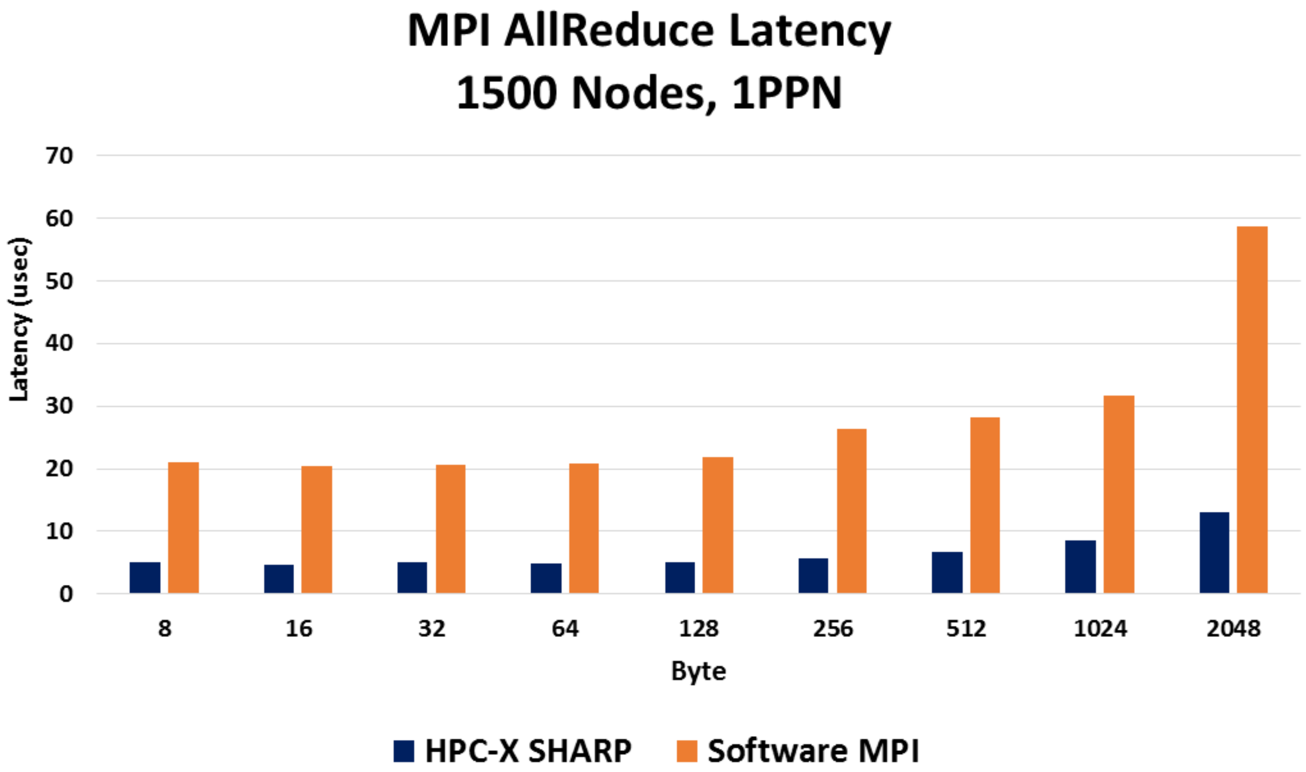


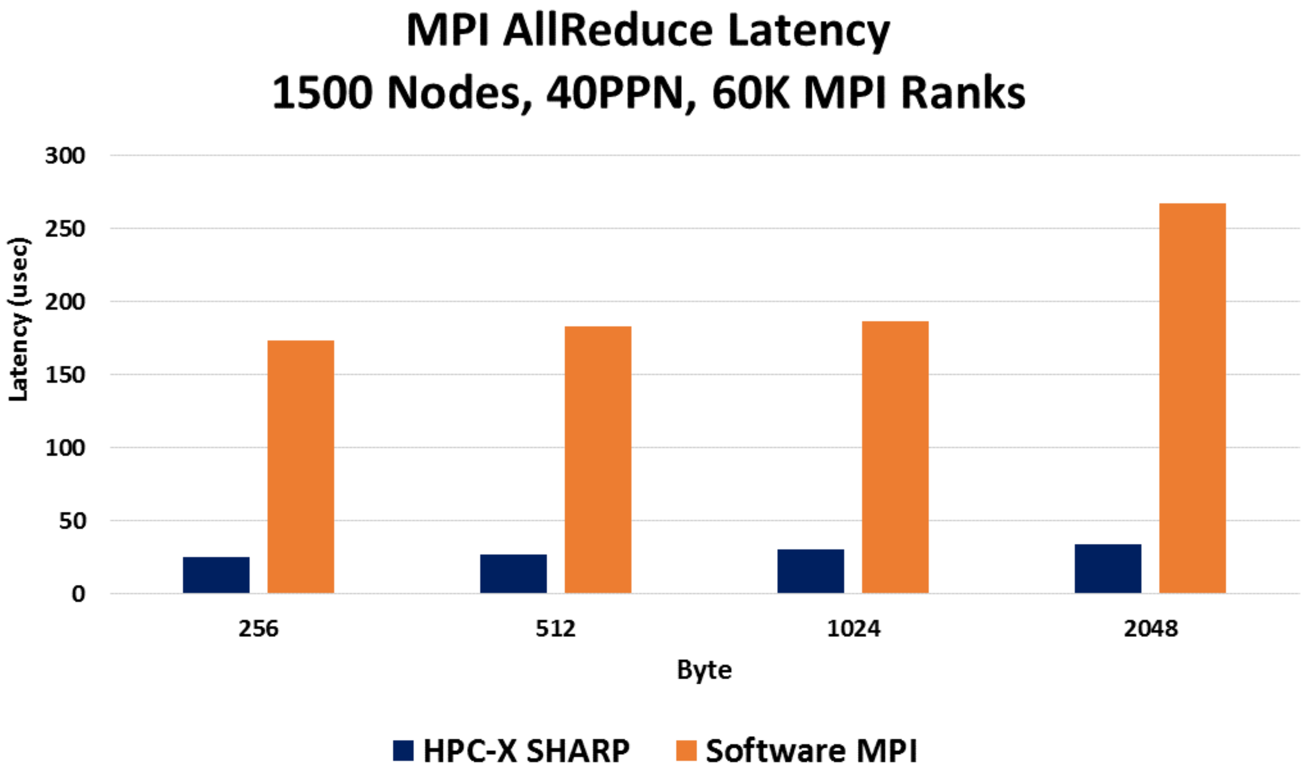
Scalable Hierarchical
Aggregation and
Reduction Protocol

SHARP enables 75% Reduction in Latency
Providing Scalable Flat Latency

SHARP AllReduce Performance Advantages

1500 Nodes, 60K MPI Ranks, Dragonfly+ Topology



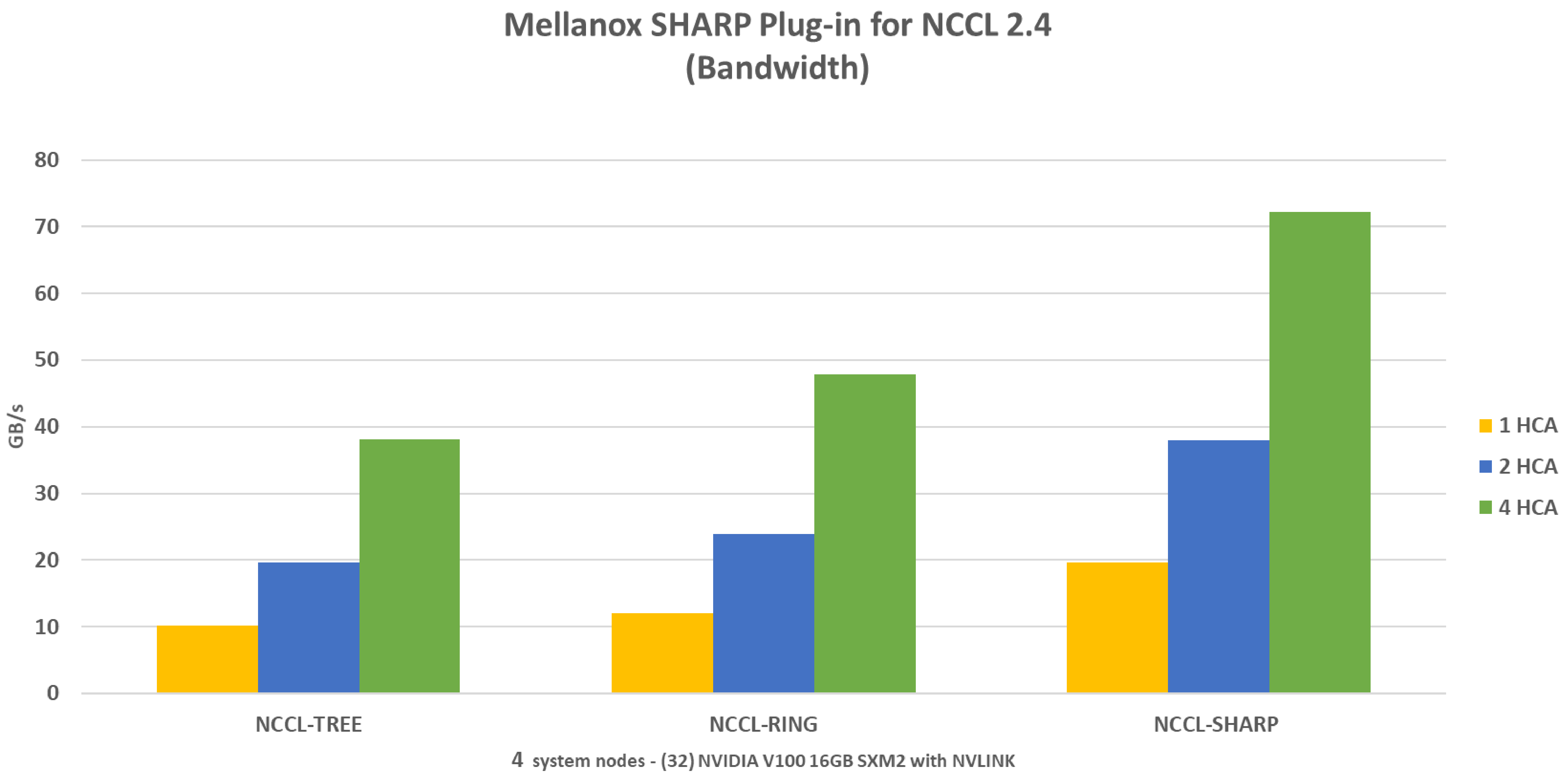


SHARP

Scalable Hierarchical Aggregation and Reduction Protocol

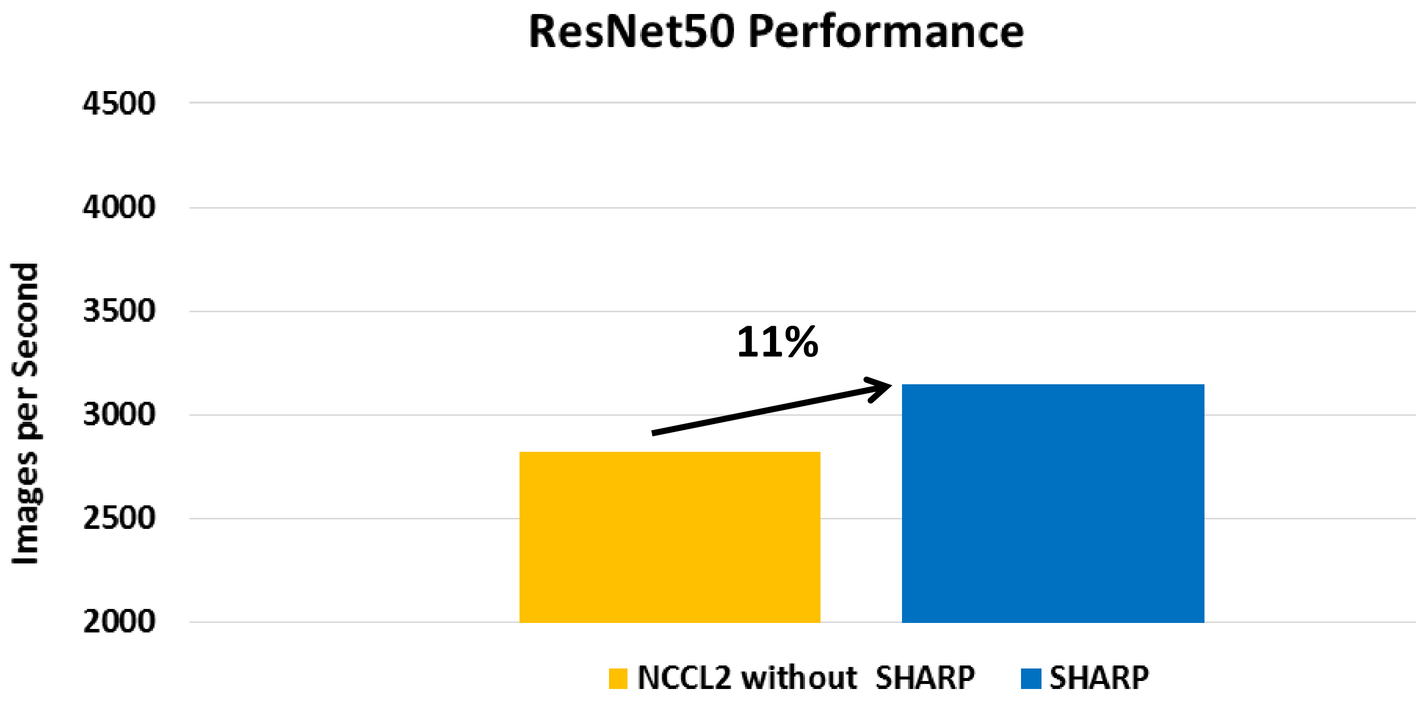
SHARP Enables Highest Performance

NCCL-SHARP Delivers Highest Performance

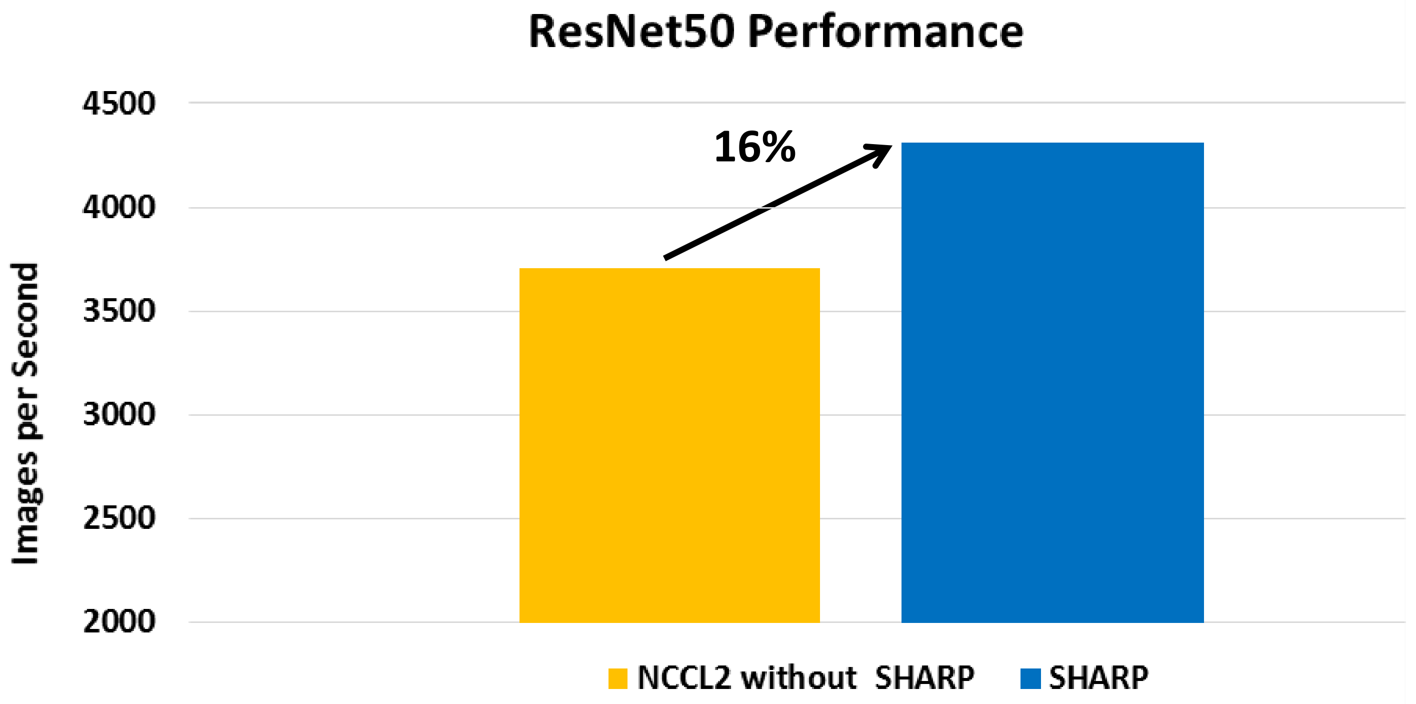


SHARP Performance Advantage for AI

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



8 Nodes, 16 GPUs, InfiniBand

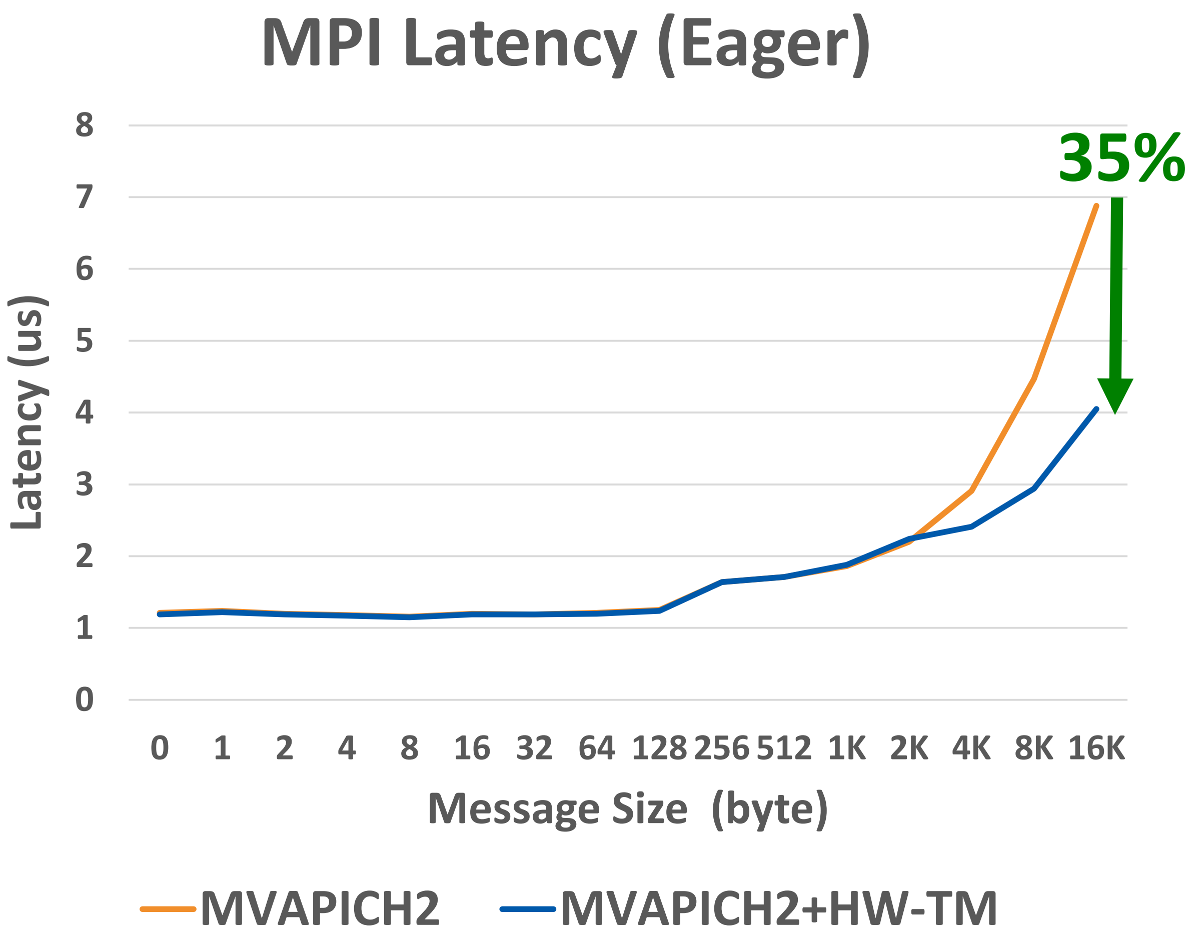


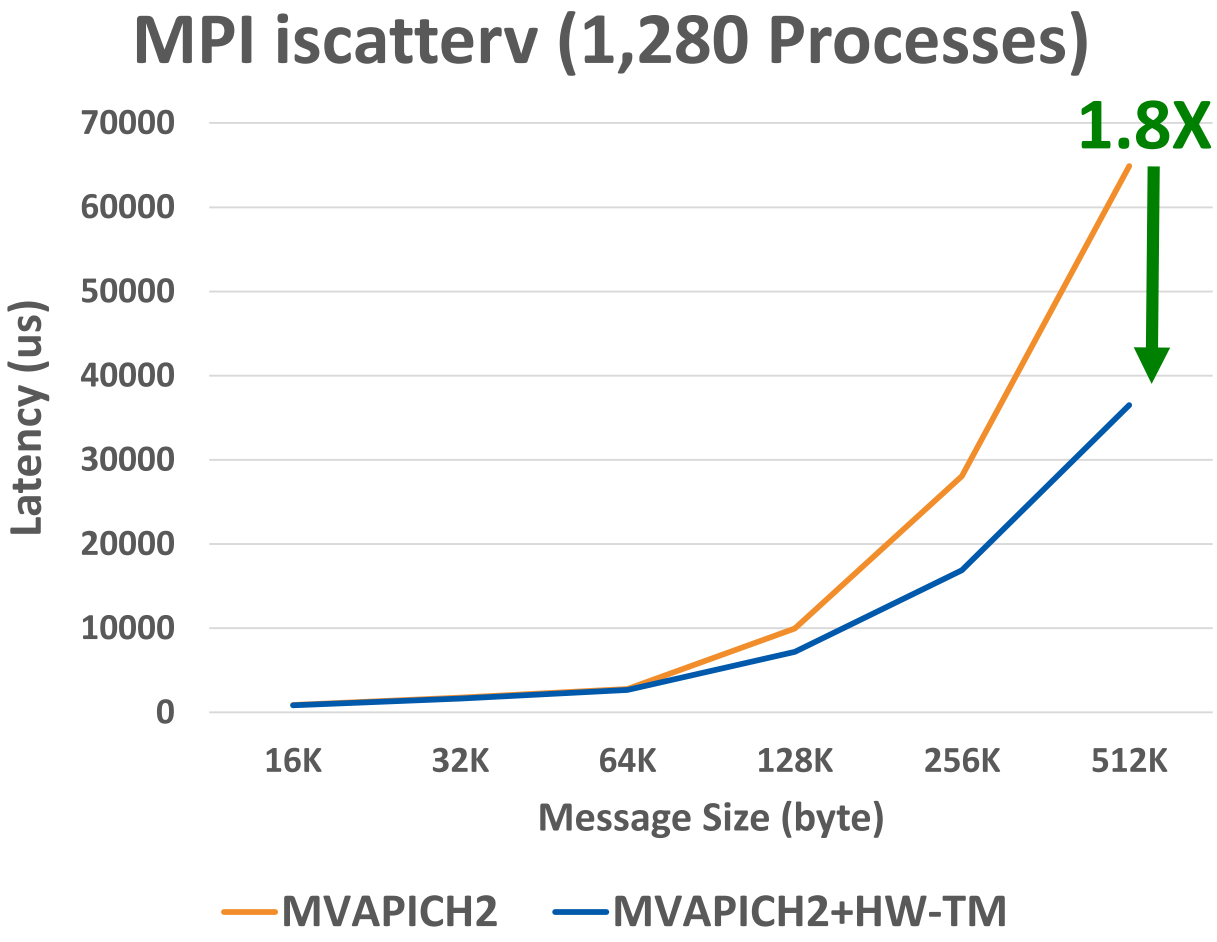
8 Nodes, 22 GPUs, InfiniBand

P100 NVIDIA GPUs, RH 7.5, Mellanox OFED 4.4, HPC-X v2.3, TensorFlow v1.11, Horovod 0.15.0

MPI Tag Matching Hardware Engine

Tag Matching Hardware Engine Performance Advantage



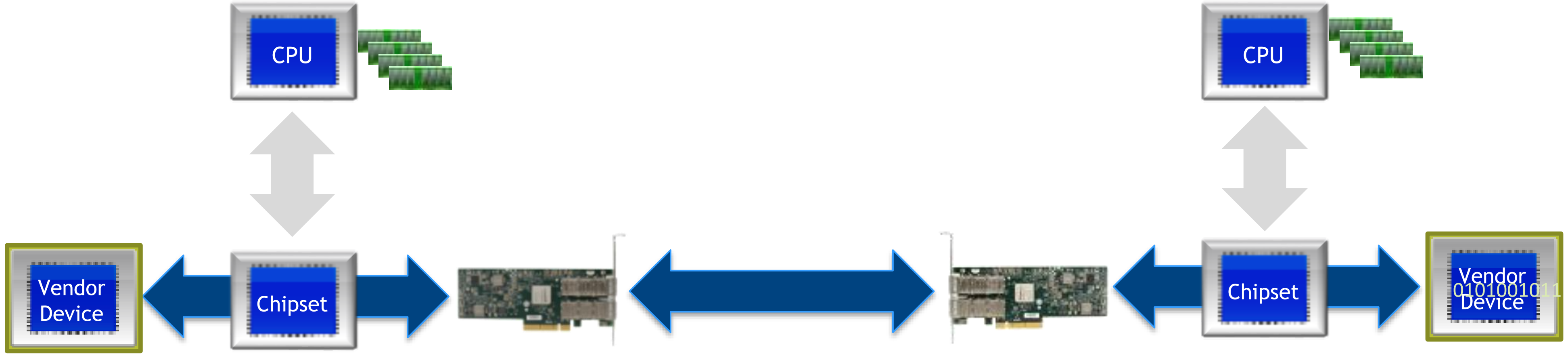


Courtesy of Dhabaleswar K. (DK) Panda
Ohio State University

GPUDirect

Mellanox PeerDirect™ Technology

- Purpose-built for acceleration of Deep Learning
- Provides significant decrease in communication latency for acceleration devices
- Peer-to-peer communications between Mellanox adapters and third-party devices
- Enables GPUDirect™ RDMA, GPUDirect™ ASYNC, ROCm and others

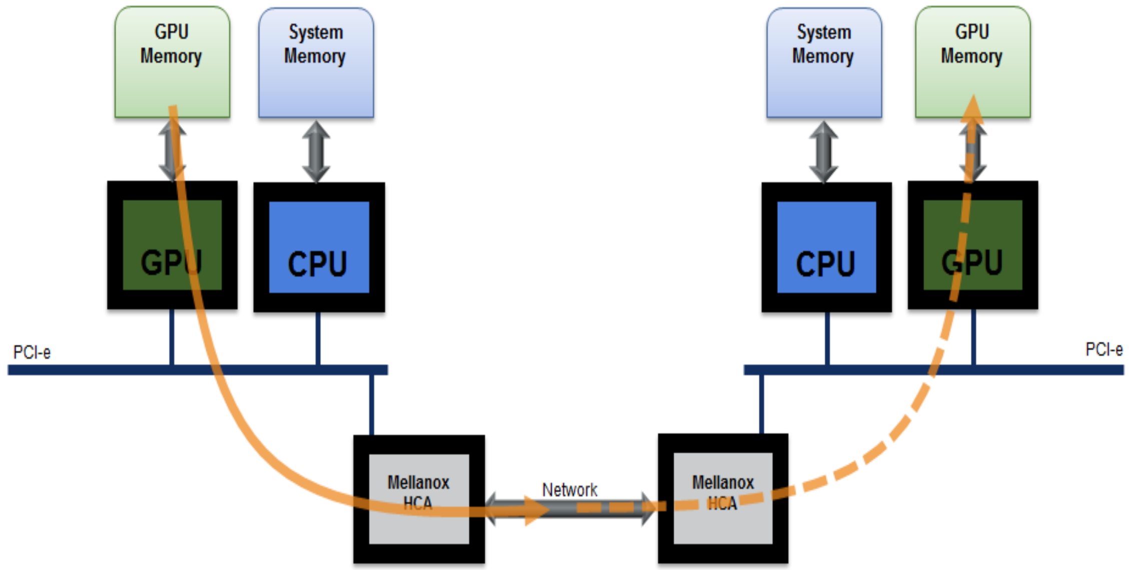


Designed for Deep Learning Acceleration

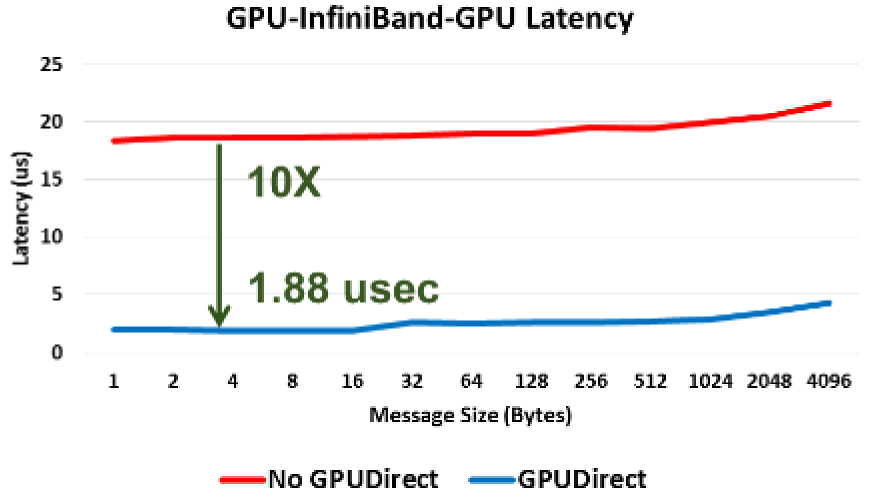
10X Higher Performance with GPUDirect™ RDMA

- Accelerates HPC and Deep Learning performance
- Lowest communication latency for GPUs

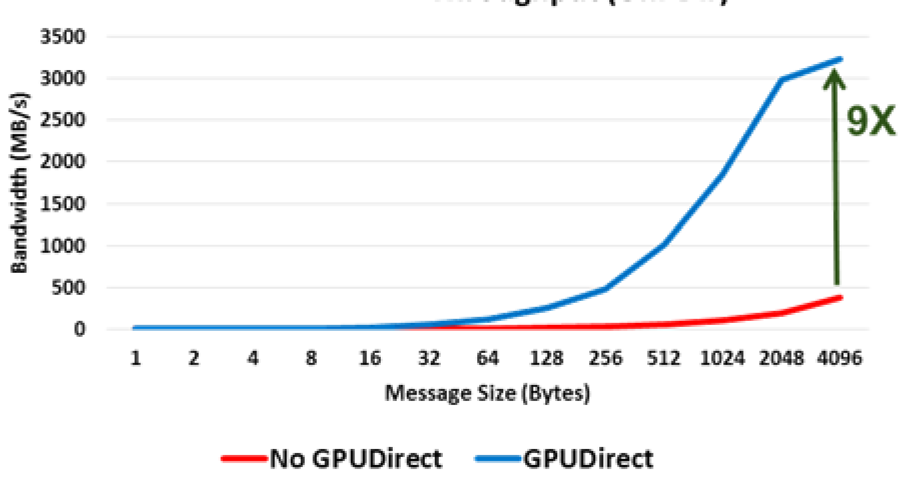
GPUDirect™ RDMA



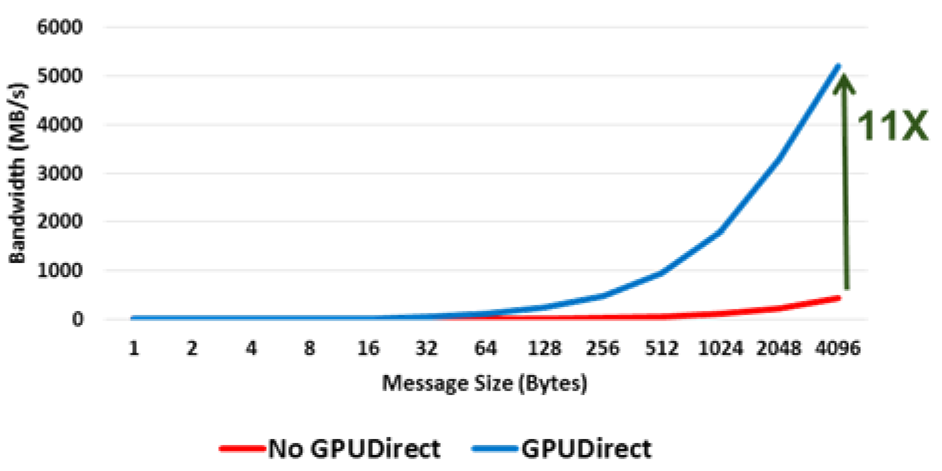
GPU-InfiniBand-GPU Latency



GPU-InfiniBand-GPU Throughput (Uni-Dir)



GPU-InfiniBand-GPU Throughput (Bi-Dir)

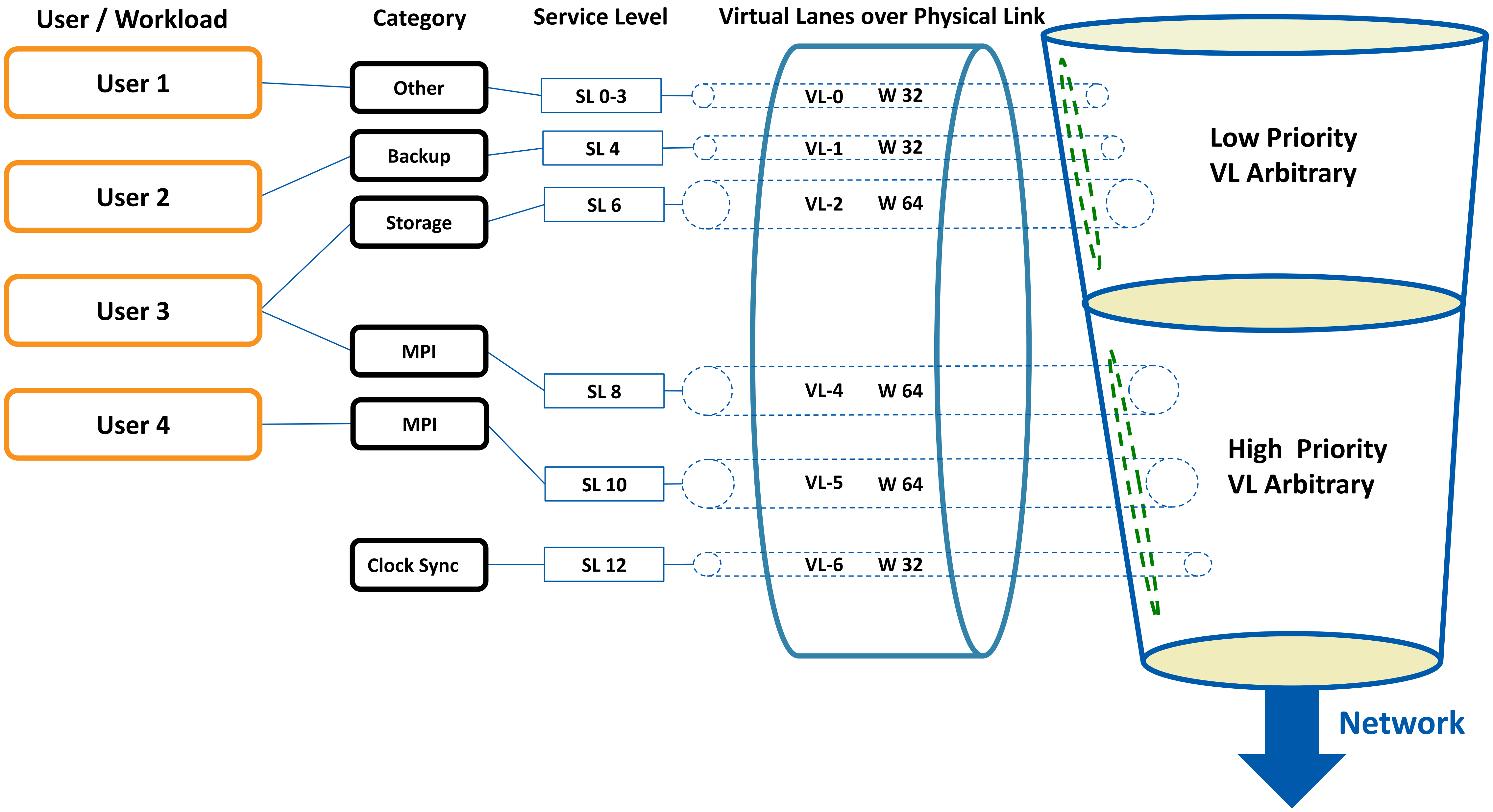


MVAPICH

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

Quality of Service

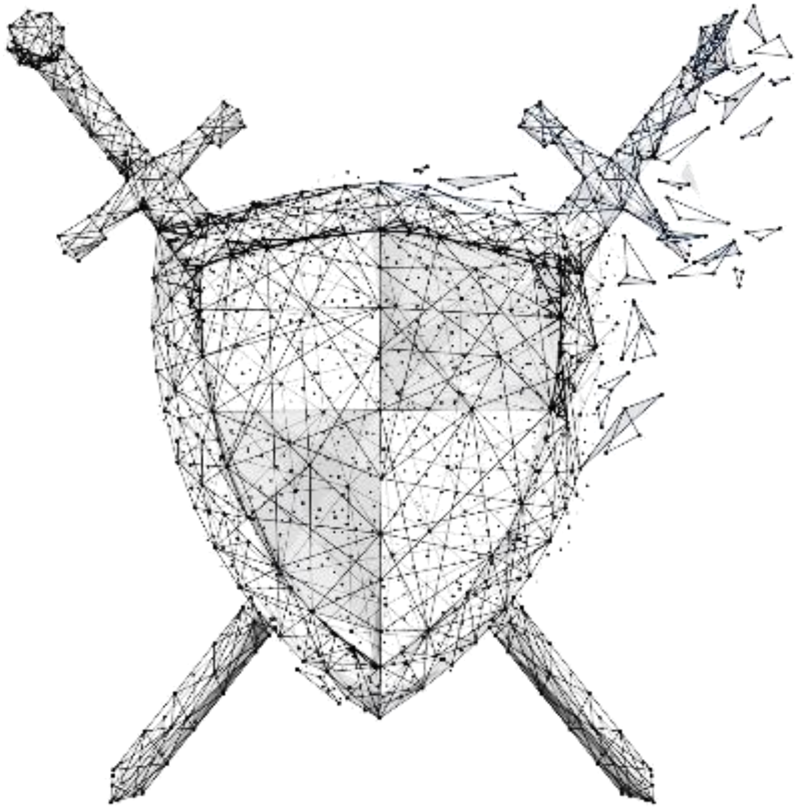
InfiniBand Quality of Service



SHIELD
Self Healing Technology

SHIELD

SELF-HEALING
INTERCONNECT

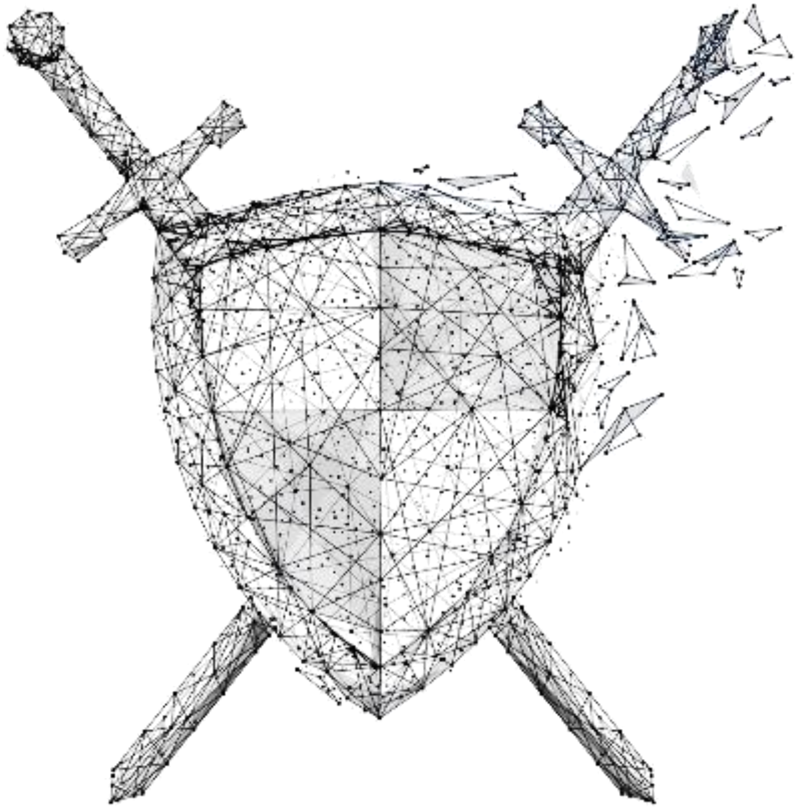


SHIELD - Self Healing Technology

Enables Unbreakable Data Centers

- The ability to overcome network failures, locally, by the switches
- Software-based solutions suffer from long delays detecting network failures
  - 5-30 seconds for 1K to 10K nodes clusters
  - Accelerates network recovery time by 5000X
  - The higher the speed or scale the greater the recovery value
- Available with EDR and HDR switches and beyond

SHIELD
SELF-HEALING
INTERCONNECT



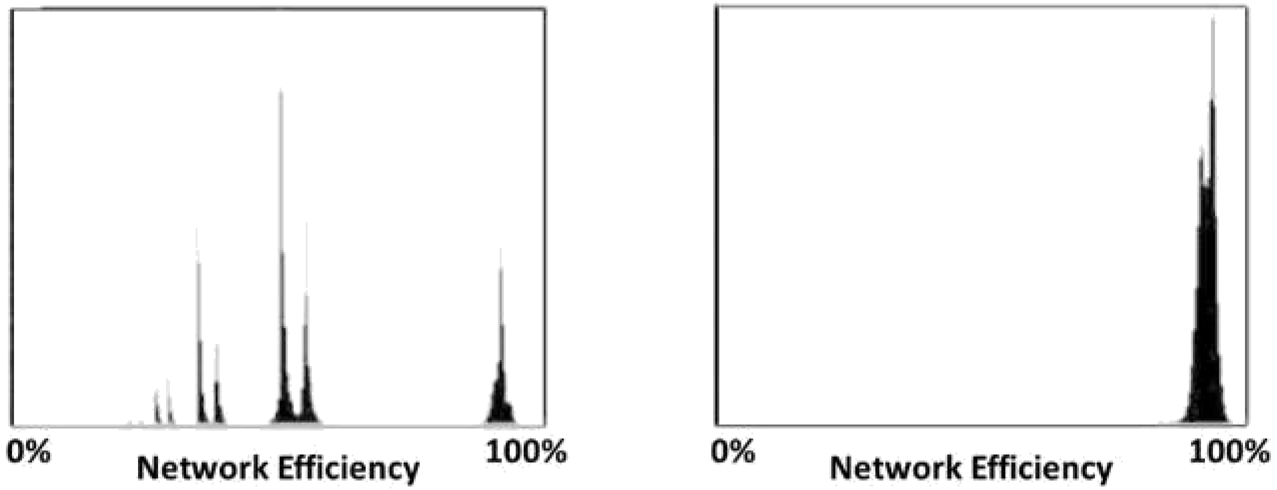
Adaptive Routing

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

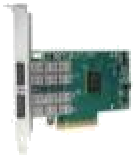
Adaptive Routing

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)

HDR InfiniBand

Highest-Performance 200Gb/s InfiniBand Solutions

Adapters	ConnectX-6	200Gb/s Adapter, 0.6us latency 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, <90ns Latency	
SoC	BlueField	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	

ConnectX-6 HDR InfiniBand Adapter

Leading Connectivity

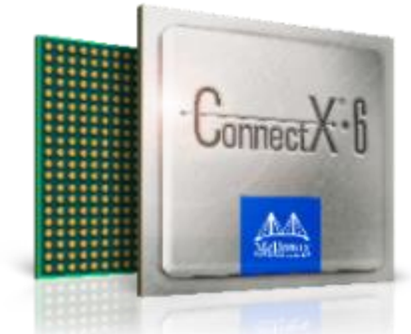
- 200Gb/s InfiniBand and Ethernet
  - HDR, HDR100, EDR (100Gb/s) and lower speeds
  - 200GbE, 100GbE and lower speeds
- Single and dual ports

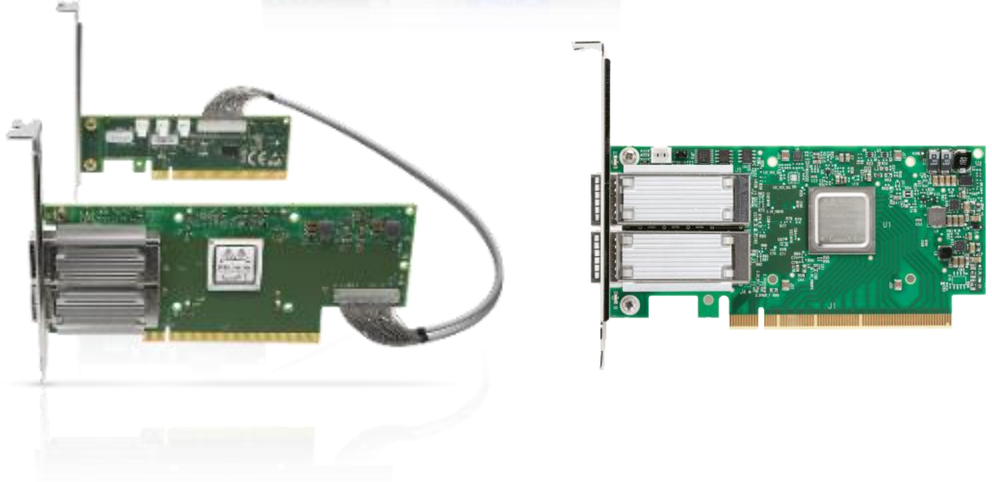
Leading Performance

- 200Gb/s throughput, 0.6usec latency, 215 million message per second
- PCIe Gen3 / Gen4, 32 lanes
- Integrated PCIe switch
- Multi-Host - up to 8 hosts, supporting 4 dual-socket servers

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

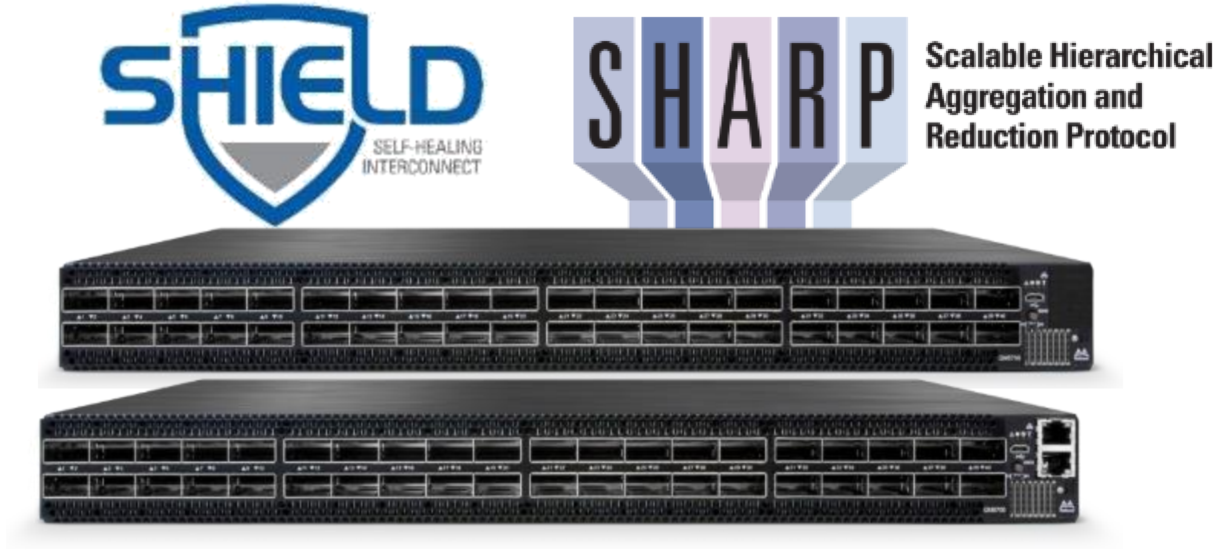




HDR InfiniBand Switches

40 QSFP56 ports

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


SHIELD
SELF-HEALING INTERCONNECT
SHARP
Scalable Hierarchical Aggregation and Reduction Protocol

800 QSFP56 ports

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

Real Time Network Visibility

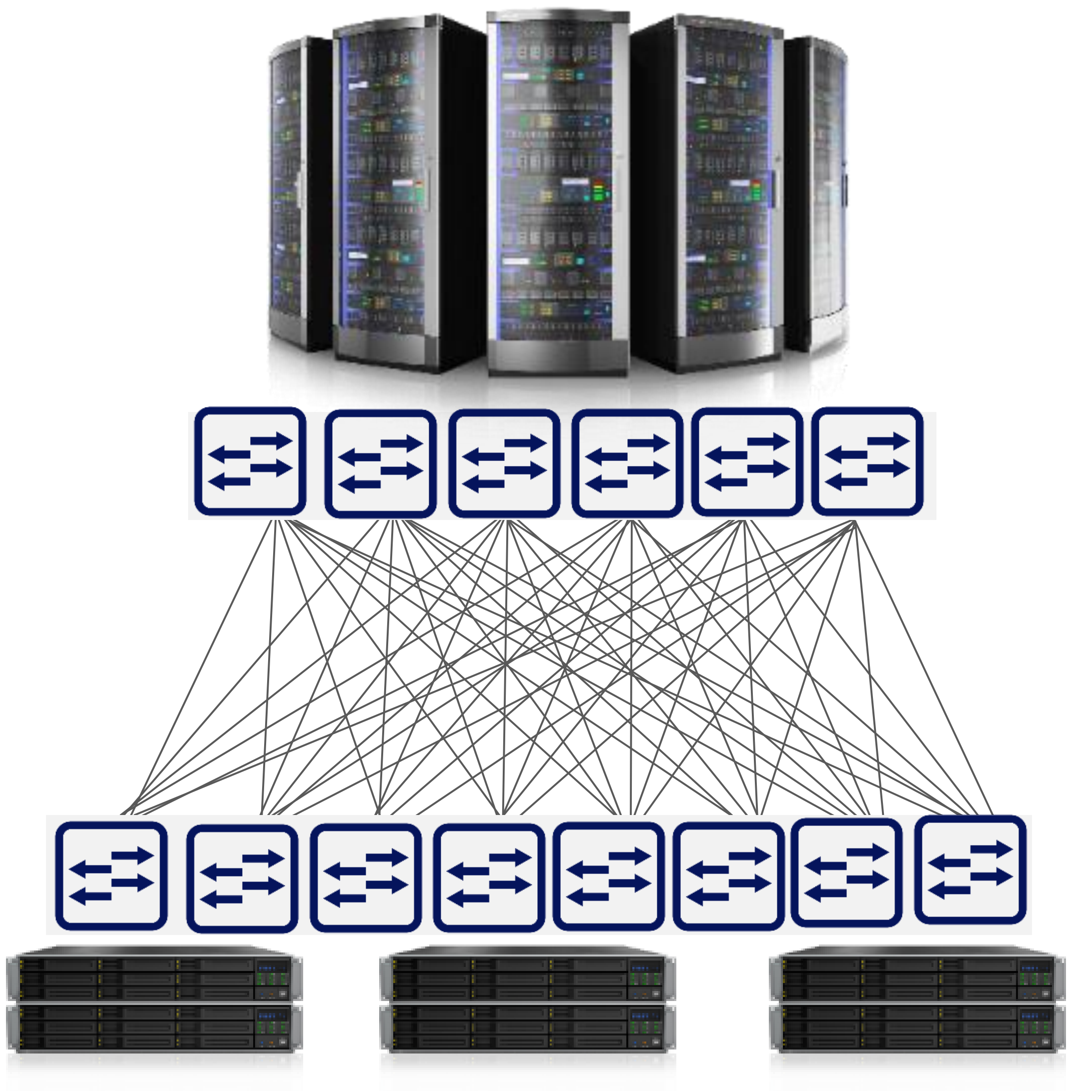
Built-in Hardware Sensors for Rich Traffic Telemetry and Data Collection

Advanced monitoring for troubleshooting

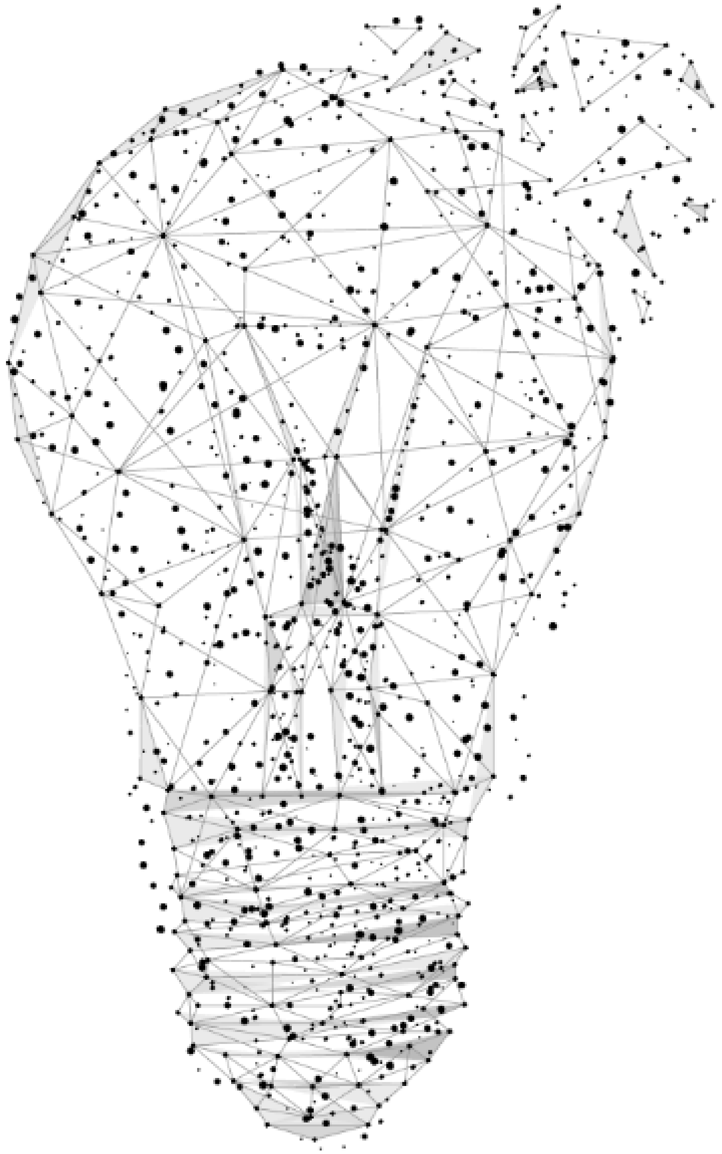
- 8 mirror agents triggered by congestion, buffer usage and latency
- Measure queue depth using histograms (64ns granularity)

Network status/health in real time

- Buffer snapshots
- Congestion notifications and buffers status

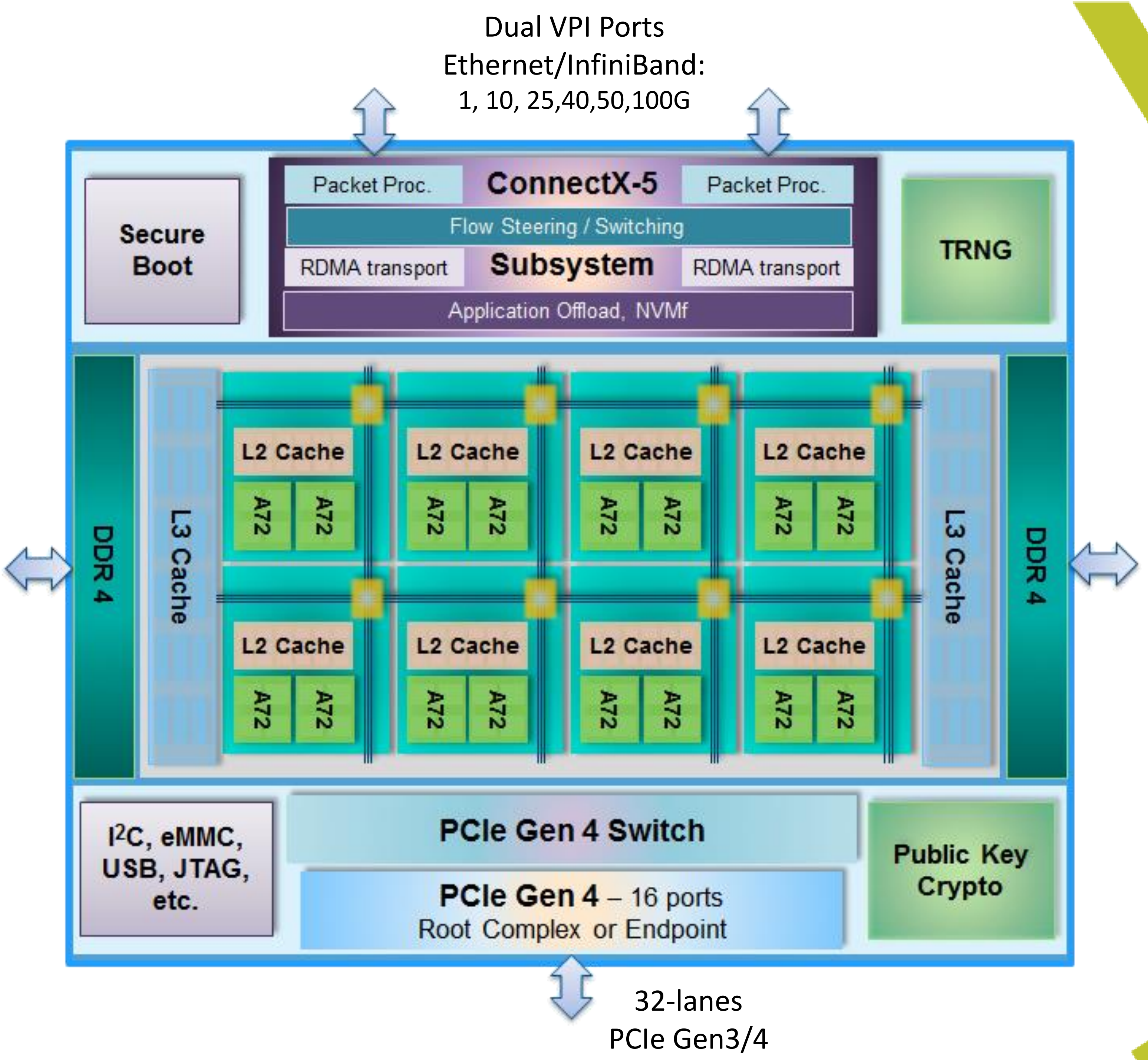


BlueField SoC Advantages and Platforms



BlueField Block Diagram

- Tile Architecture - 16 ARM® A72 CPUs subsystem
  - SkyMesh™ fully coherent low-latency interconnect
  - 8MB L2 Cache, 8 Tiles
- Dual Port 100g IO Controller, based on ConnectX-5
  - Dual 100Gb/s Ethernet/InfiniBand, compatible with ConnectX-5
  - NVMe-oF hardware accelerator
  - High-end Networking Offloads: RDMA, Erasure Coding, T10-DIF
- Fully Integrated PCIe switch
  - 32 Bifurcated PCI Gen3/4 lanes (up to 200Gb/s)
  - Root Complex or Endpoint modes
  - 2x16, 4x8, 8x4 or 16x2 configurations
- Memory Controllers
  - 2x Channels DDR4 Memory Controllers w/ ECC
  - NVDIMM-N Support



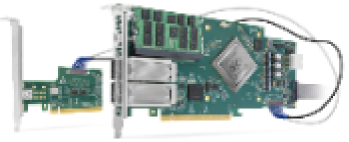
BlueField for Smart Solutions

BlueField SoC (System on Chip)

- SoC: Compute, networking and PCIe connectivity
  - Dual port VPI EDR/100GbE
  - 16 Arm cores
  - 32 lanes of PCIe switch gen3/4

Storage Solutions

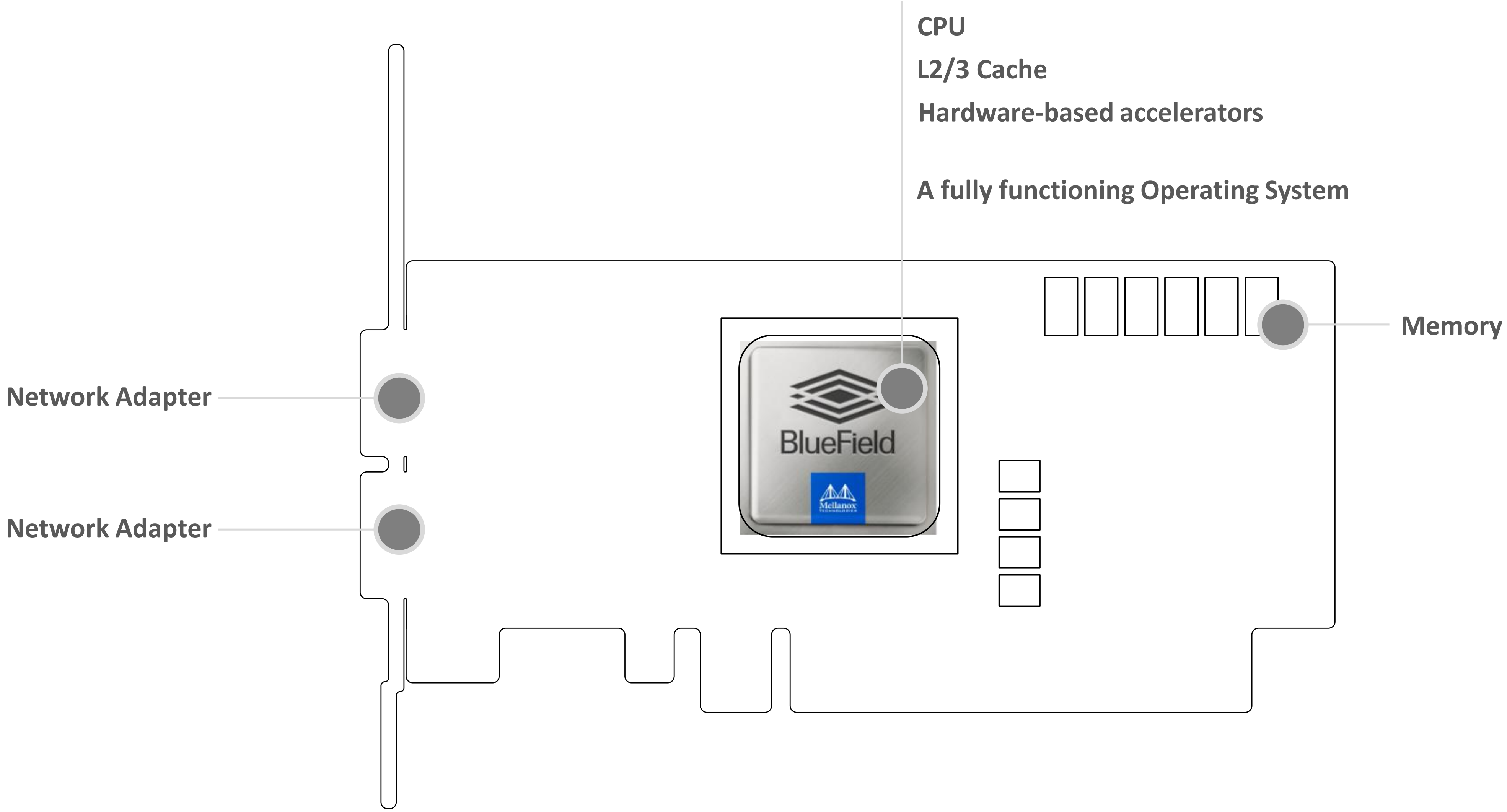
- NVMe-based storage platforms
  - RDMA, NVMe over Fabrics, RAID, Signature offload
- Partner's solutions based on BlueField storage controller



Smart Adapters

- In-network computing and collective offloads
- Co-processor running proprietary smart algorithms
- Security and privacy algorithms

BlueField Smart Adapter is a Computer



CPU

L2/3 Cache

Hardware-based accelerators

A fully functioning Operating System

Memory

Network Adapter

Network Adapter

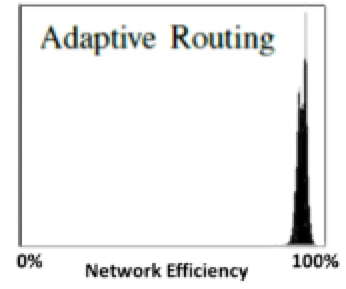
Highest Performance and Scalability for Exascale Platforms

Mellanox TECHNOLOGIES

OAK RIDGE National Laboratory

SHARP SHIELD SELF-HEALING

96% Network Utilization



Summit's MpiGraph

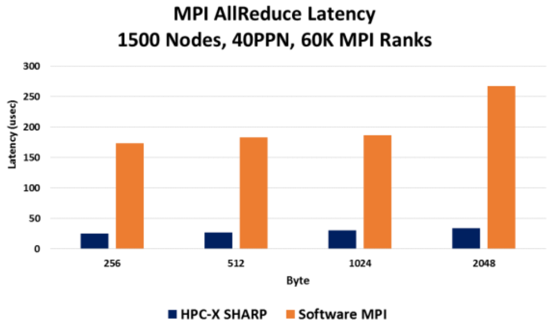
UNIVERSITY OF TORONTO

SciNet

SHARP

Niagara
Canada's Most-Powerful Research Supercomputer
Built to Fuel Canadian Innovation and Discovery
SciNet

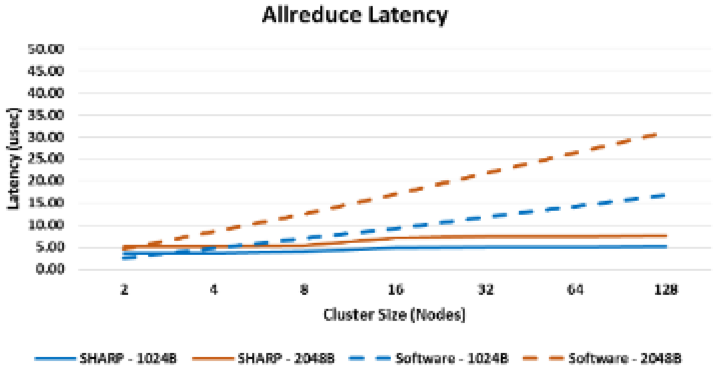
7X Higher Performance



SHARP

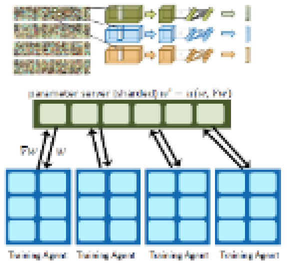
Scalable Hierarchical Aggregation and Reduction Protocol

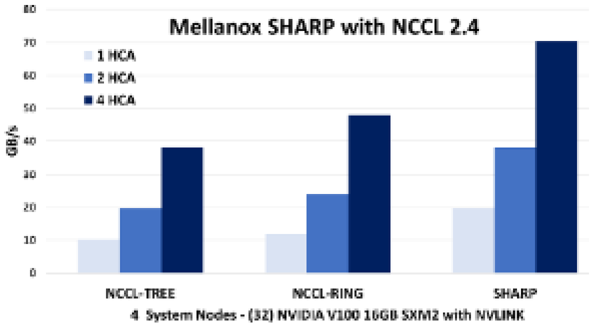
Flat Latency



Deep Learning

2X Higher Performance

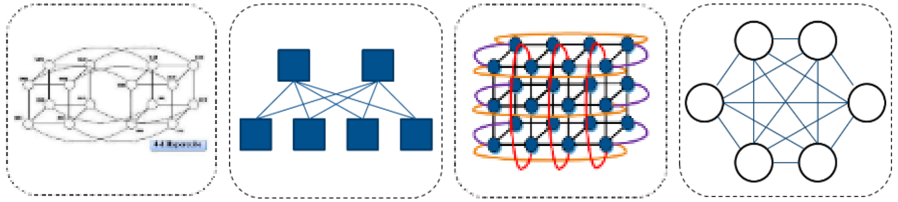


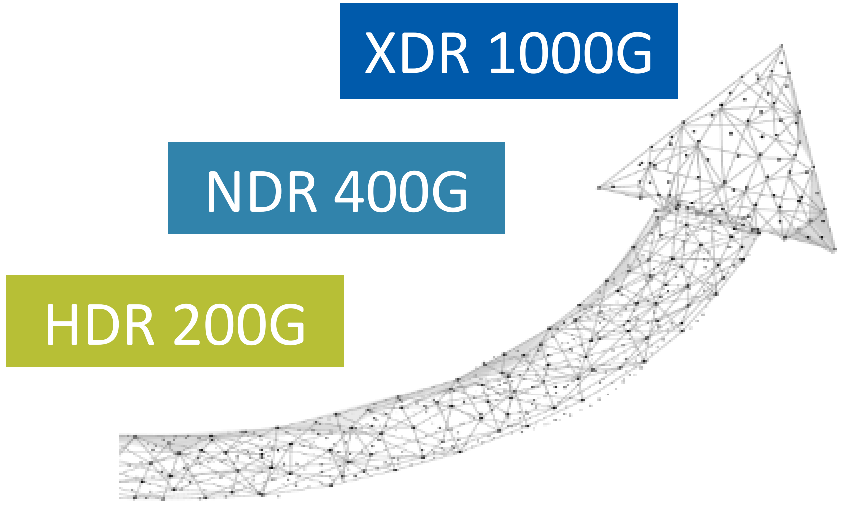


Mellanox SHARP with NCCL 2.4
GB/s
80
70
60
50
40
30
20
10
0
1 HCA
2 HCA
4 HCA
NCCL-TREE
NCCL-RING
SHARP
4 System Nodes - (32) NVIDIA V100 16GB SXM2 with NVLINK

SHIELD SELF-HEALING

5000X Higher Resiliency





XDR 1000G

NDR 400G

HDR 200G

Mellanox TECHNOLOGIES

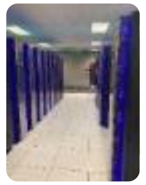
SUPERCONNECTING the #1 Supercomputers

TACC
TEXAS ADVANCED COMPUTING CENTER

FRONTERA

5 TOP500 The List.

MISSISSIPPI STATE UNIVERSITY



NOAA

62 TOP500 The List.

CSC

FINNISH METEOROLOGICAL INSTITUTE

166 TOP500 The List.

筑波大学
University of Tsukuba

264 TOP500 The List.

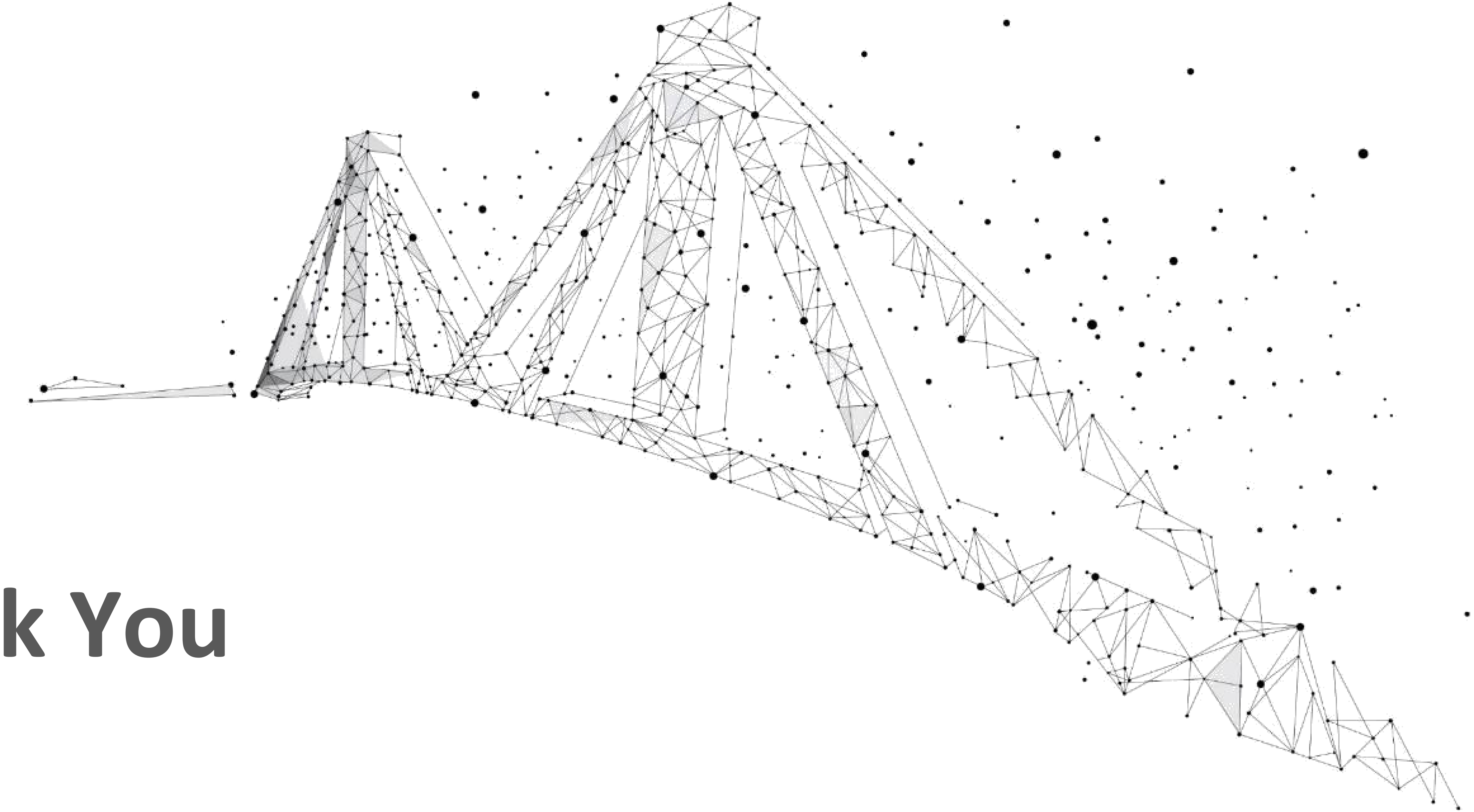
UNIVERSITY OF MICHIGAN

World's First
HDR InfiniBand
Supercomputer

सी-डैक
CDAC

India's National
Supercomputing
Program

HDR 200G InfiniBand Accelerated Supercomputers



Thank You