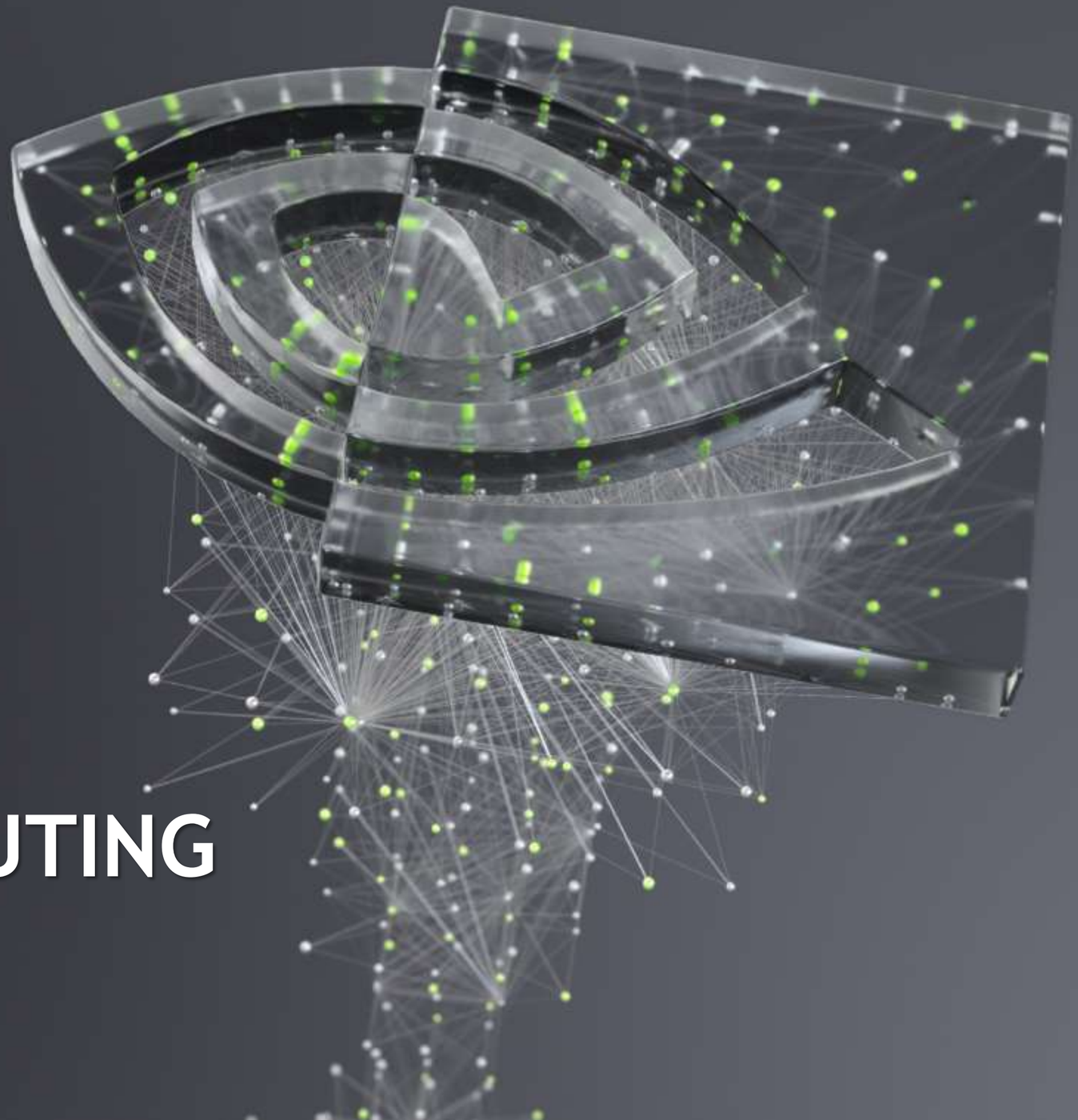
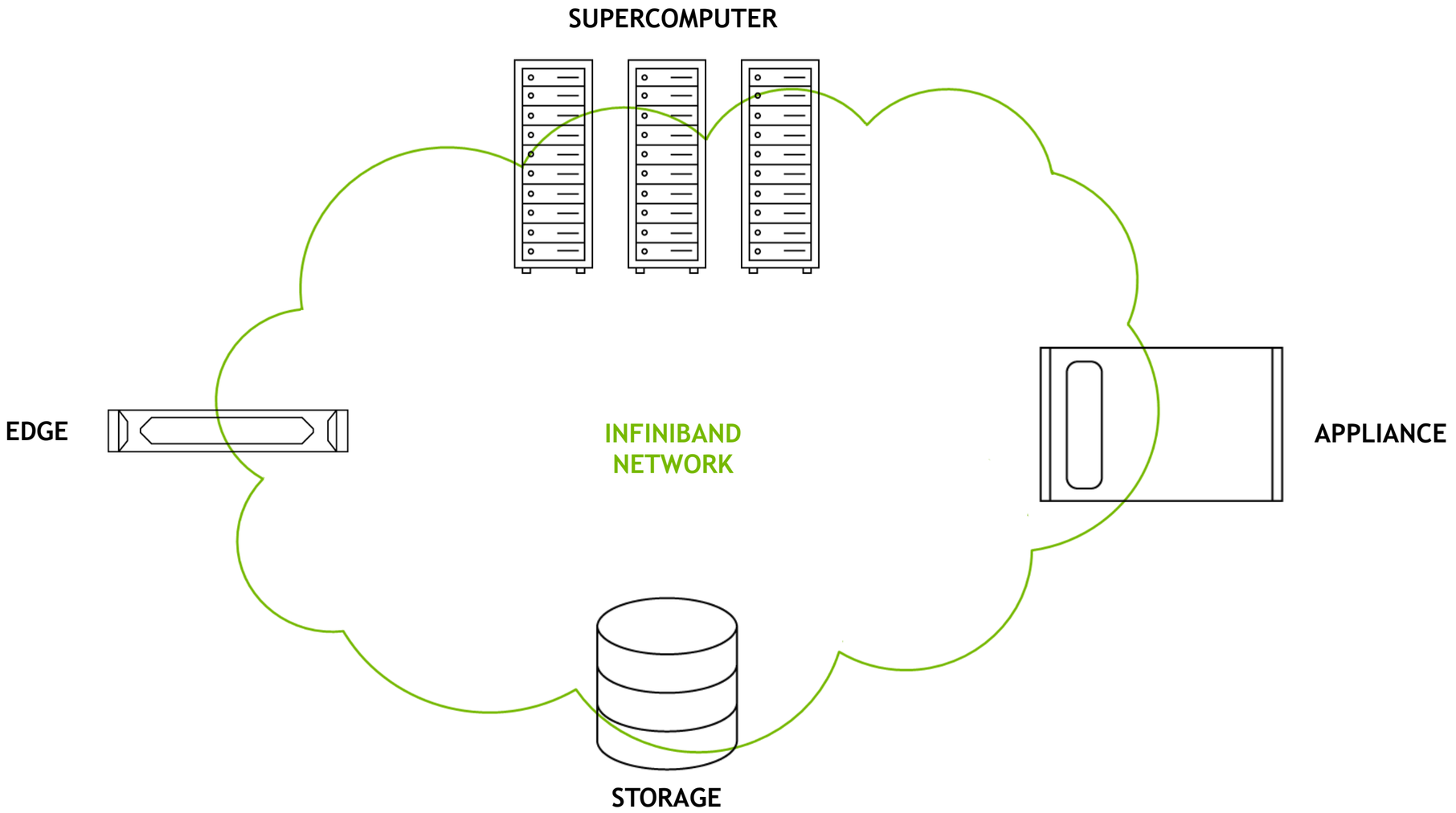


INFINIBAND IN-NETWORK COMPUTING

MUG, August 2020

THE NEW SCIENTIFIC COMPUTING WORLD



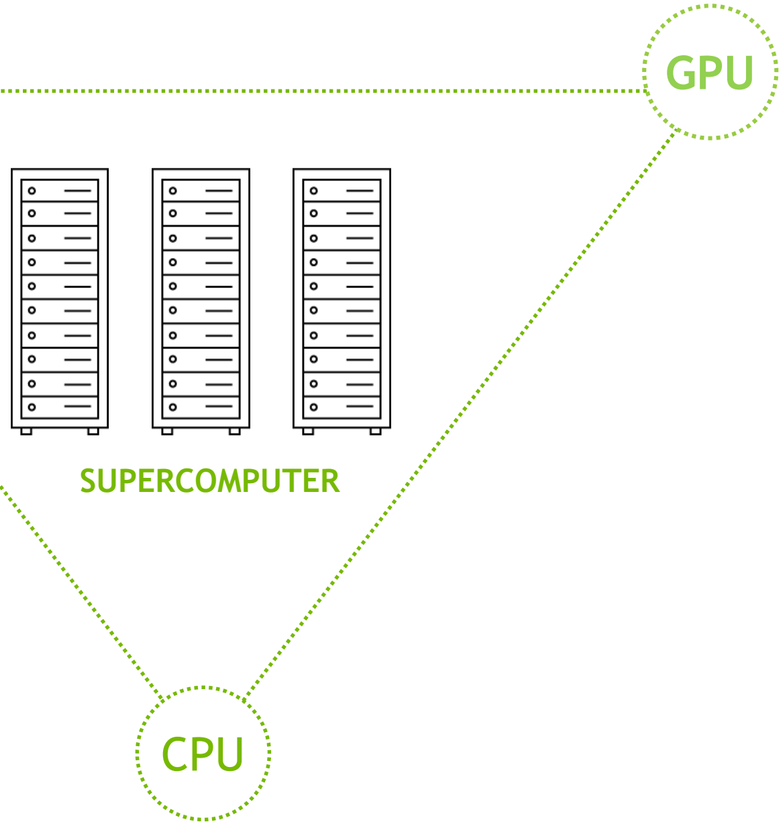
IN-NETWORK COMPUTING ACCELERATED SUPERCOMPUTING

INFINIBAND

Data Processing Unit

Software Defined
Hardware-Accelerated
In-network computing

Pre-configured DPUs
Programmable DPUs



IN-NETWORK COMPUTING ACCELERATED SUPERCOMPUTING

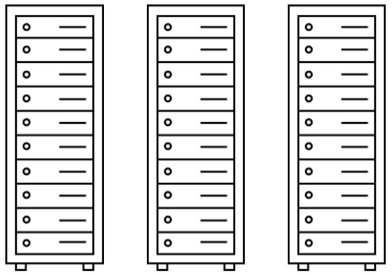
INFINIBAND

Data Processing Unit

DPU

Software Defined
Hardware-Accelerated
In-network computing

Pre-configured DPUs
Programmable DPUs



SUPERCOMPUTER

Programmable

DPU cores, MMU
Data pre-processing
User-defined algorithms

Pre-Configured

SHARP (data reductions)
MPI Tag-Matching
SHIELD (self healing resiliency)
NVMe over fabric
Data security and tenant isolations

Speed of Light

200G end-to-end, extremely low latency
RDMA, GPU Direct RDMA, GPU Direct storage
Enhanced Adaptive Routing and Congestion Control
Smart topologies

IN-NETWORK COMPUTING ACCELERATED SUPERCOMPUTING

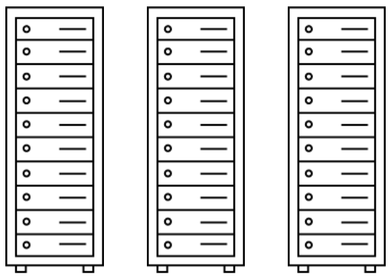
INFINIBAND

Data Processing Unit

DPU

Software Defined
Hardware-Accelerated
In-network computing

Pre-configured DPUs
Programmable DPUs



SUPERCOMPUTER

UCF Unified Communication Framework

OpenSNAPI

Applications / Services Acceleration

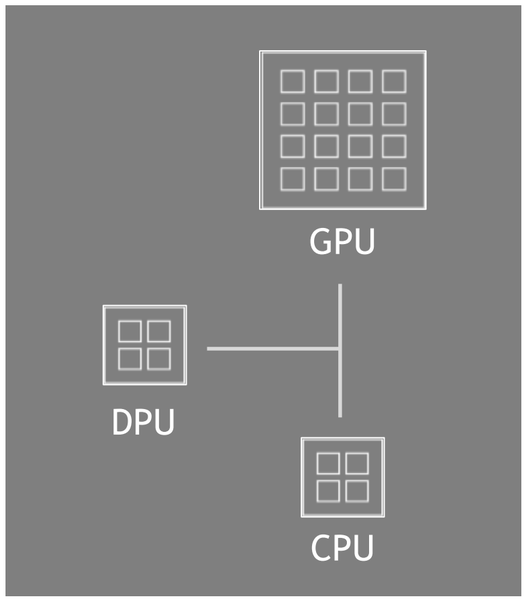
Applications and Services

Switching Routing | Analytics, Telemetry | Elastic Storage | ..... | QoS, Isolation, Protection

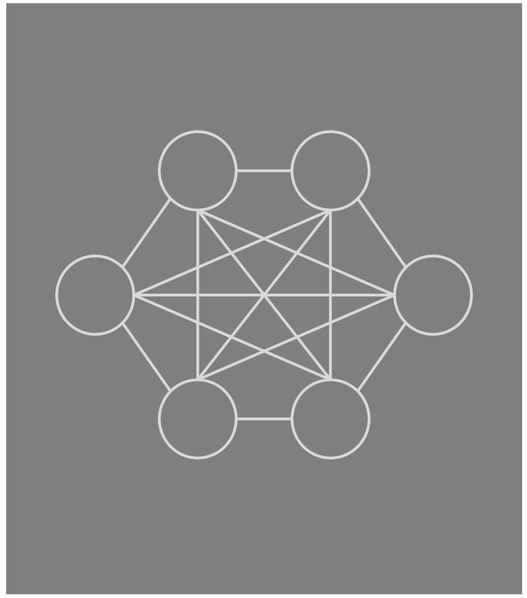
Hardware Acceleration SDK Libraries

Communication | Computation | Storage | ..... | Security

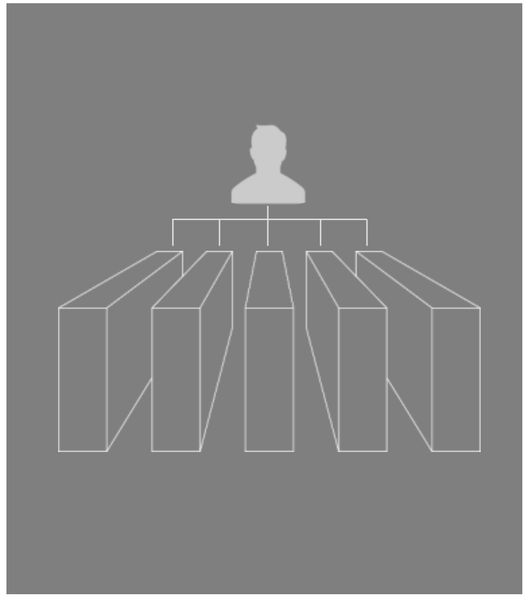
INFINIBAND TECHNOLOGY FUNDAMENTALS



Smart End-Point



Architected to Scale



Centralized Management

Standard

HDR 200G INFINIBAND ACCELERATES NEXT GENERATION HPC AND AI SUPERCOMPUTERS (EXAMPLES)

Australian National University

9 PetaFLOPS
3K HDR Nodes
Dragonfly+ Topology

CSC

FINNISH METEOROLOGICAL INSTITUTE

7.5 PetaFLOPS
2K HDR Nodes
Dragonfly+ Topology

DKRZ
DEUTSCHES KLIMARECHENZENTRUM

3K HDR Nodes
16 PetaFLOPS
Dragonfly+ Topology

ECMWF
EUROPEAN CENTRE FOR MEDIUM RANGE WEATHER FORECASTS

8K HDR Nodes
Dragonfly+ Topology

35.5 PetaFLOPS
2K HDR Nodes
Fat-Tree Topology

HLRS
High-Performance Computing Center | Stuttgart

23 PetaFLOPS
5.6K HDR Nodes
Dragonfly+ Topology

NVIDIA

27.6 PetaFLOPS
3K HDR Nodes
Fat-Tree Topology

Microsoft Azure

HPC/AI Cloud
HDR InfiniBand

NSF
SDSC
SAN DIEGO SUPERCOMPUTER CENTER
PITTSBURGH SUPERCOMPUTING CENTER
PURDUE UNIVERSITY

HDR
Supercomputers

TACC
TEXAS ADVANCED COMPUTING CENTER

23.5 PetaFLOPS
8K HDR Nodes
Fat-Tree Topology

NVIDIA

SCALABLE HIERARCHICAL AGGREGATION AND REDUCTION PROTOCOL (SHARP)

SCALABLE HIERARCHICAL AGGREGATION AND REDUCTION PROTOCOL (SHARP)

In-network Tree based aggregation mechanism

Multiple simultaneous outstanding operations

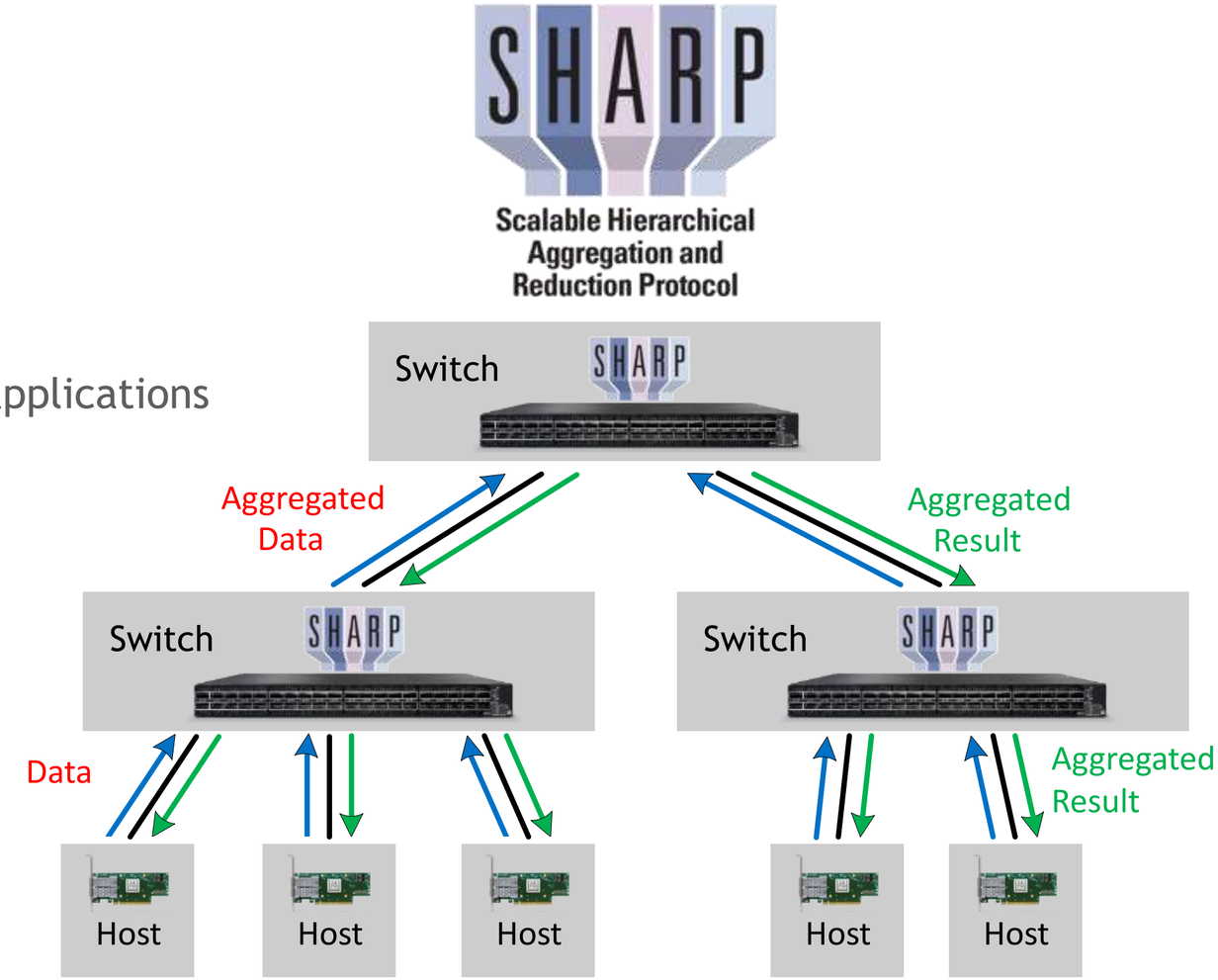
For HPC (MPI / SHMEM) and Distributed Machine Learning applications

Scalable High Performance Collective Offload

Barrier, Reduce, All-Reduce, Broadcast and more

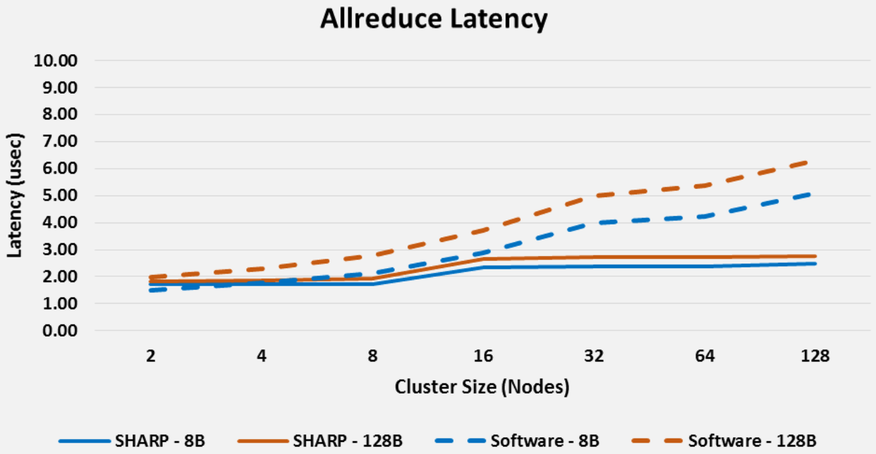
Sum, Min, Max, Min-loc, max-loc, OR, XOR, AND

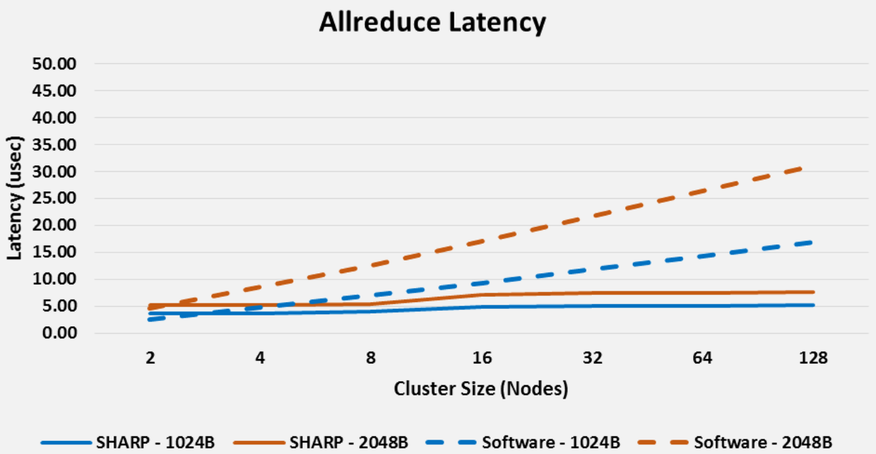
Integer and Floating-Point, 16/32/64 bits

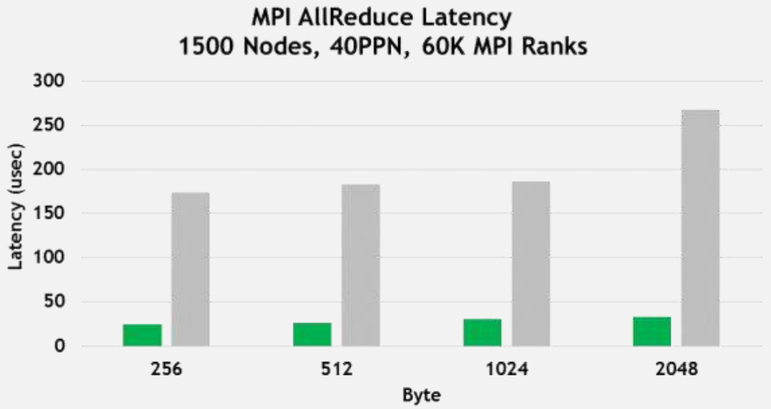


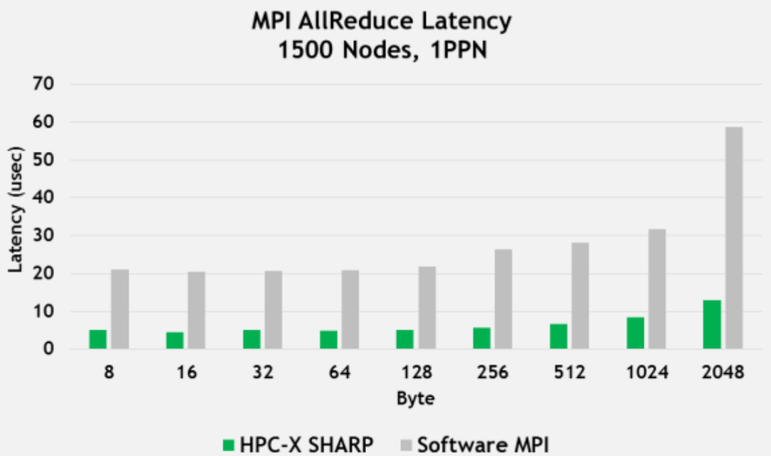
SHARP ALLREDUCE PERFORMANCE ADVANTAGES

Providing Flat Latency, 7X Higher Performance



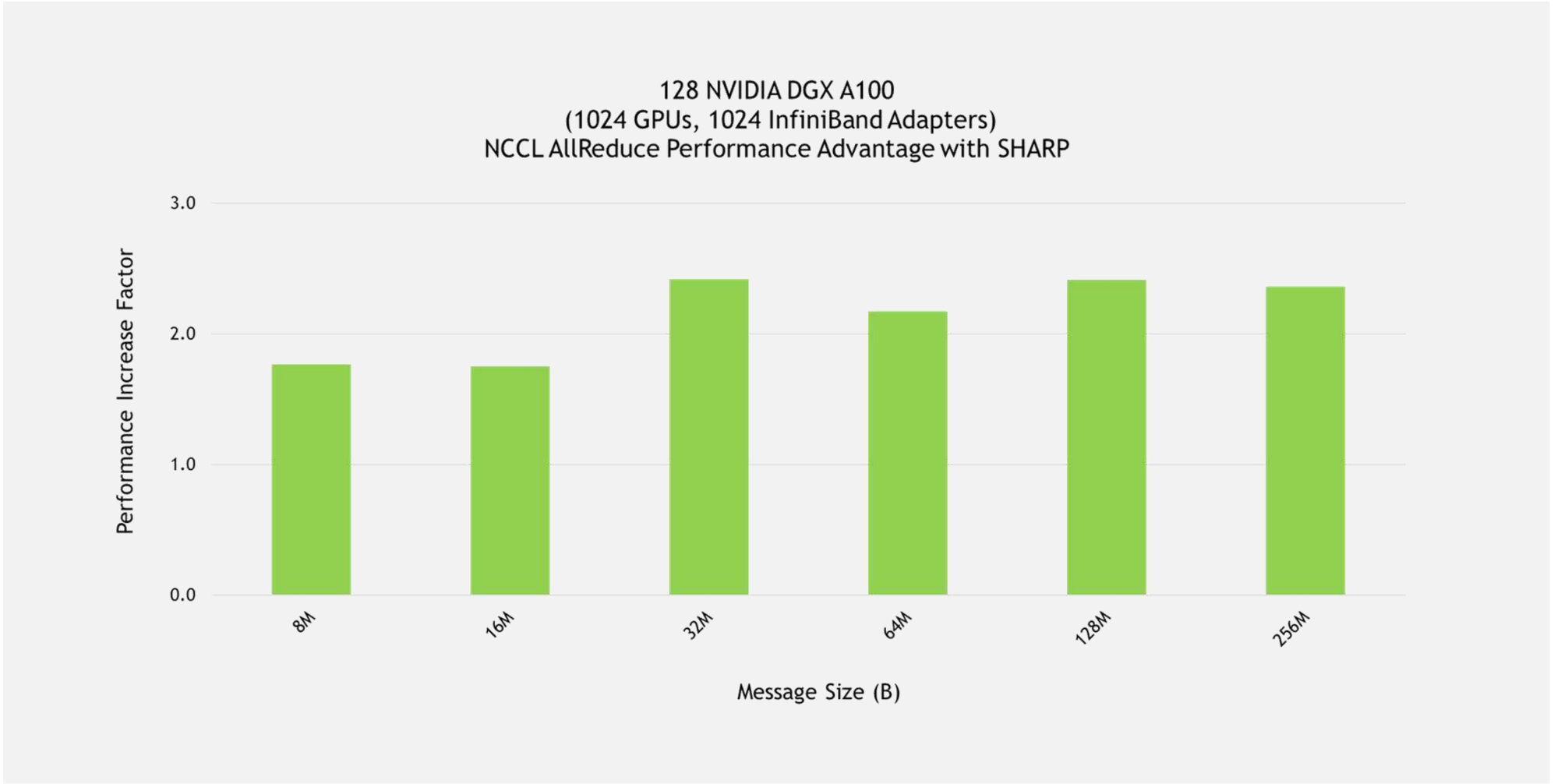






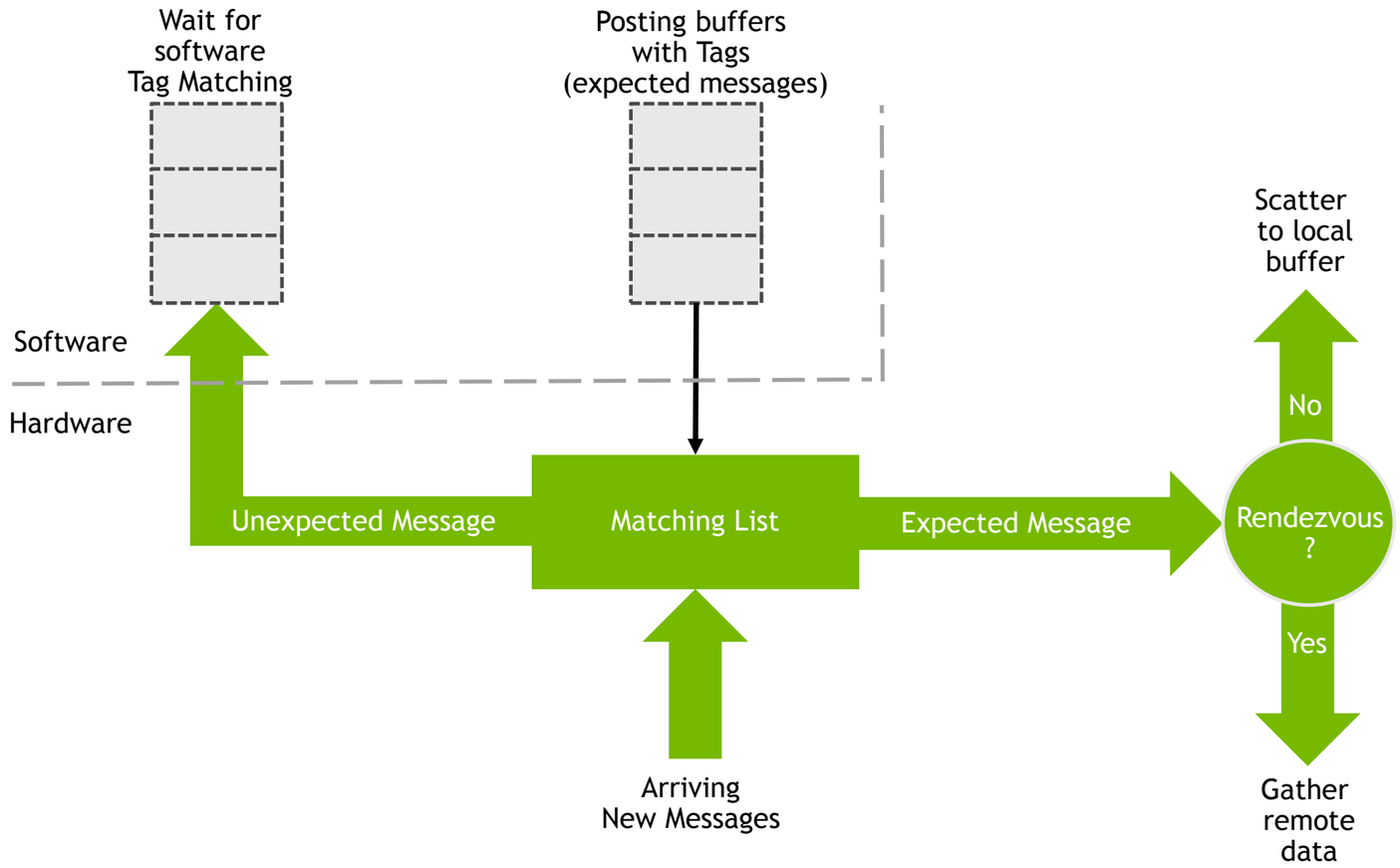
INFINIBAND SHARP AI PERFORMANCE ADVANTAGE

2.5X Higher Performance



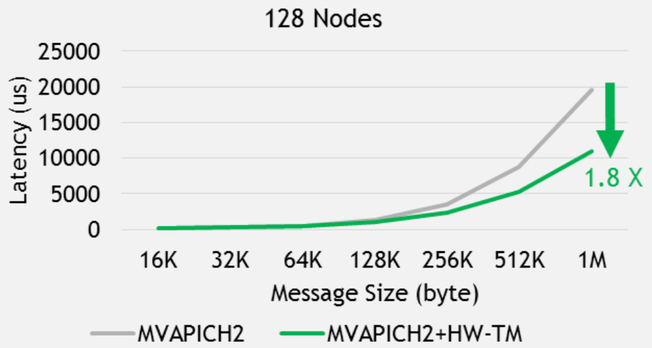
MPI TAG MATCHING HARDWARE ENGINE

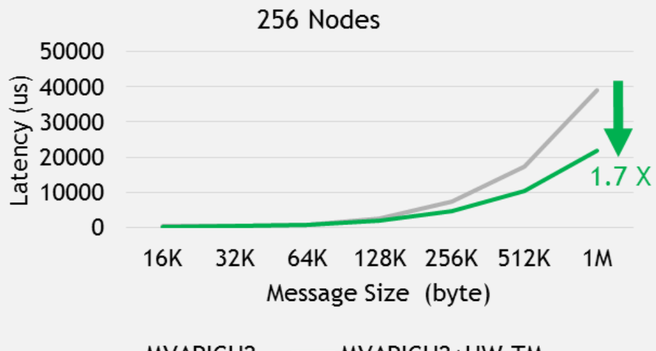
INFINIBAND MPI TAG MATCHING HARDWARE ENGINE

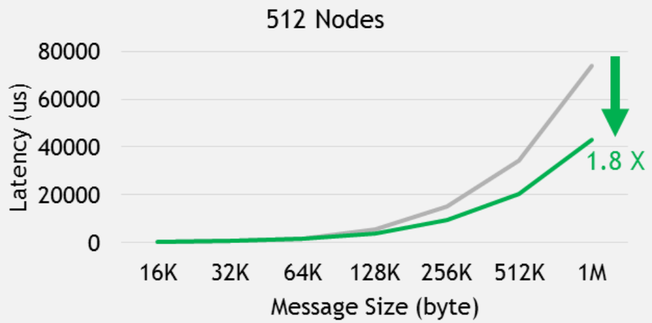


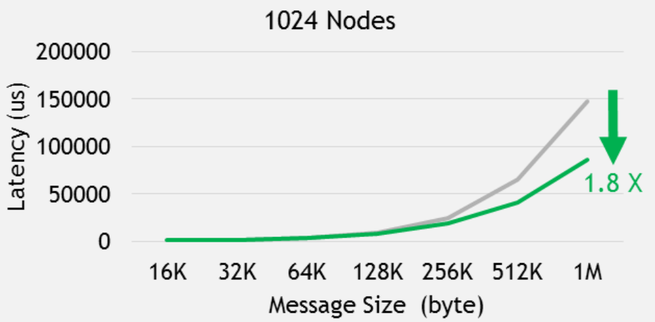
HARDWARE TAG MATCHING PERFORMANCE ADVANTAGES

1.8X Higher MPI_Iscatterv Performance on TACC Frontera





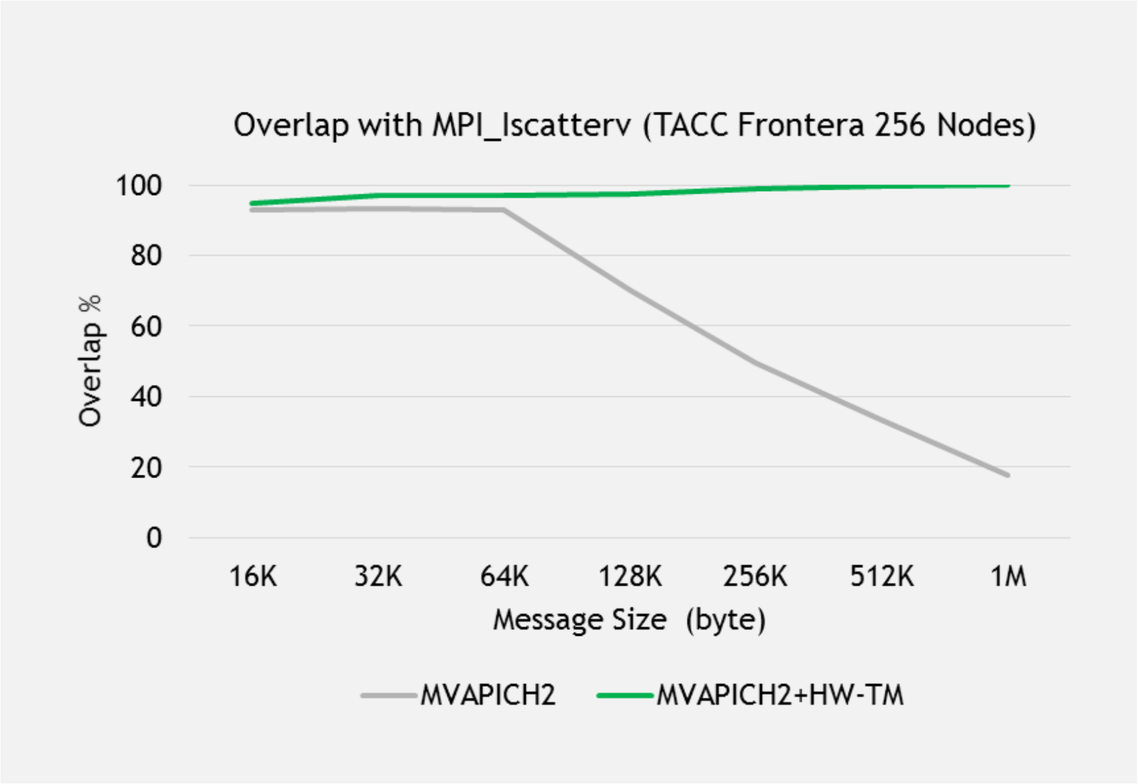


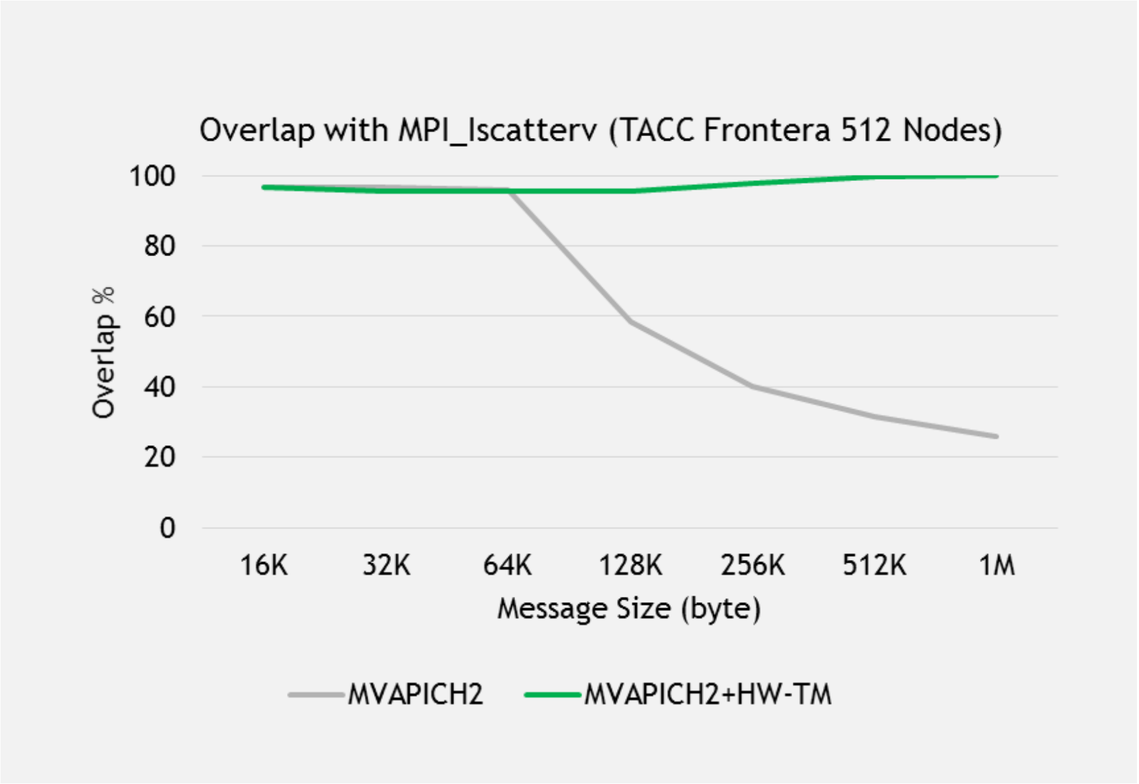


Courtesy of Dhabaleswar K. (DK) Panda
Ohio State University

HARDWARE TAG MATCHING PERFORMANCE ADVANTAGES

Nearly 100% Compute – Communication Overlap



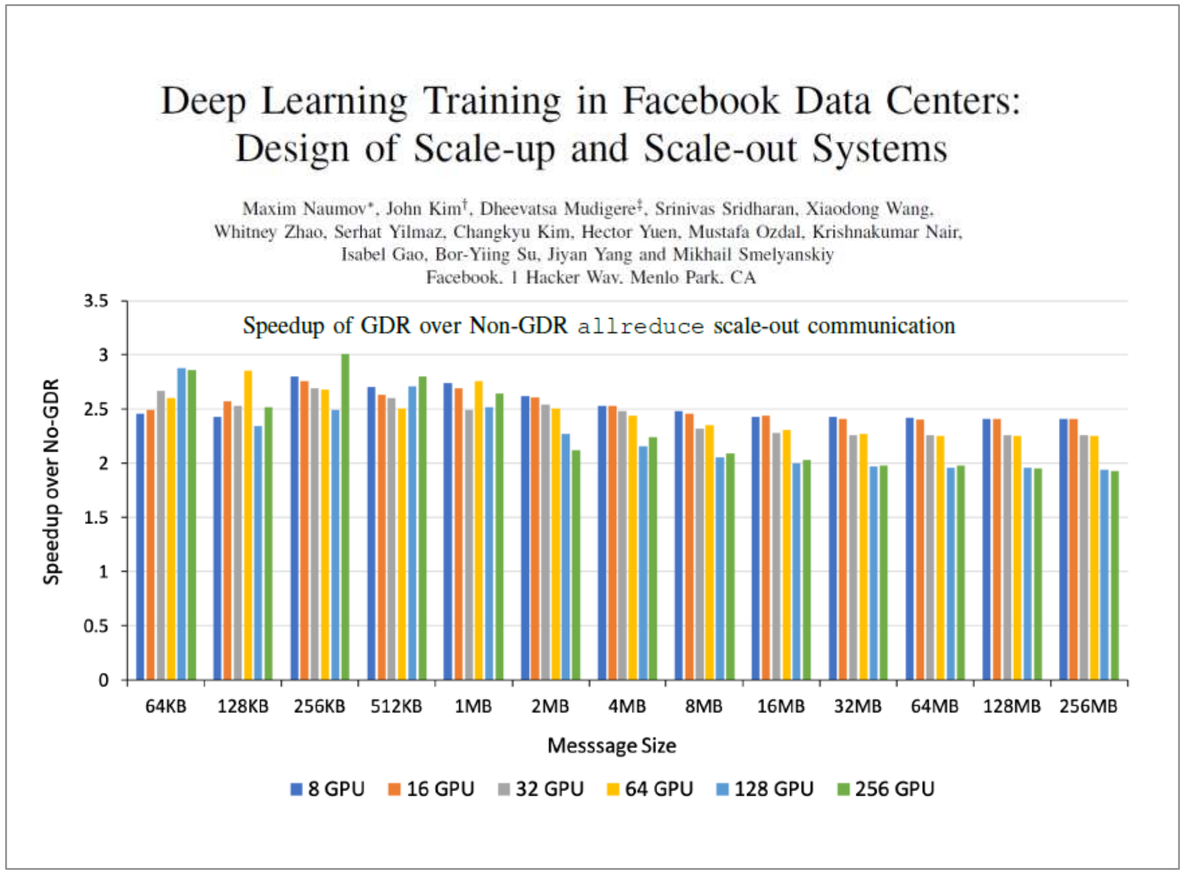


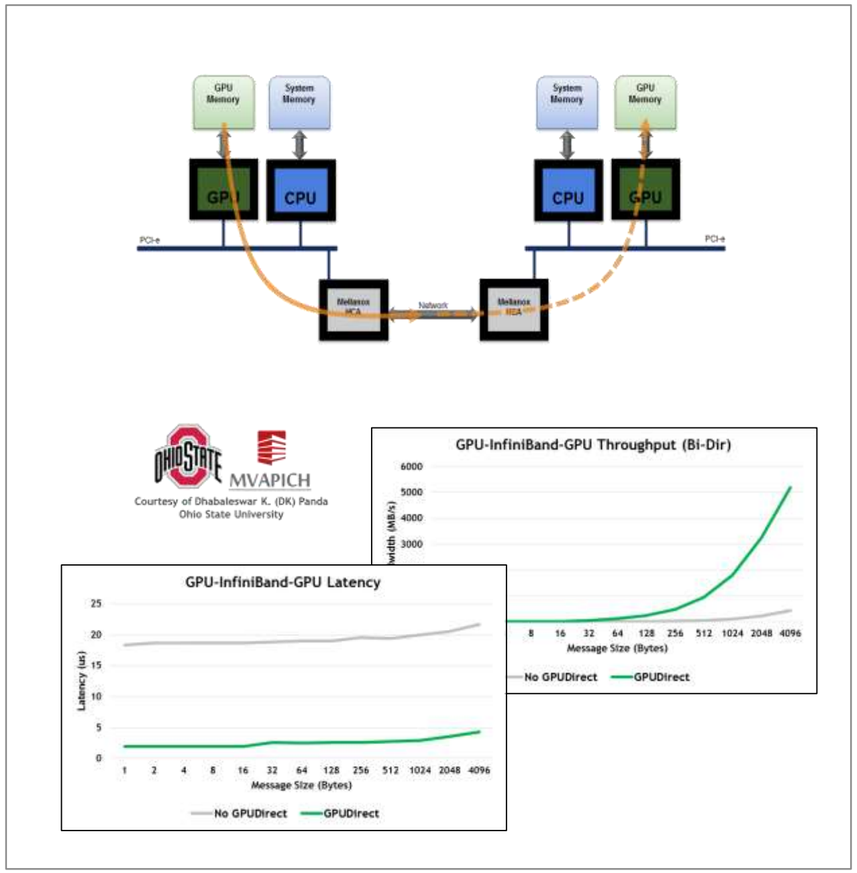
Courtesy of Dhabaleswar K. (DK) Panda
Ohio State University

GPUDIRECT

EFFICIENT COMMUNICATION FOR ACCELERATED TRAINING

10X Better Latency & Bandwidth, 3X Faster Deep Learning

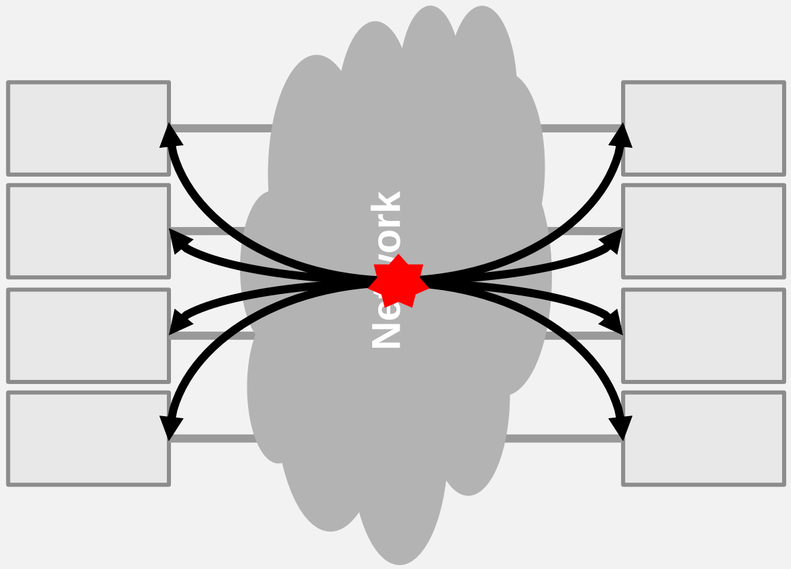




QUALITY OF SERVICE

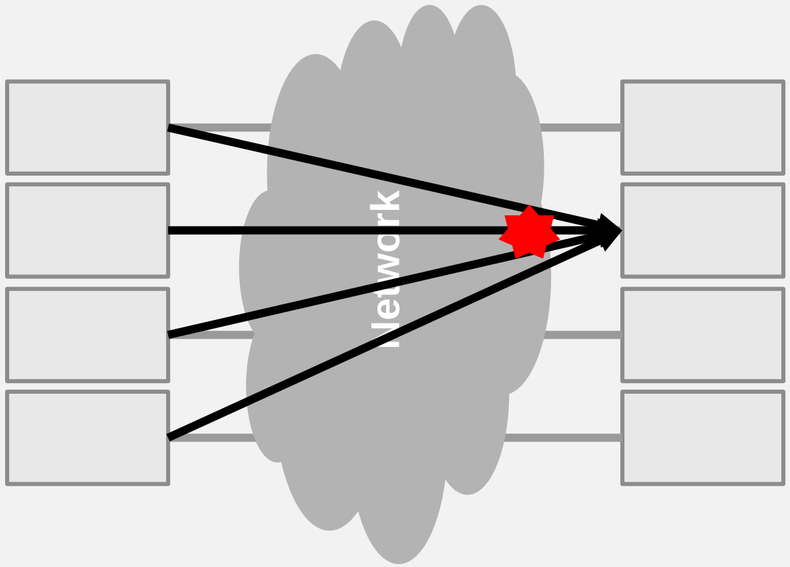
NETWORK CONGESTION TYPES

In-network Congestion



Solution: Adaptive Routing

In-cast Congestion

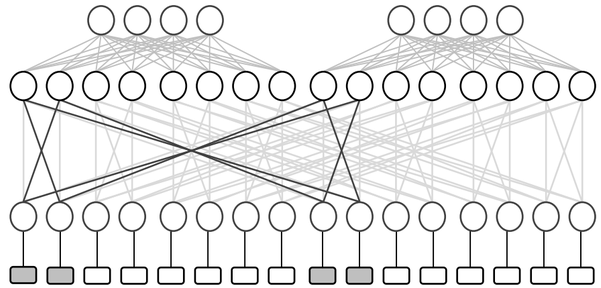


Solutions: Congestion Control

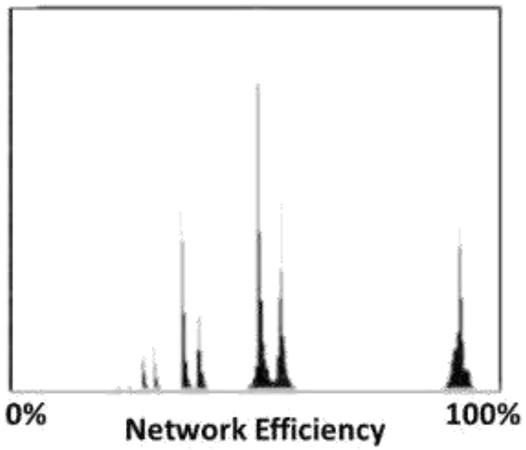
IN-NETWORK CONGESTION: ADAPTIVE ROUTING

The Design, Deployment, and Evaluation of the CORAL Pre-Exascale Systems

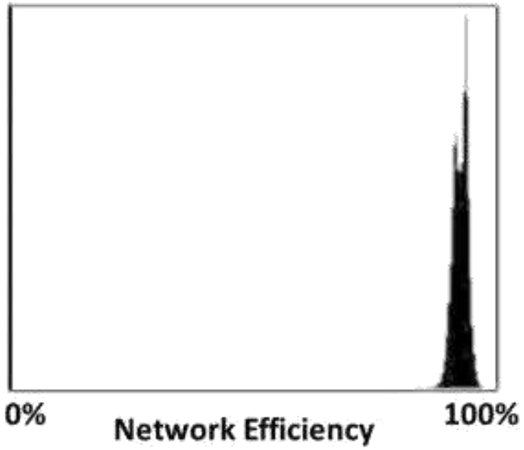
Sudharshan S. Vazhkudai†, Bronis R. de Supinski‡, Arthur S. Bland†, Al Geist†, James Sexton*, Jim Kahle*, Christopher J. Zimmer†, Scott Atchley†, Sarp Oral†, Don E. Maxwell†, Veronica G. Vergara Larrea†, Adam Bertsch‡, Robin Goldstone‡, Wayne Joubert†, Chris Chambreau‡, David Appelhans*, Robert Blackmore*, Ben Casses‡, George Chochia*, Gene Davison*, Matthew A. Ezell†, Tom Gooding*, Elsa Gonsiorowski‡, Leopold Grinberg*, Bill Hanson*, Bill Hartner*, Ian Karlin‡, Matthew L. Leininger‡, Dustin Leverman†, Chris Marroquin*, Adam Moody‡, Martin Ohmacht*, Ramesh Pankajakshan‡, Fernando Pizzano*, James H. Rogers†, Bryan Rosenburg*, Drew Schmidt†, Mallikarjun Shankar†, Feiyi Wang†, Py Watson‡, Bob Walkup*, Lance D. Weems‡, Junqi Yin†
† Oak Ridge National Laboratory, ‡ Lawrence Livermore National Laboratory, * IBM
{vazhkudaiss@ornl.gov, bronis@llnl.gov}



mpiGraph: Static vs. Adaptive Routing

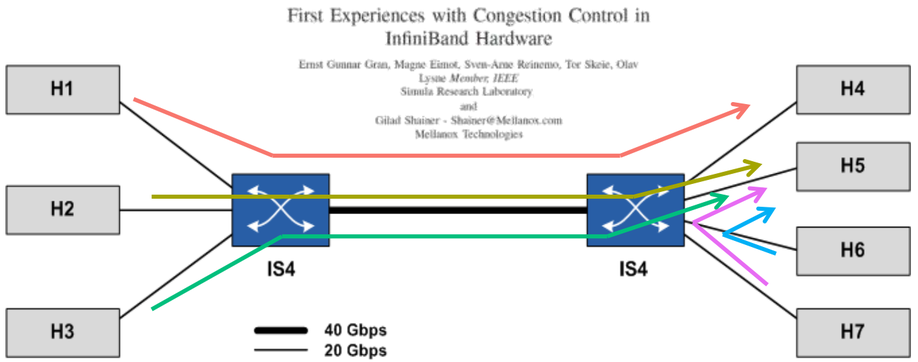


Static Routing

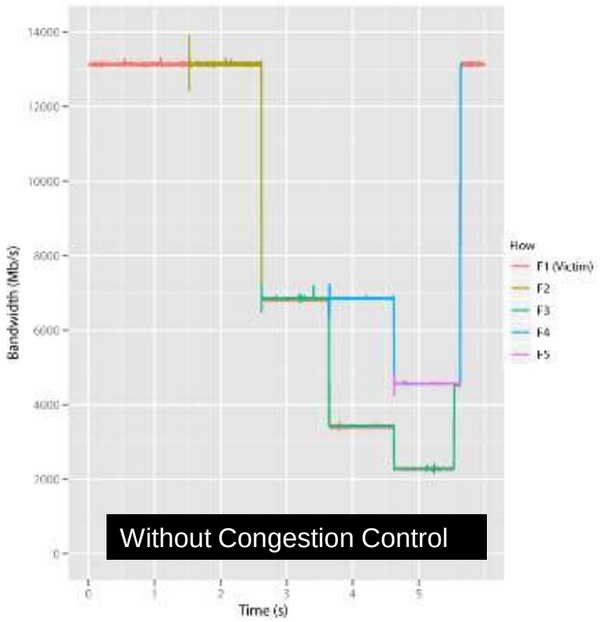


Adaptive Routing

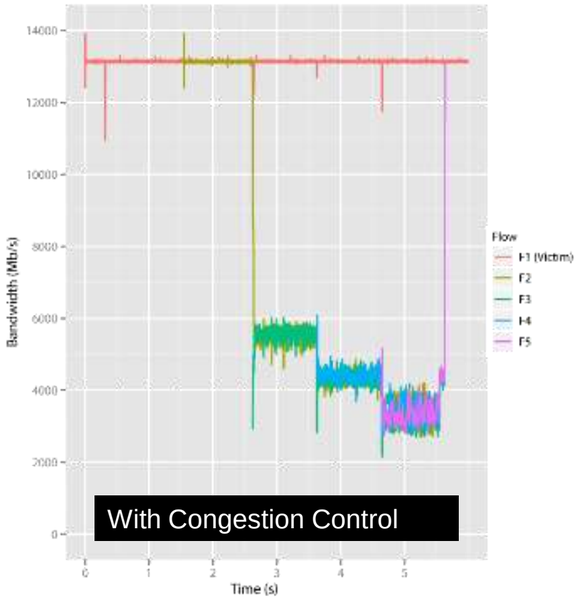
INFINIBAND CONGESTION CONTROL

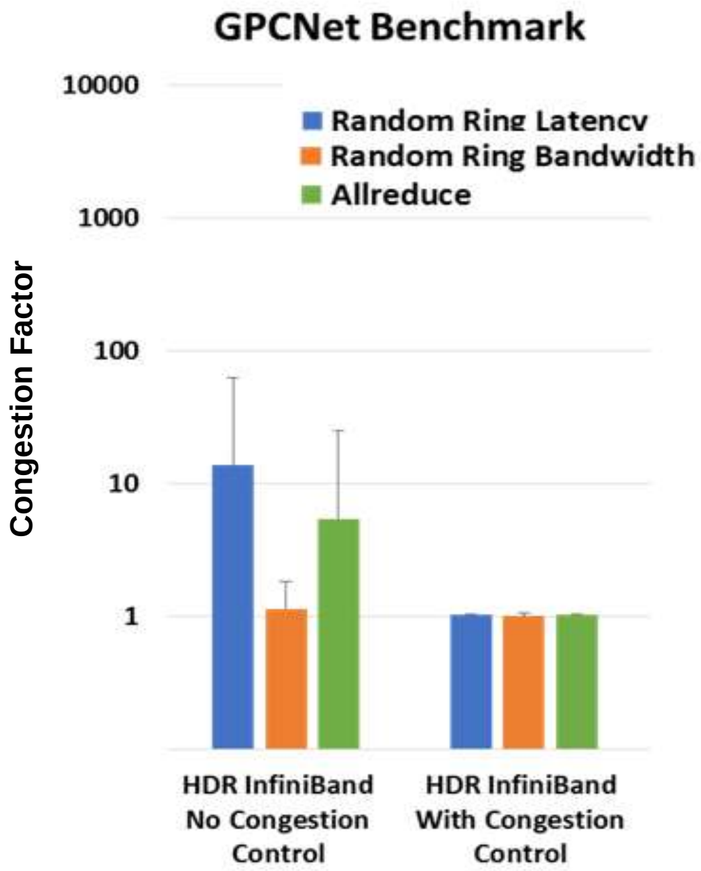


Congestion – Throughput loss



No congestion – highest throughput!

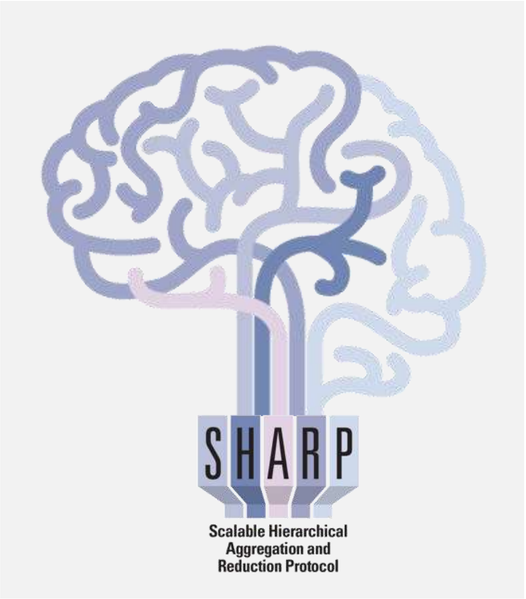




INFINIBAND ACCELERATED SUPERCOMPUTING

Speed of Light

200Gb/s Data Throughput
RDMA and GPUDirect RDMA
3X Better (Lower) Latency



SHARP AI Technology

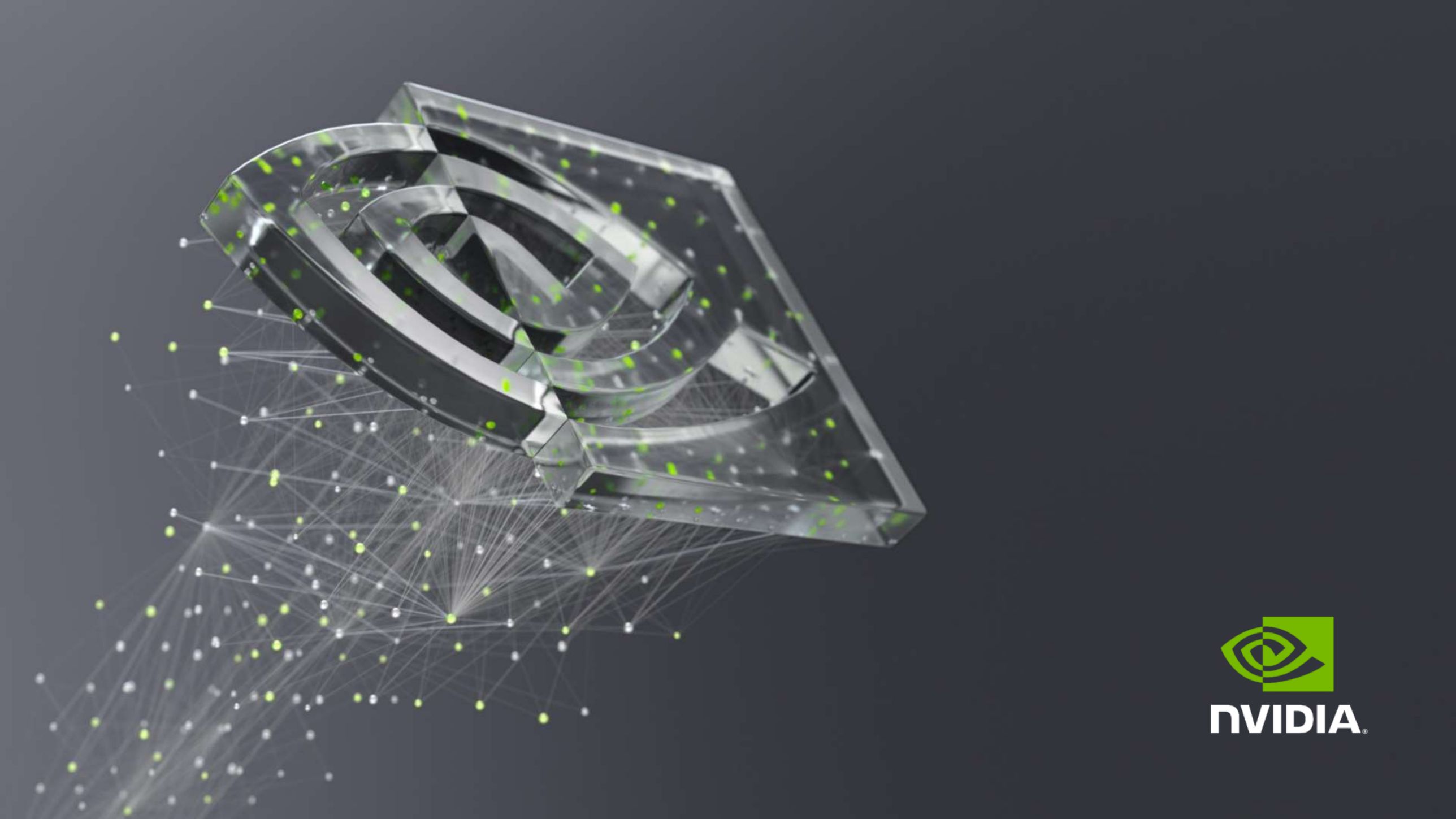
AI Acceleration Engines
2.5X Higher AI Performance

SHIELD AI Technology

Self Healing Network
1000X Faster Recovery Time

UFM Cyber AI

Data Center Cyber Intelligence
and Analytics


NVIDIA