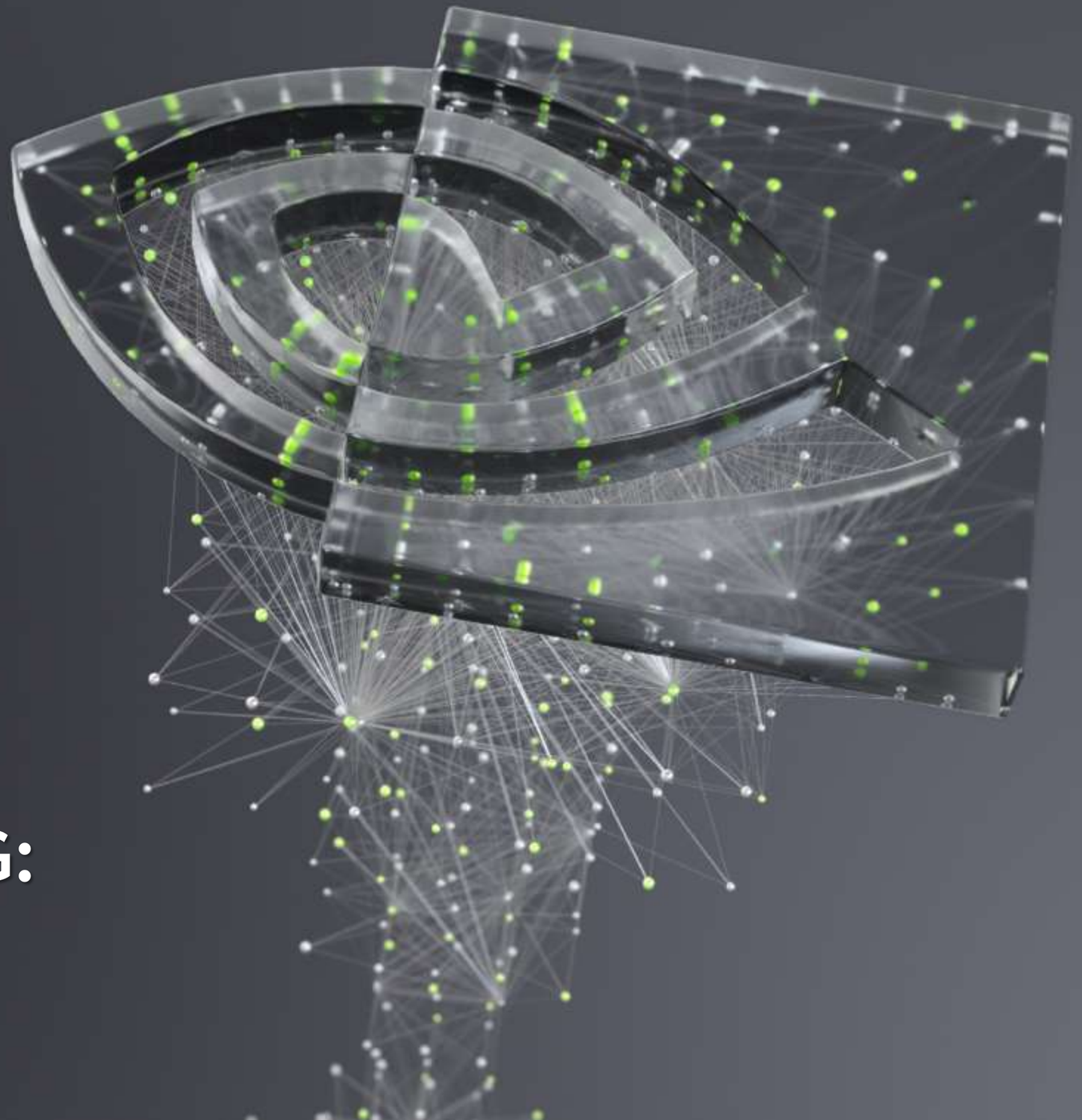
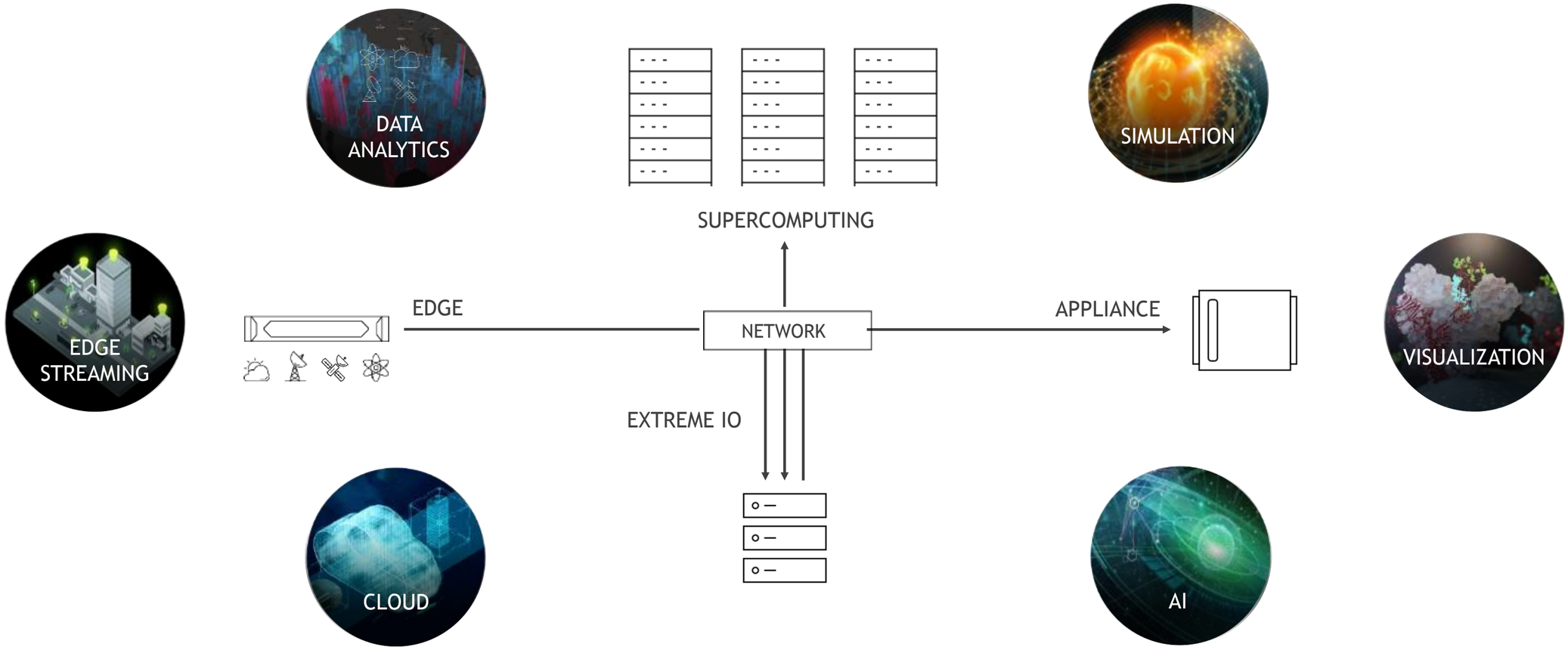


IN-NETWORK COMPUTING:

January 2020

EXPANDING UNIVERSE OF COMPUTING



NVIDIA PLATFORM POWERING THE EXASCALE AI SUPERCOMPUTERS

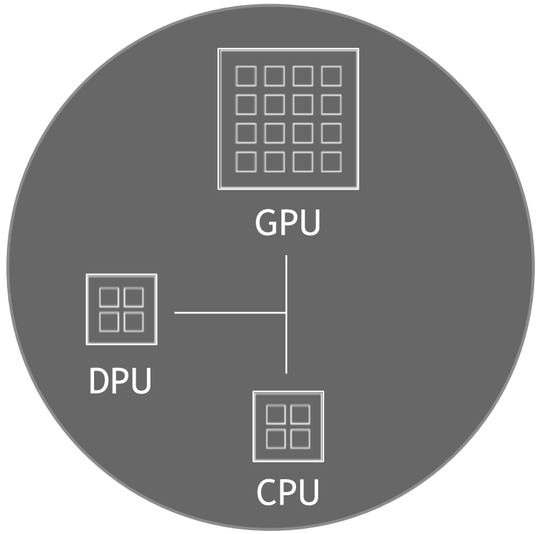
NVIDIA
2.8 EFLOPS AI Perf
HDR 200G InfiniBand

Jülich
2.3 EFLOPS AI Perf HDR
HDR 200G InfiniBand

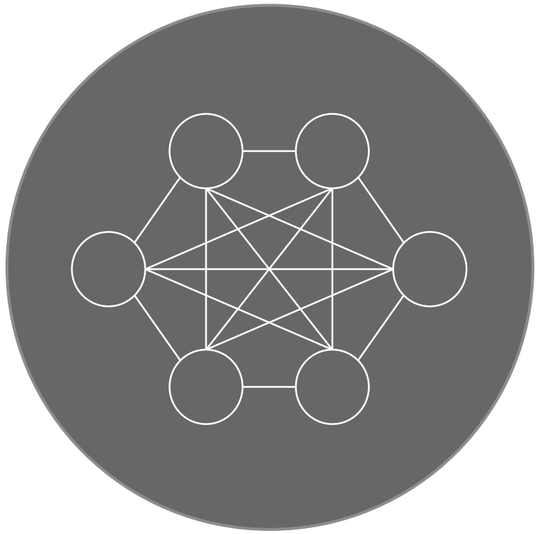
CINECA
10 EFLOPS AI Perf
HDR 200G InfiniBand

ABCI
600 PFLOPS AI Perf
HDR 200G InfiniBand

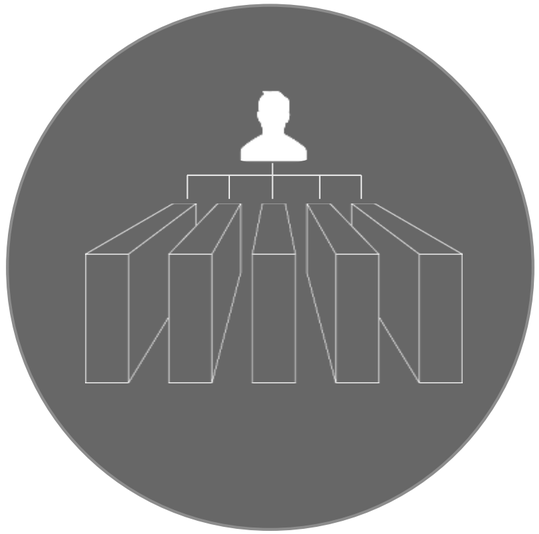
INFINIBAND TECHNOLOGY FUNDAMENTALS



Smart Networking



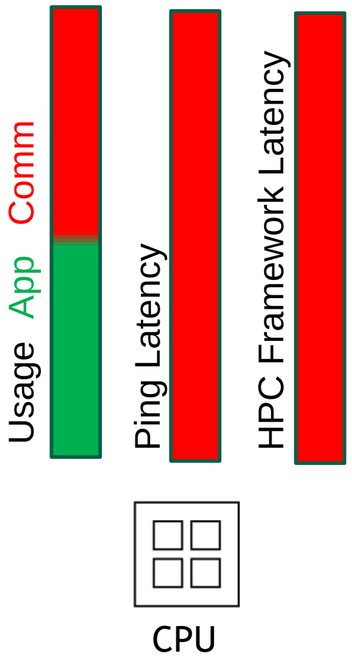
Architected to Scale

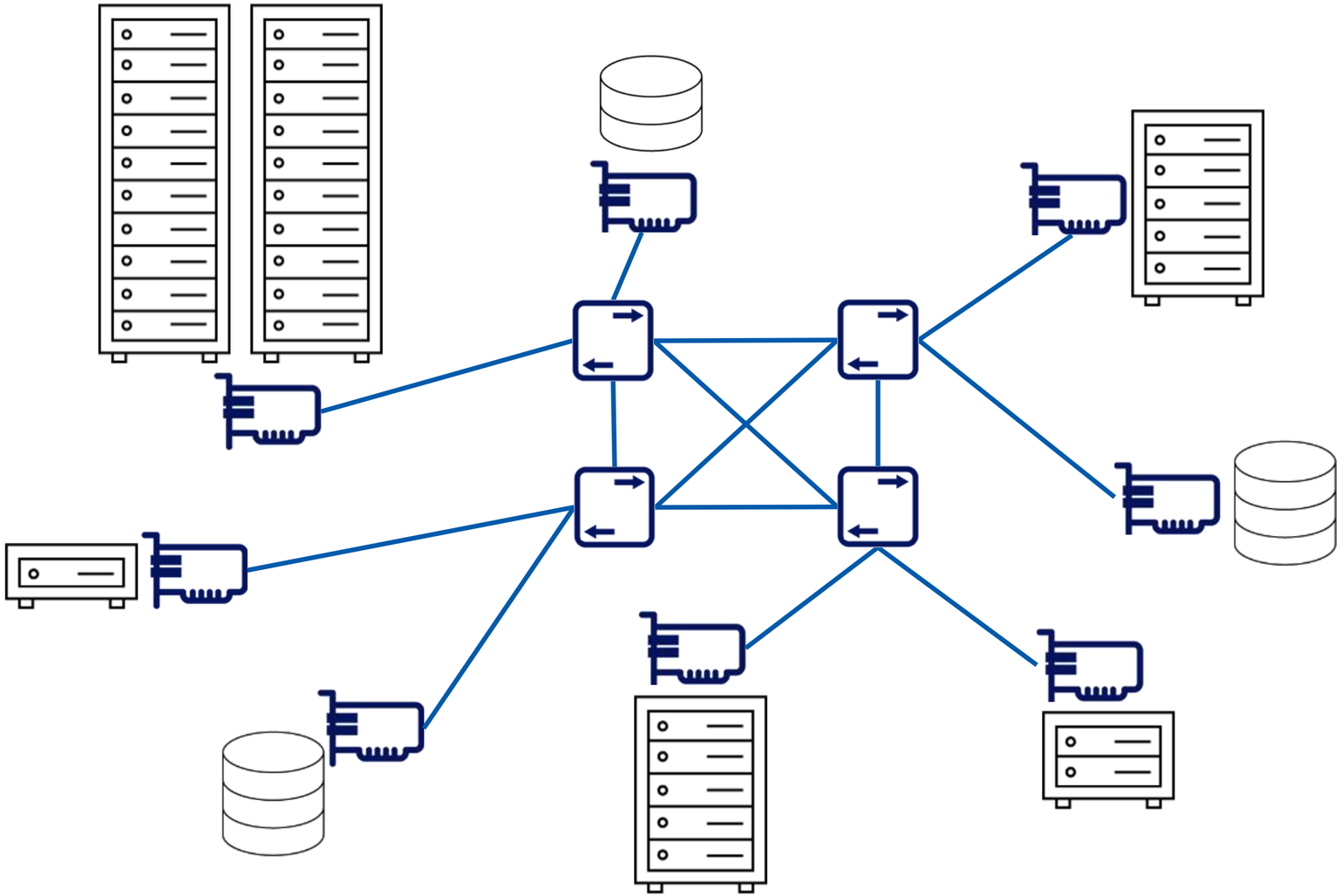


Centralized Management

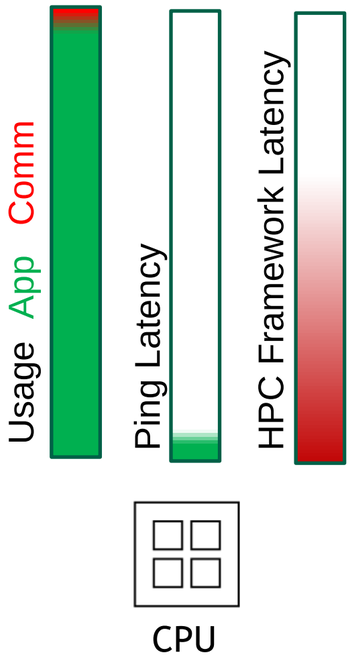
Standard

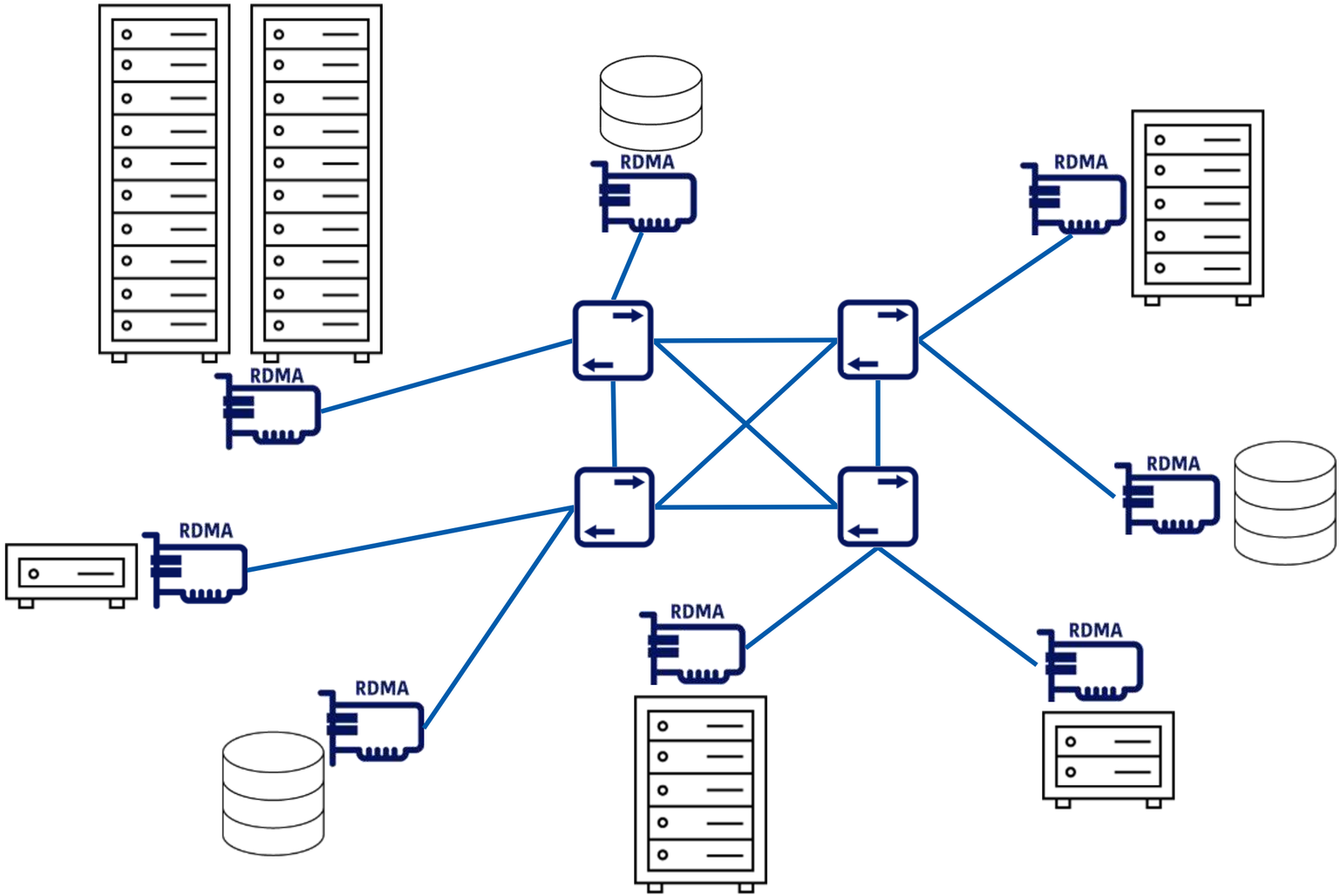
TRADITIONAL HIGH PERFORMANCE DATA CENTER





RDMA-ACCELERATED DATA CENTER

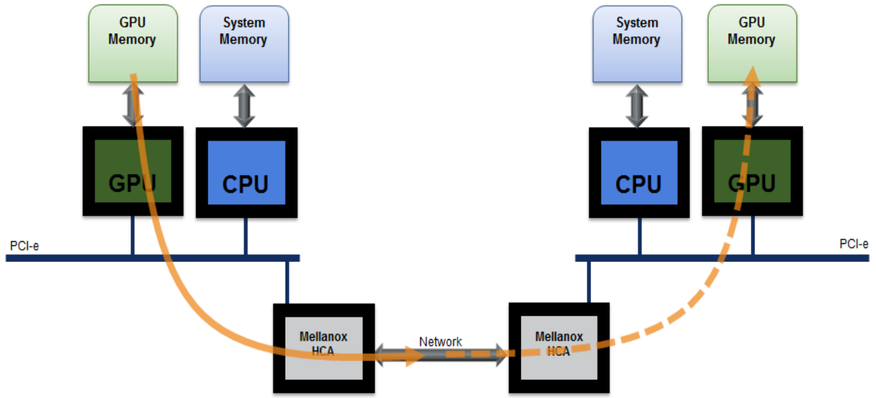




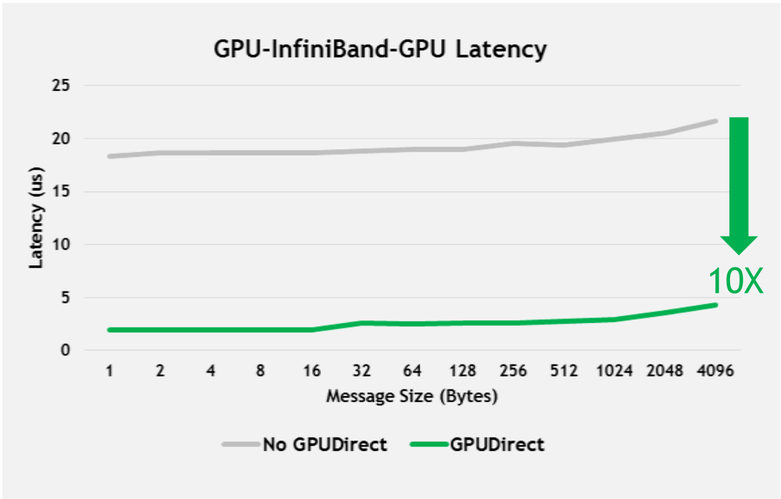
10X HIGHER PERFORMANCE WITH GPUDIRECT™ RDMA

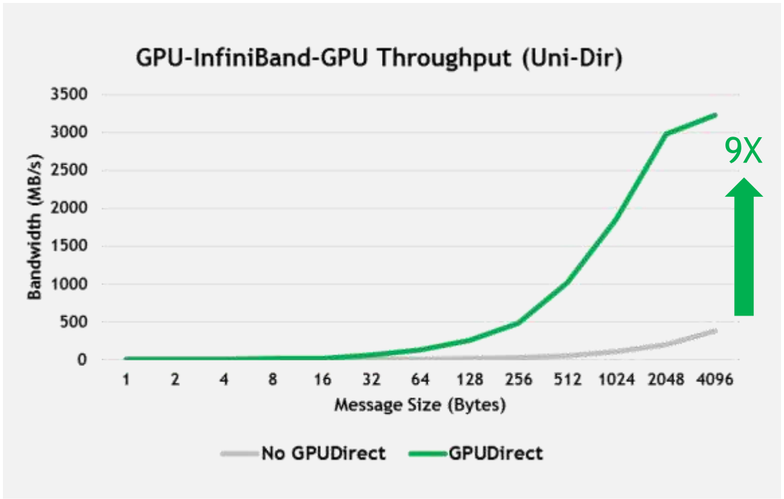
Accelerates HPC and Deep Learning performance

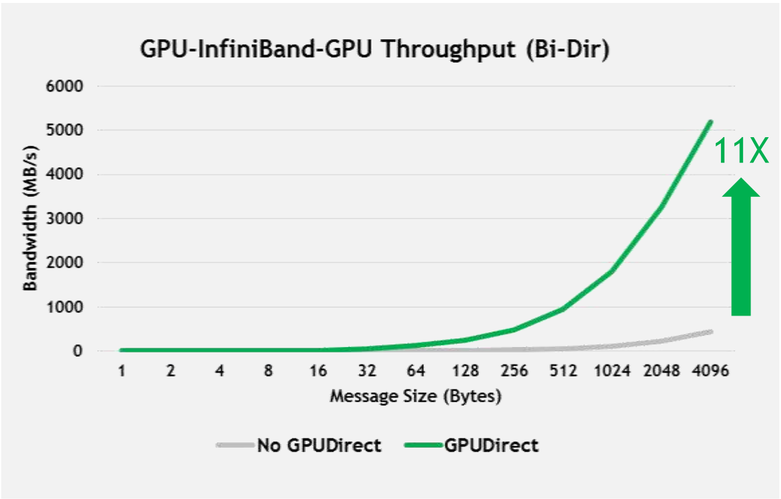
Lowest communication latency for GPUs



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

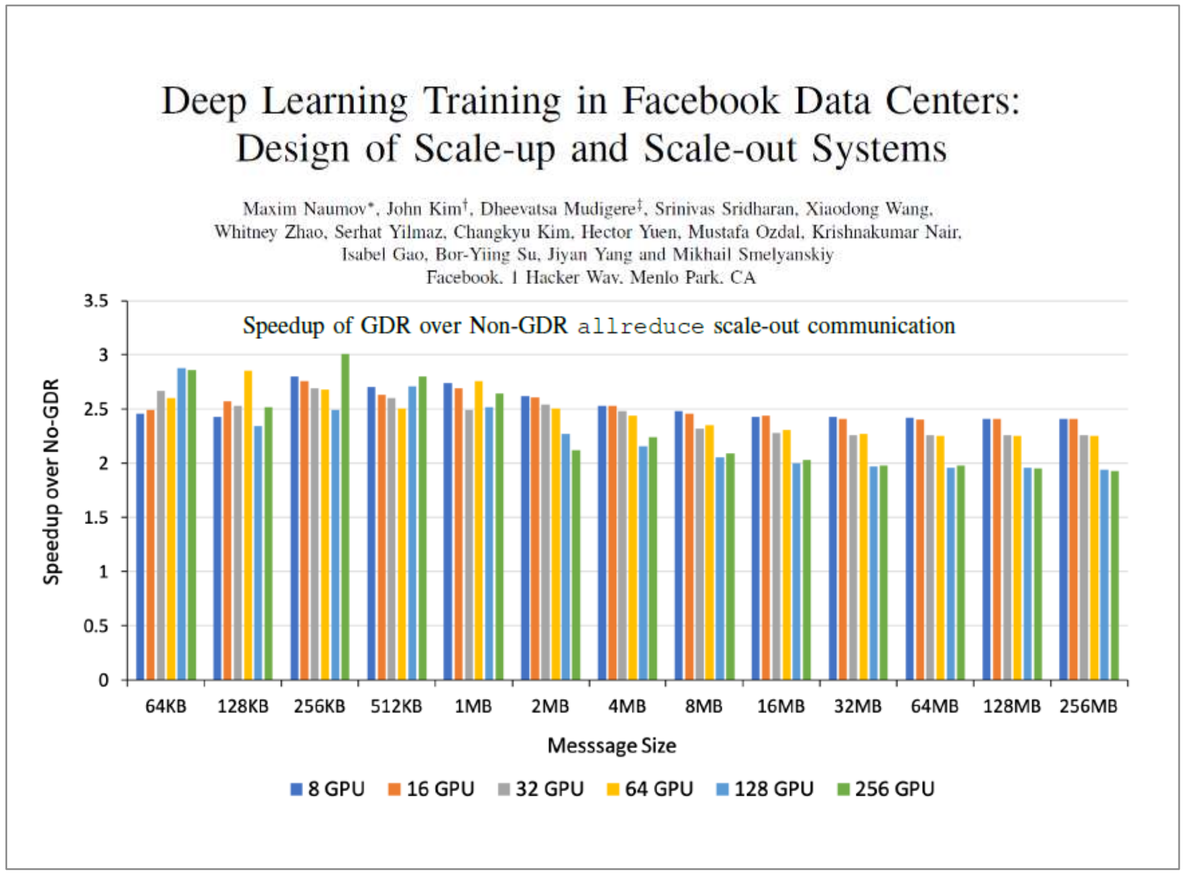


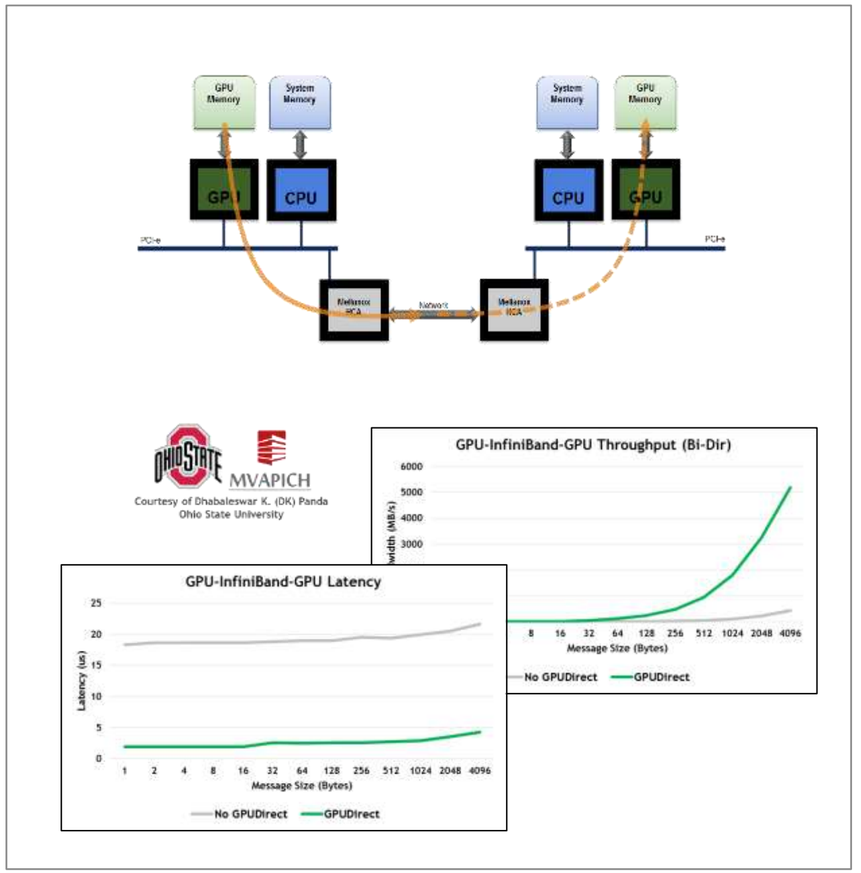




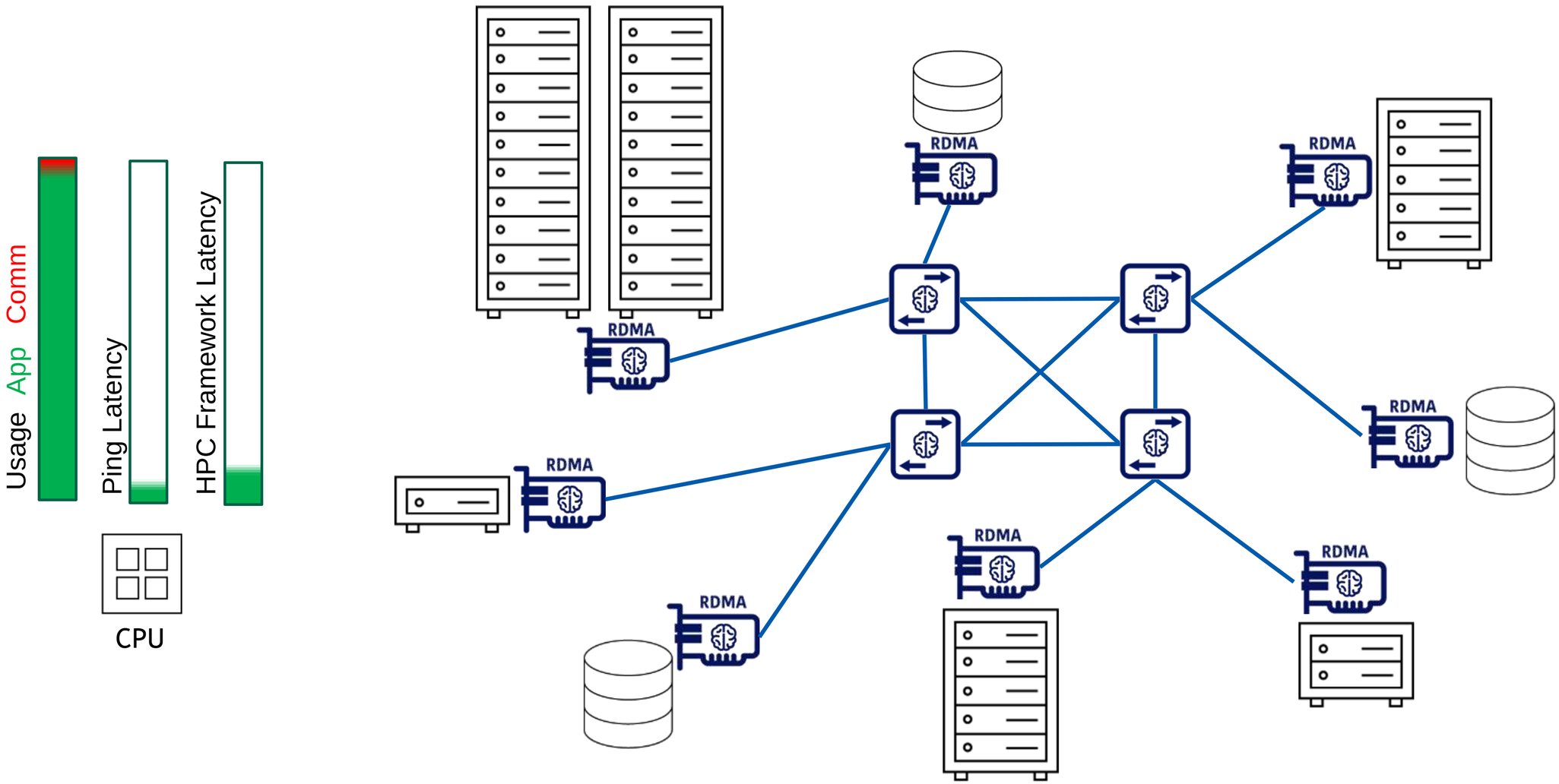
EFFICIENT COMMUNICATION FOR ACCELERATED TRAINING

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





IN-NETWORK COMPUTING-ACCELERATED DATA CENTER



SCALABLE HIERARCHICAL AGGREGATION AND REDUCTION PROTOCOL (SHARP)

In-network Tree based aggregation mechanism

Multiple simultaneous outstanding operations

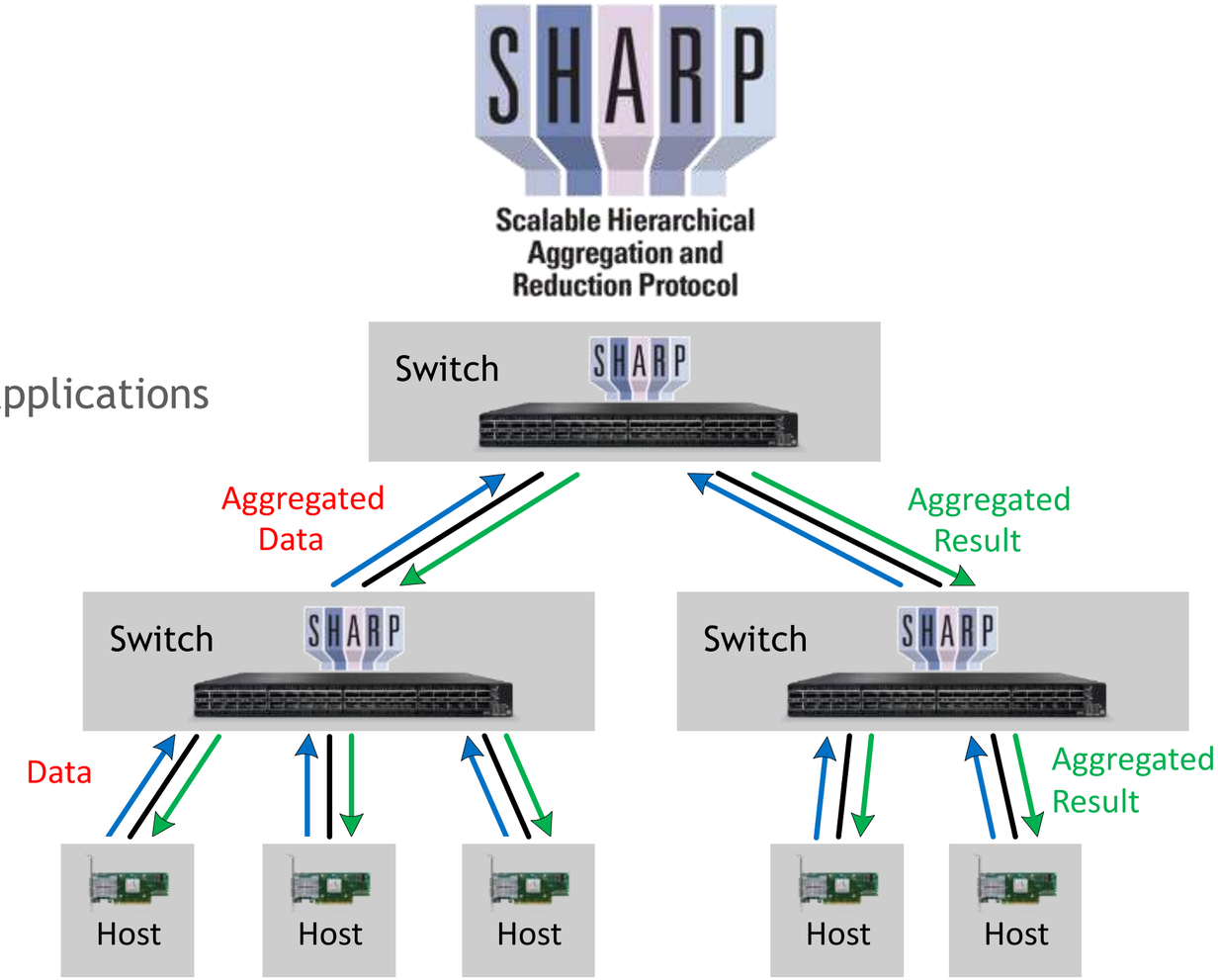
For HPC (MPI / SHMEM) and Distributed Machine Learning applications

Scalable High Performance Collective Offload

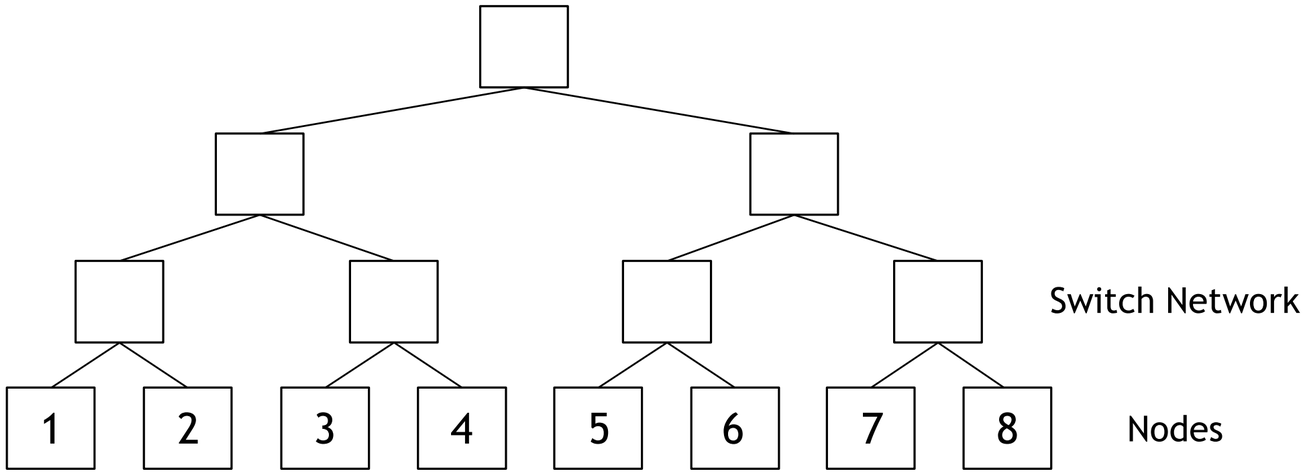
Barrier, Reduce, All-Reduce, Broadcast and more

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

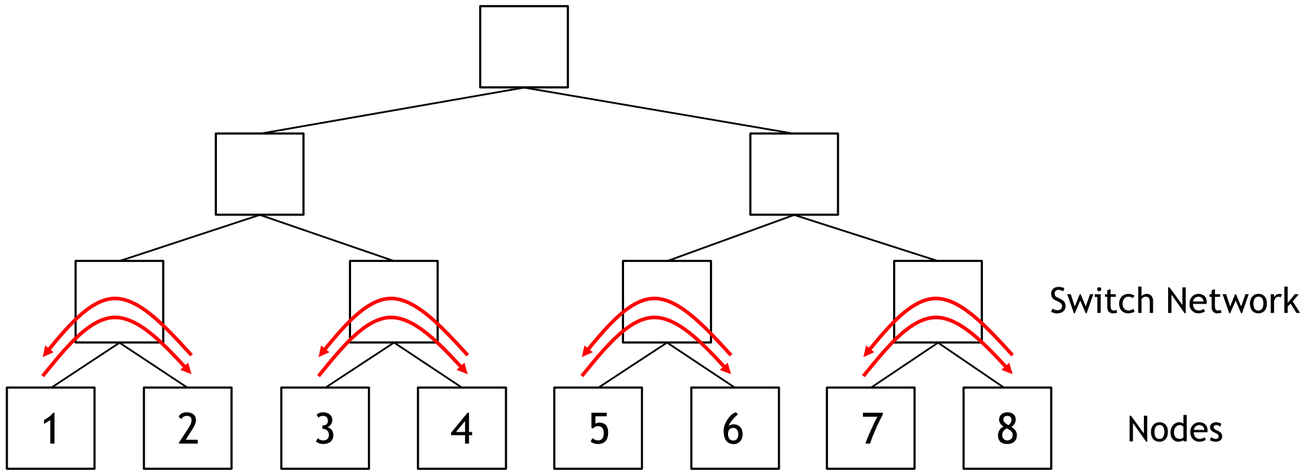
Integer and Floating-Point, 16/32/64 bits



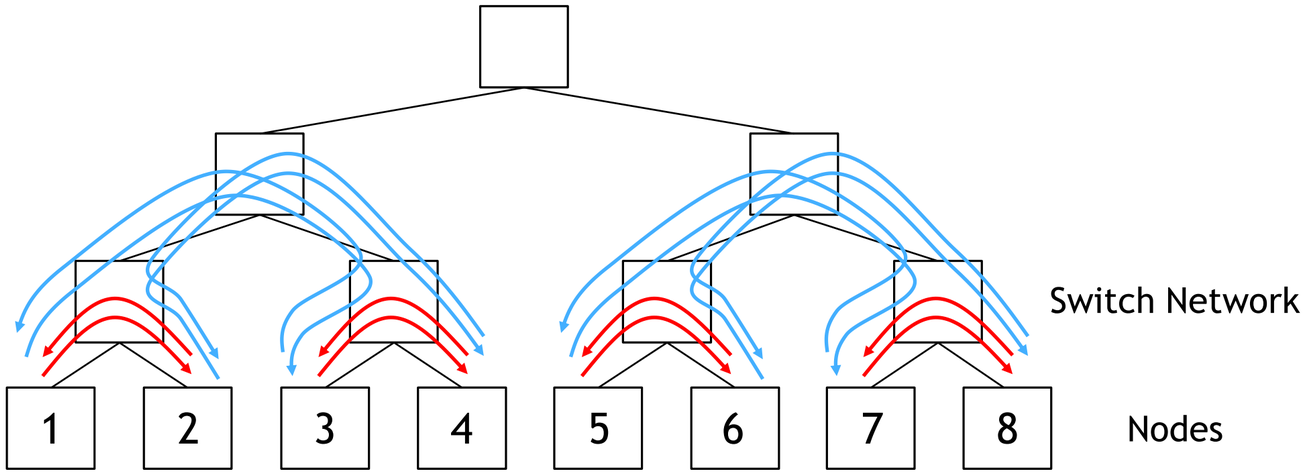
DATA AGGREGATION



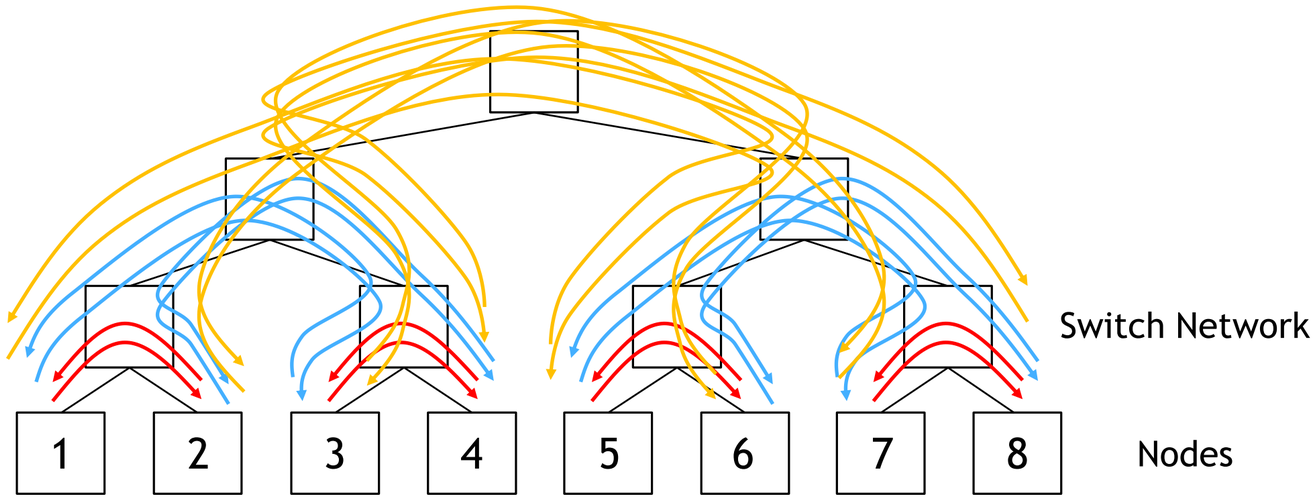
TRADITIONAL DATA AGGREGATION



TRADITIONAL DATA AGGREGATION



TRADITIONAL DATA AGGREGATION

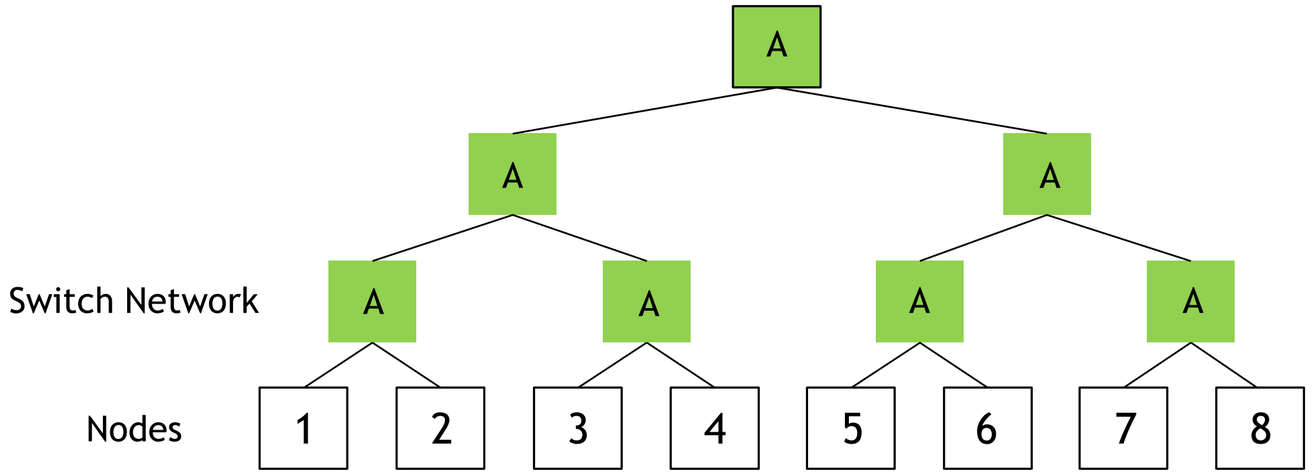


Switch Network

Nodes

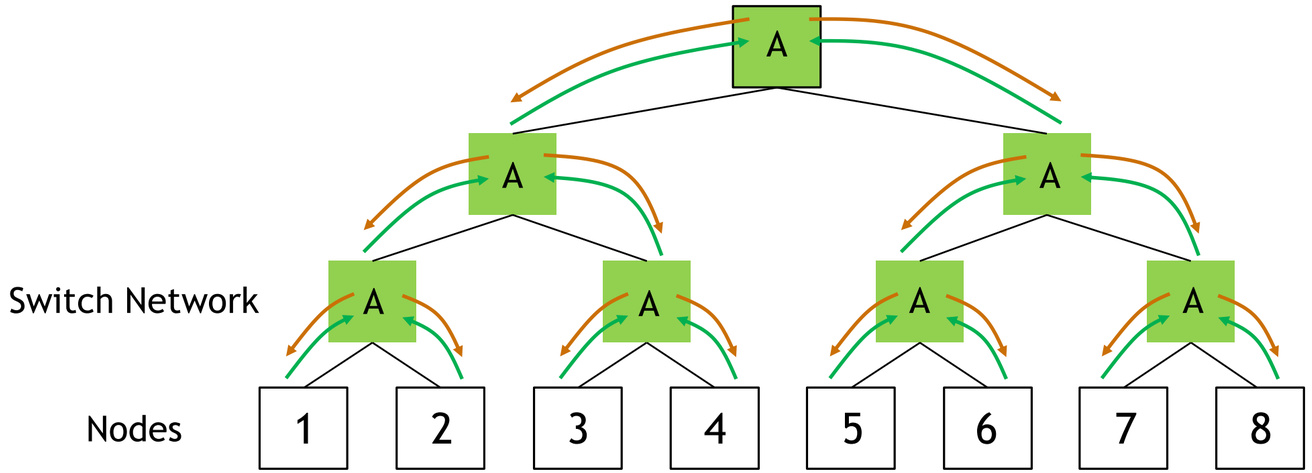
High latency
High amount of transferred data
CPU/GPU overhead

SHARP IN-NETWORK COMPUTING AGGREGATION



Scalable Hierarchical Aggregation and Reduction Protocol

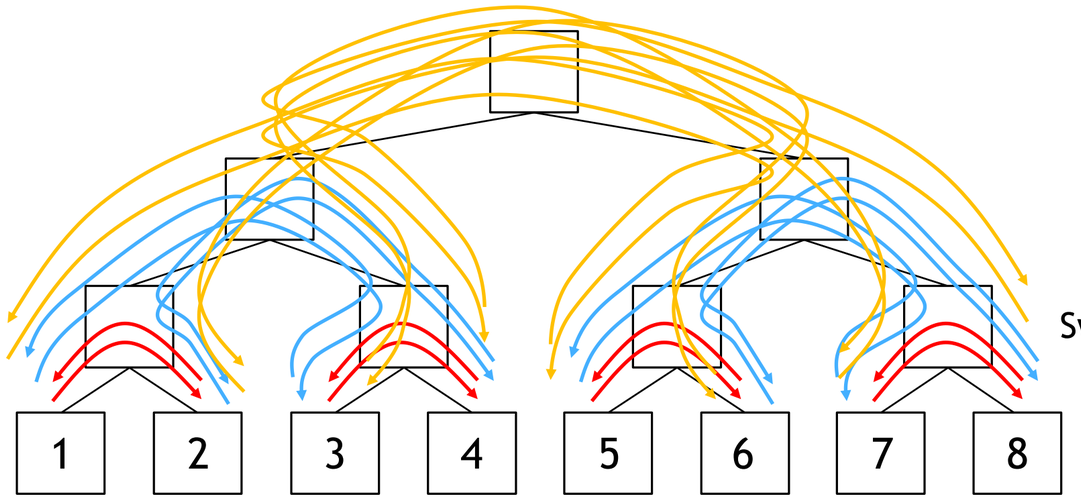
SHARP IN-NETWORK COMPUTING AGGREGATION



Low latency
Optimized data motion
No CPU/GPU calculation latency addition

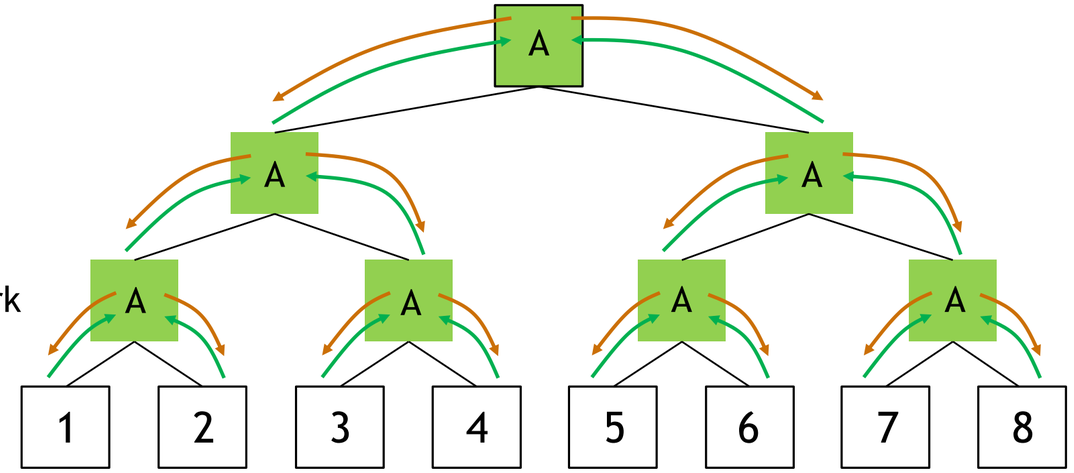
SHARP
Scalable Hierarchical Aggregation and Reduction Protocol

SHARP IN-NETWORK COMPUTING AGGREGATION



Switch Network

Nodes



High latency
High amount of transferred data
CPU/GPU overhead

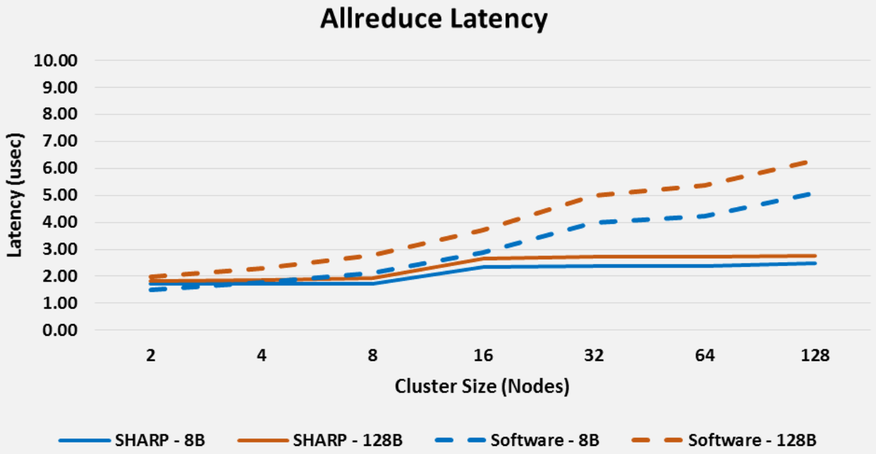
Low latency
Optimized data motion
No CPU/GPU calculation latency addition

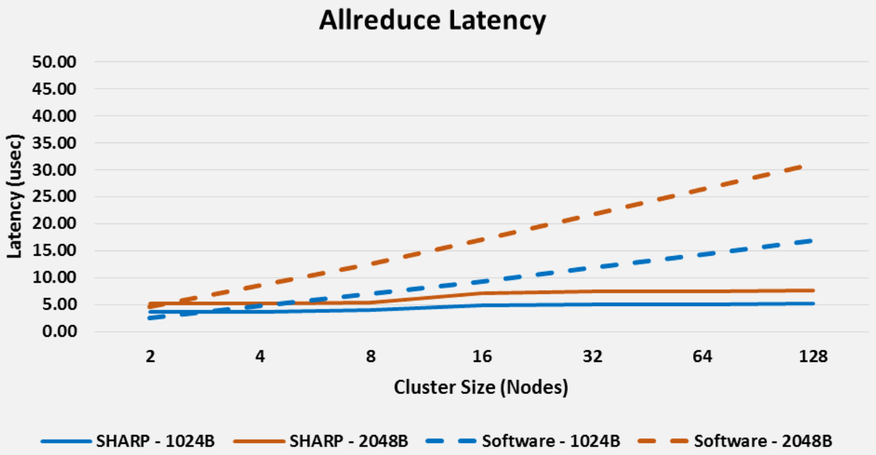
SHARP
Scalable Hierarchical Aggregation and Reduction Protocol

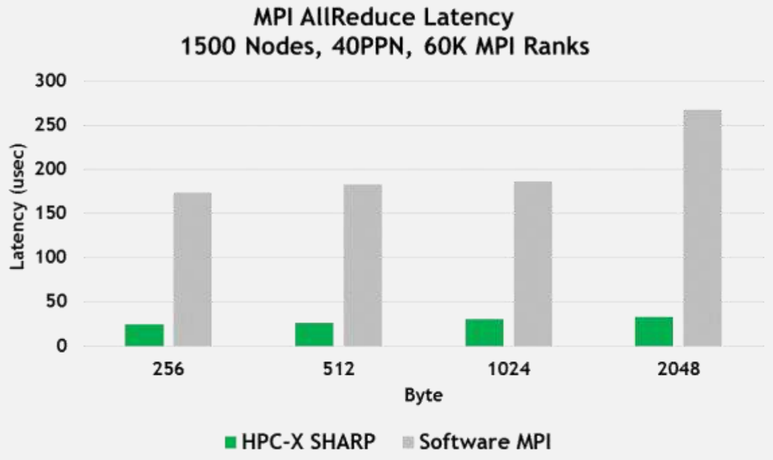
NVIDIA

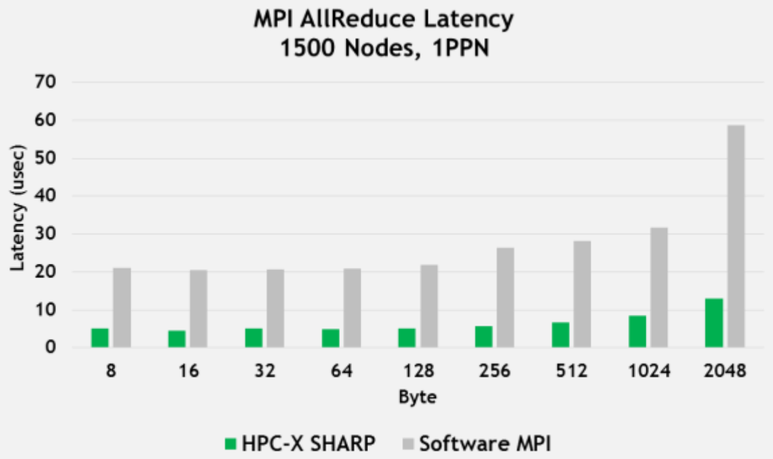
SHARP ALLREDUCE PERFORMANCE ADVANTAGES

Providing Flat Latency, 7X Higher Performance



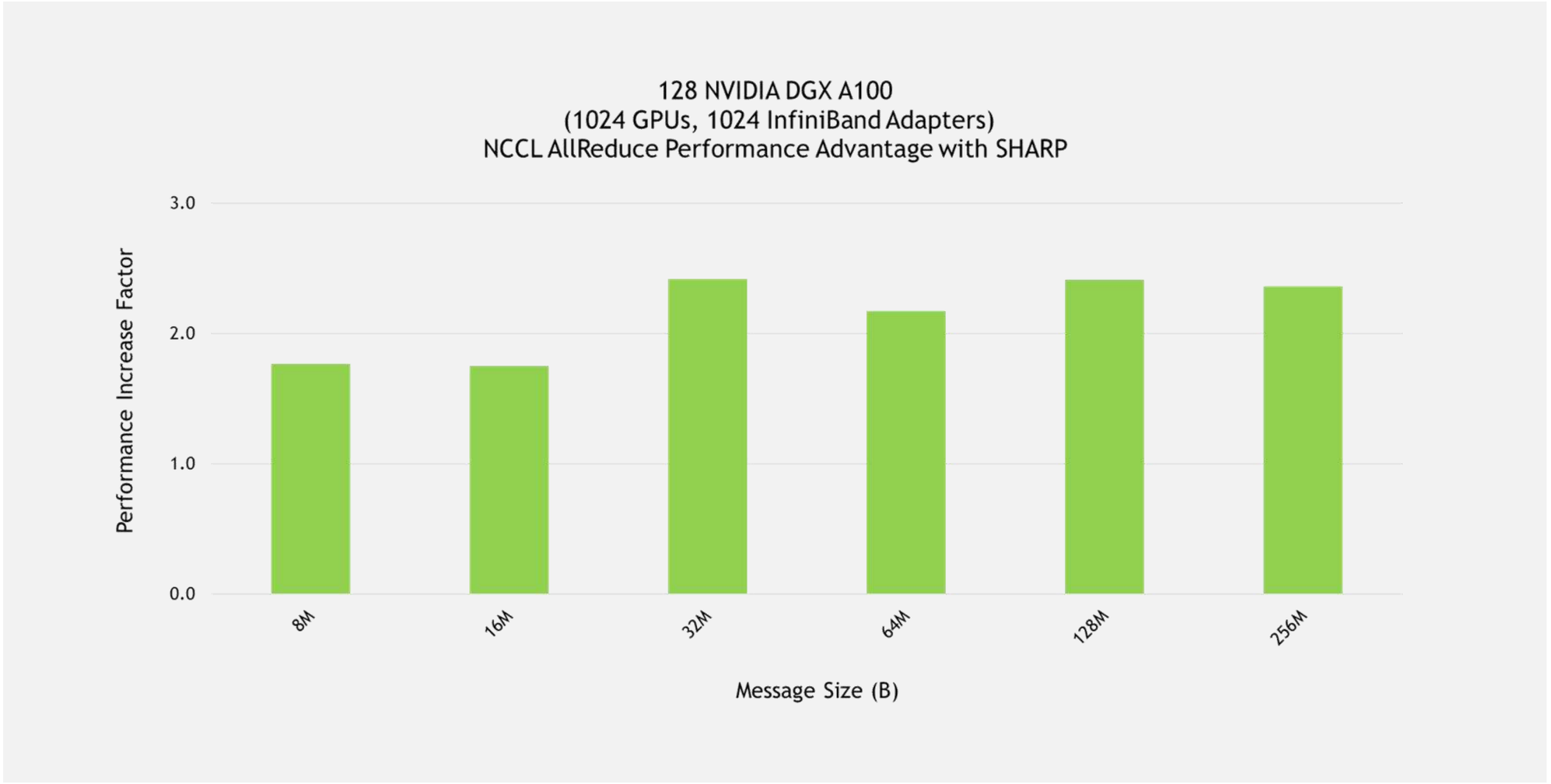




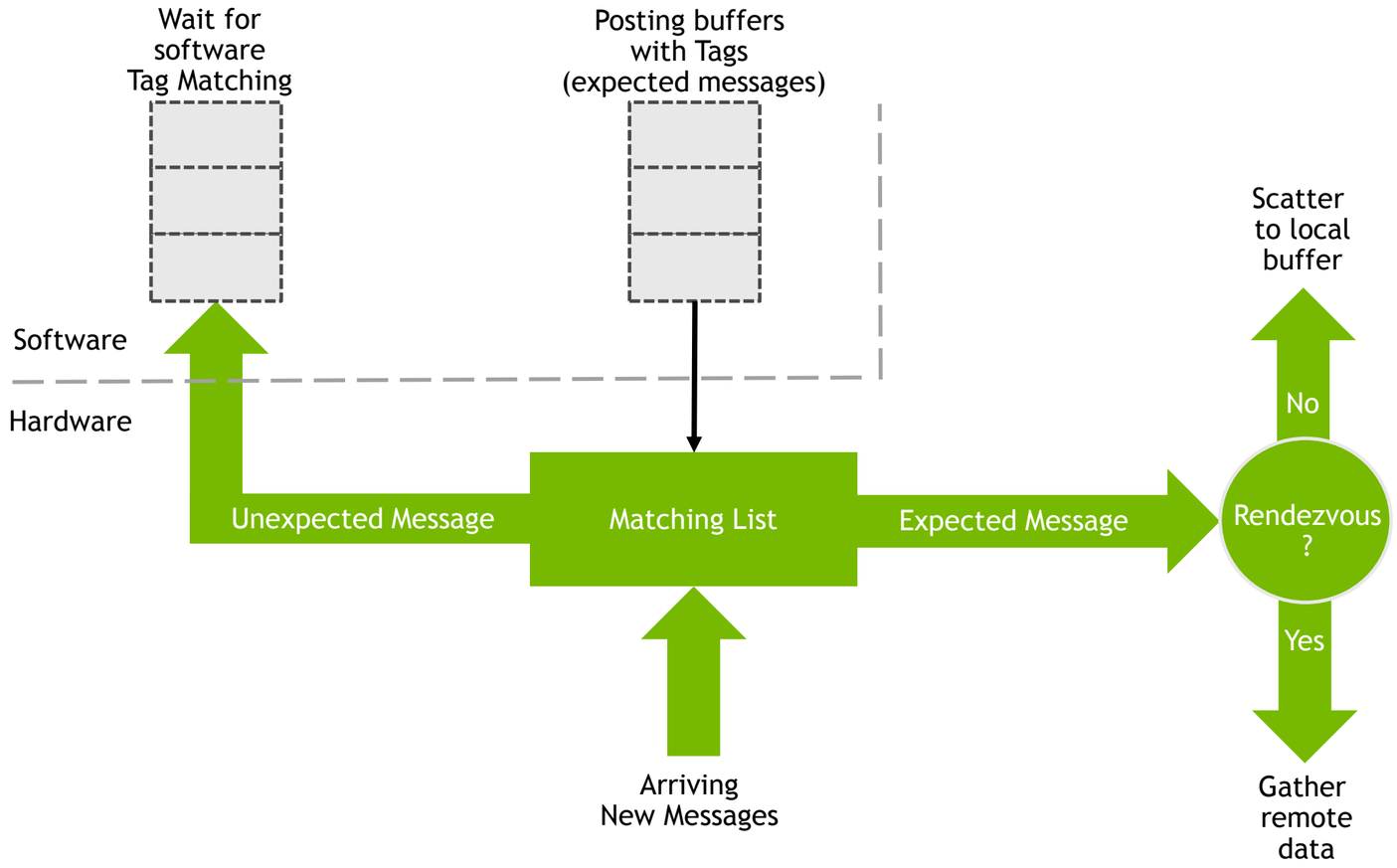


INFINIBAND SHARP AI PERFORMANCE ADVANTAGE

2.5X Higher Performance

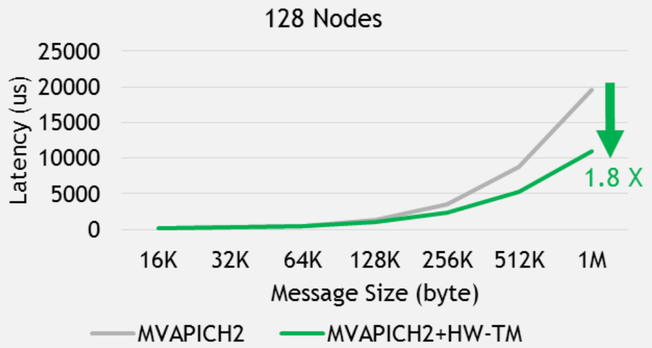


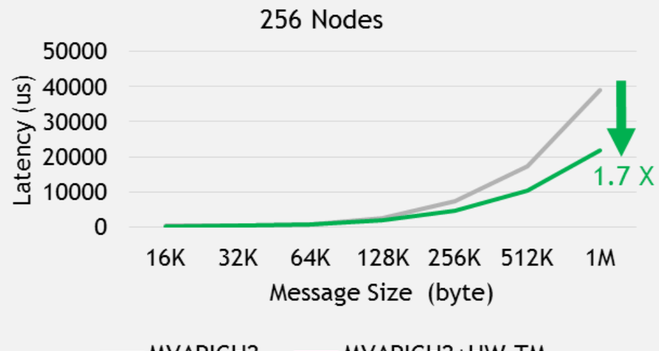
INFINIBAND MPI TAG MATCHING HARDWARE ENGINE

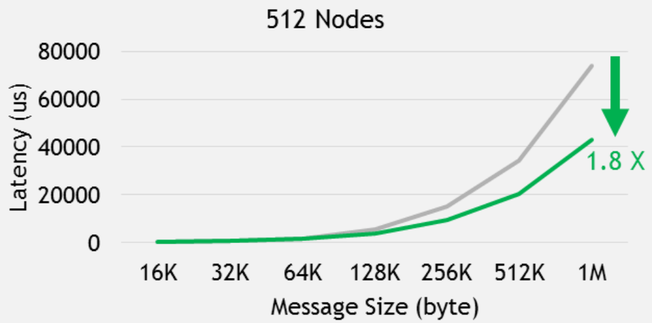


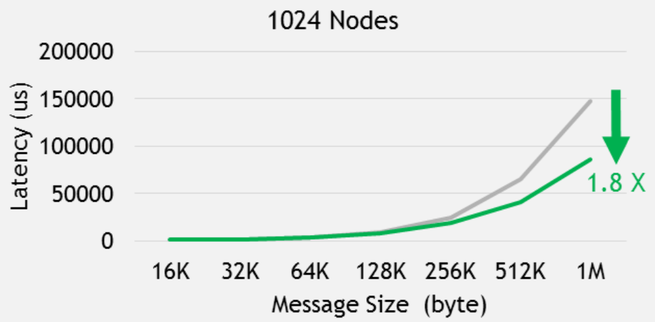
HARDWARE TAG MATCHING PERFORMANCE ADVANTAGES

1.8X Higher MPI_Iscatterv Performance on TACC Frontera





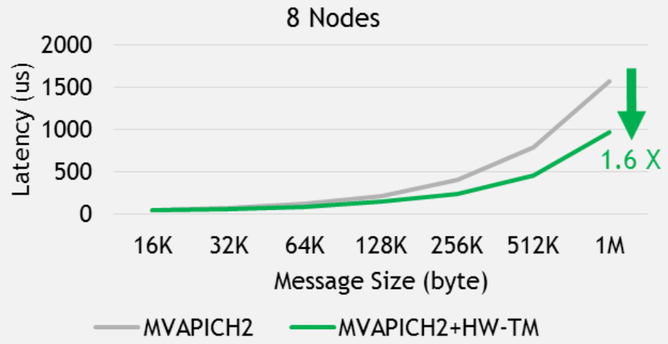


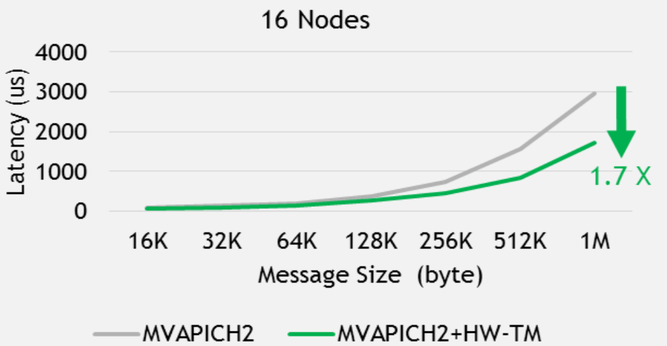


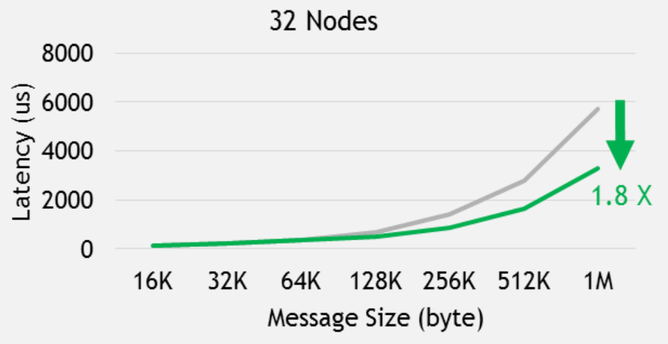
Courtesy of Dhabaleswar K. (DK) Panda
Ohio State University

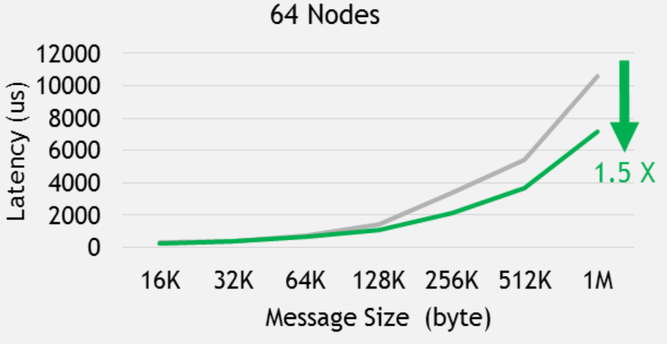
HARDWARE TAG MATCHING PERFORMANCE ADVANTAGES

1.8X higher MPI_Ialltoall Performance on TACC Frontera





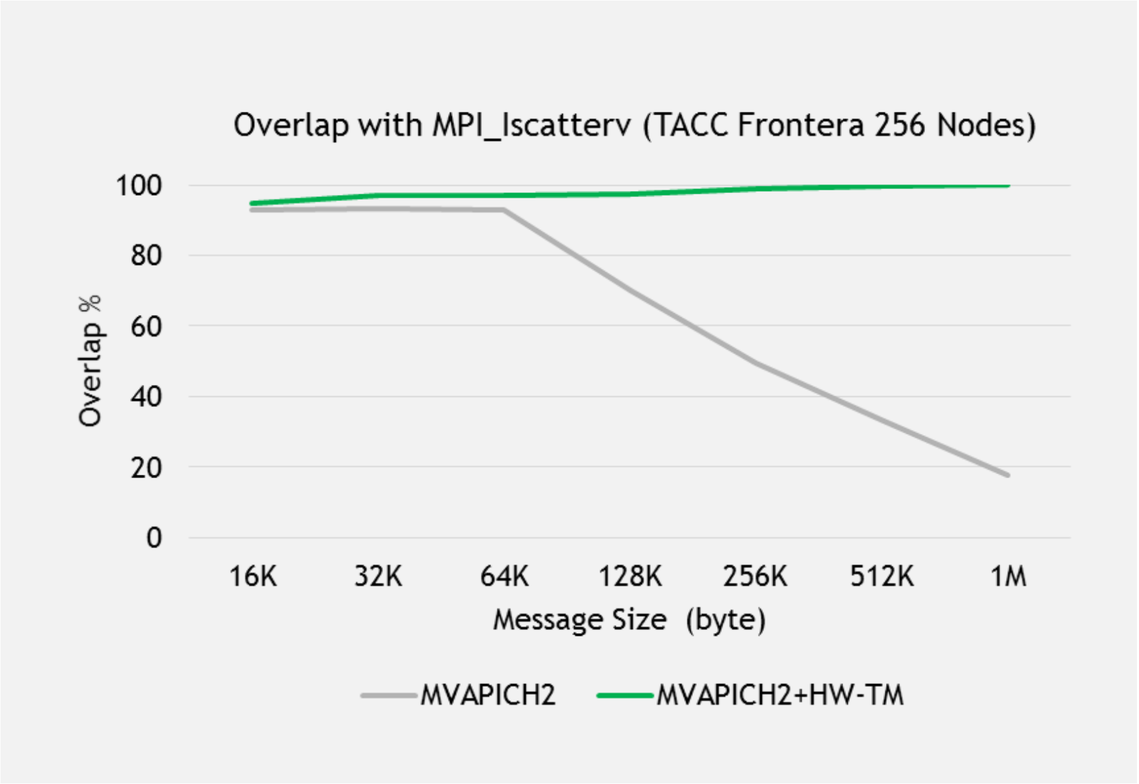


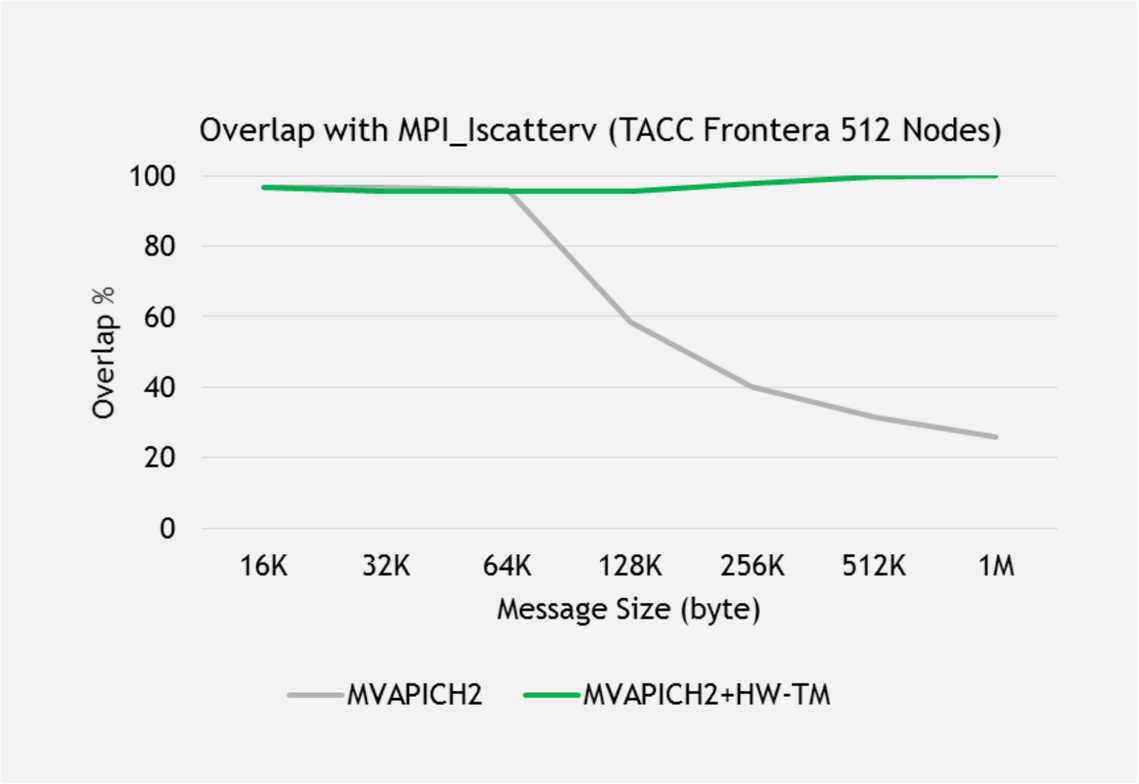


Courtesy of Dhabaleswar K. (DK) Panda
Ohio State University

HARDWARE TAG MATCHING PERFORMANCE ADVANTAGES

Nearly 100% Compute – Communication Overlap

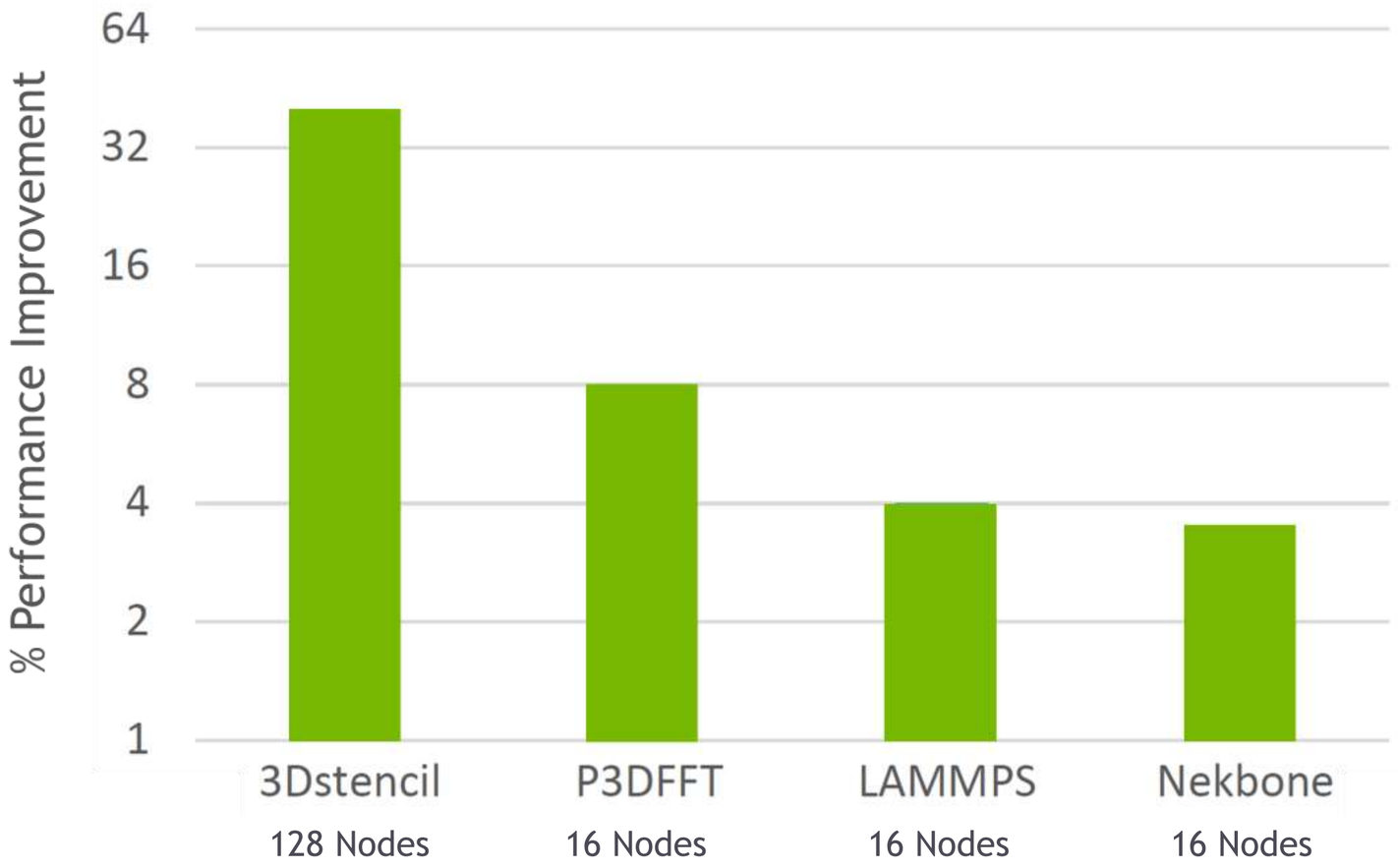




Courtesy of Dhabaleswar K. (DK) Panda
Ohio State University

HARDWARE TAG MATCHING PERFORMANCE ADVANTAGES

Maximizing communication / computations overlap leads to higher applications performance

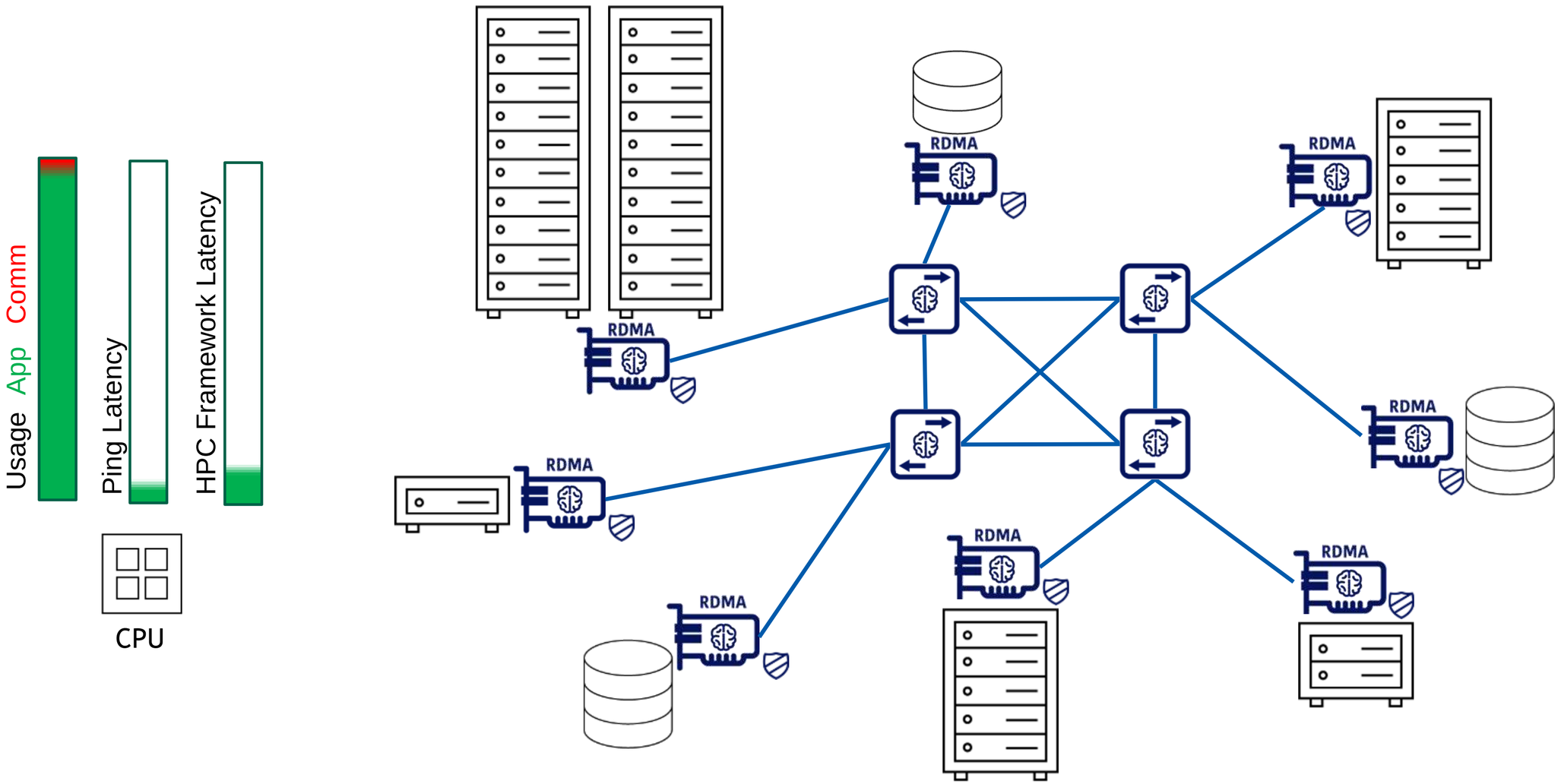


TACC
TEXAS ADVANCED COMPUTING CENTER

OHIO STATE
MVAPICH

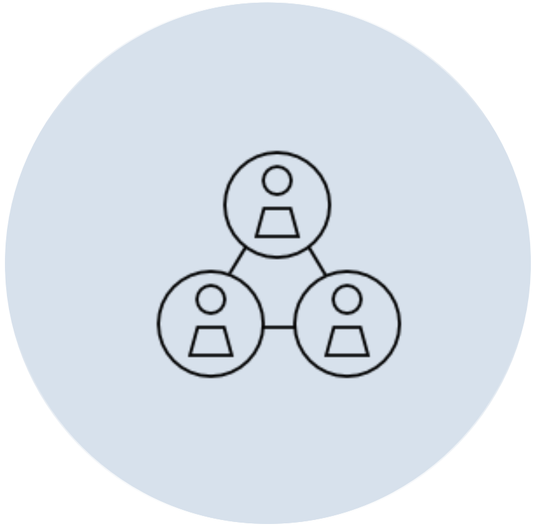
Courtesy of Dhabaleswar K. (DK) Panda
Ohio State University

SECURED IN-NETWORK COMPUTING DATA CENTER



CLOUD NATIVE SUPERCOMPUTER

BARE-METAL PERFORMANCE

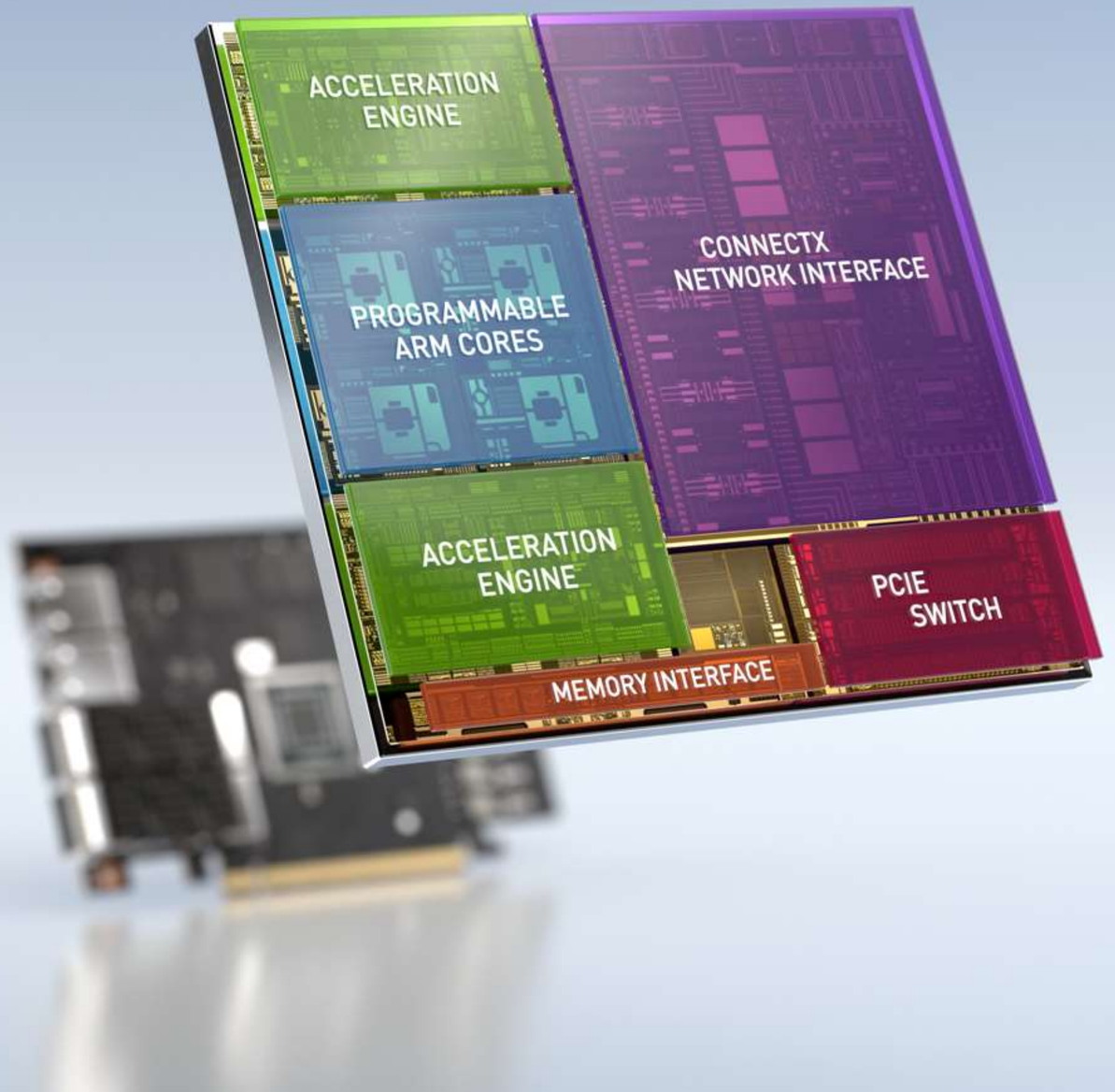


MULTI TENANCY

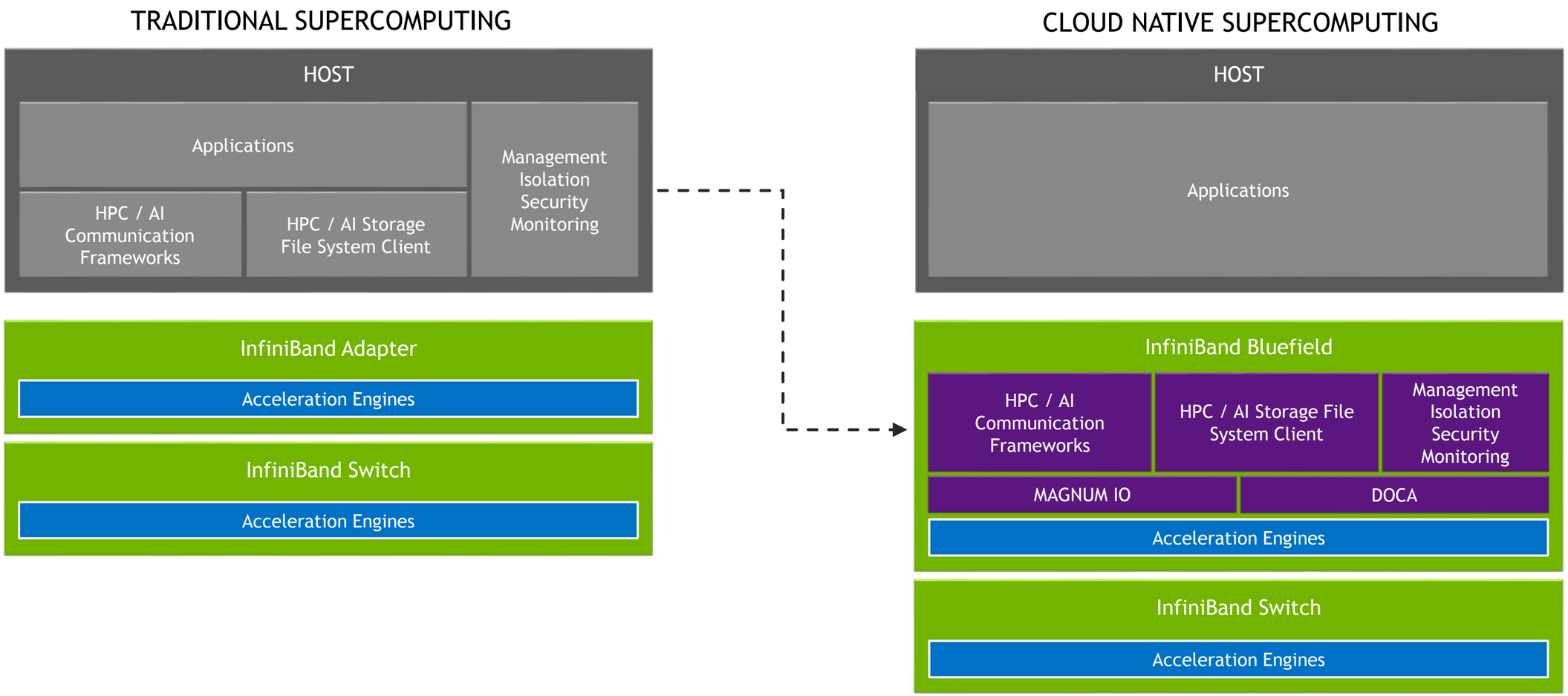
SECURITY

CONFIGURABLE SERVICES

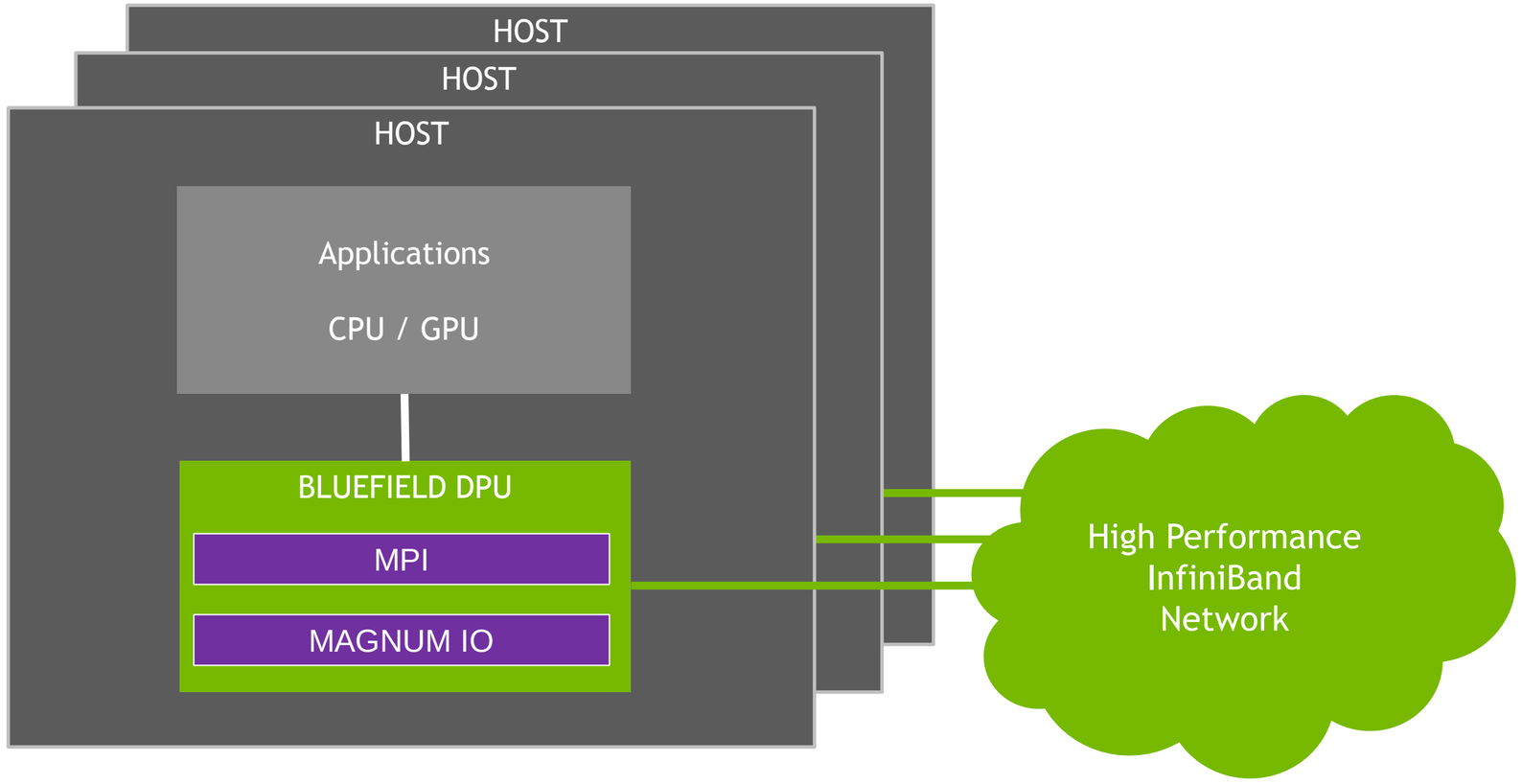
BLUEFIELD DPU – THE CLOUD-NATIVE SUPERCOMPUTING INFRASTRUCTURE PLATFORM

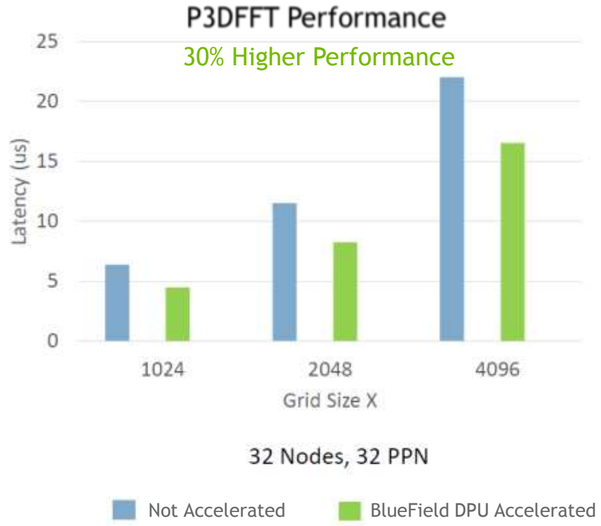


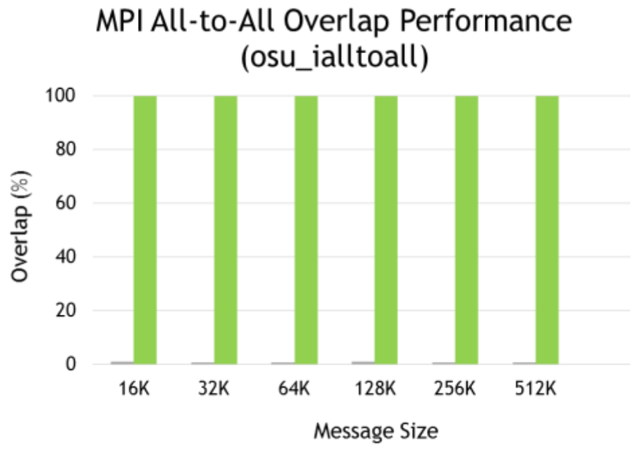
BLUEFIELD DPU - THE CLOUD NATIVE SUPERCOMPUTING INFRASTRUCTURE PLATFORM



BLUEFIELD DPU - HPC AND AI COMMUNICATION FRAMEWORKS OFFLOAD







Courtesy of Ohio State University MVAPICH team and X-ScaleSolutions

MVAPICH

X-ScaleSolutions

Eight servers, Dual Socket Intel® Xeon® 16-core CPUs E5-2697A V4 @ 2.60 GHz (32 processes per node), NVIDIA BlueField-2 HDR100 DPUs and ConnectX-6 HDR100 adapters, NVIDIA Mellanox HDR Quantum Switch QM7800 40-Port 200Gb/s HDR InfiniBand, 256GB DDR4 2400MHz RDIMMs memory and 1TB 7.2K RPM SATA 2.5" hard drive per node.

UFM CYBER-AI

Management, Monitoring, Orchestration, Cyber Intelligence and Analytics

Network setup, connectivity validation and secure cable management

Automated network discovery and network provisioning

Network telemetry and traffic monitoring, congestion discovery

Performance, health and fault monitoring

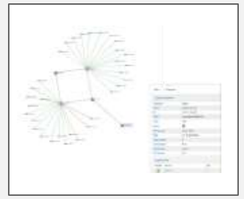
Centralized management for global software updates and configuration

Job scheduler provisioning, network provisioning

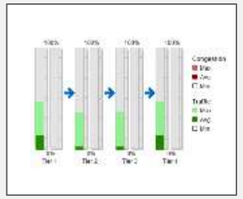
Detects performance degradations, anomalies and usage changes

Provides alerts of abnormal system and application behavior

Provides alerts for potential system failures

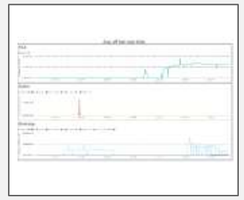


Network Validation

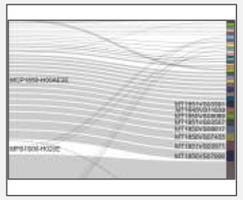


Congestion Mapping

Prediction Dashboard



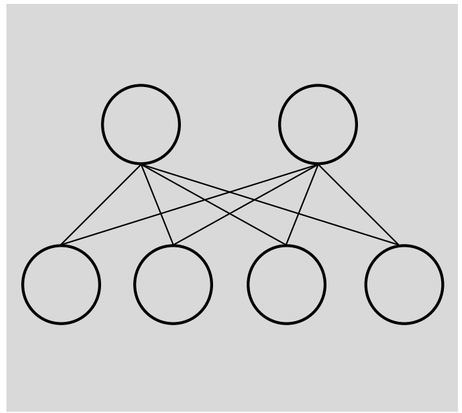
Real-Time Analysis



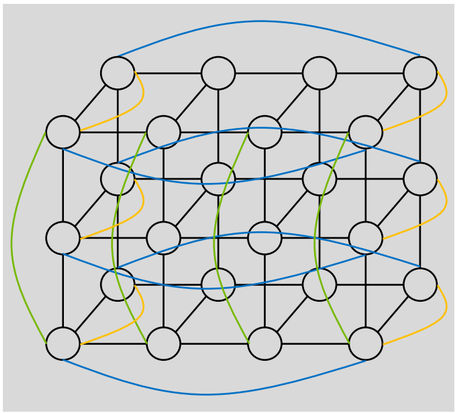
Performance Monitoring

Secure Cable Management

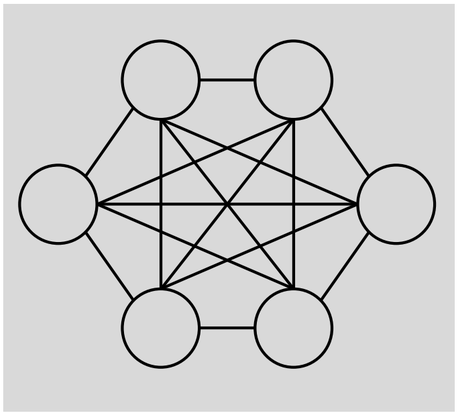
SUPPORTING VARIETY OF TOPOLOGIES



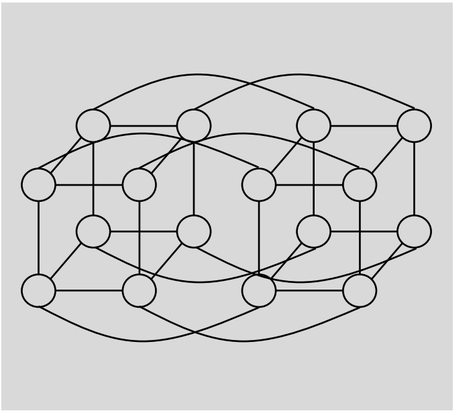
Fat Tree



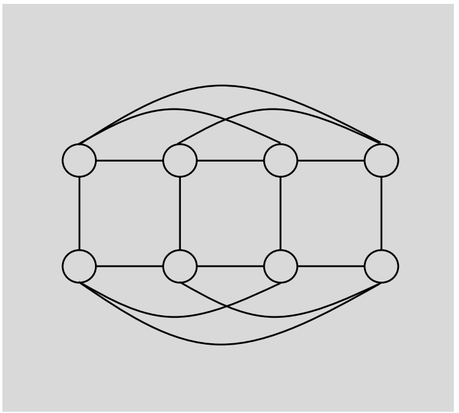
Torus



Dragonfly



Hypercube



HyperX

NETWORK TOPOLOGIES

Fat Tree
A common topology
Full-bandwidth (1:1) is idea for capability clusters – when a single job must provide super performance

Dragonfly+
Tradeoff cost versus worst traffic pattern bandwidth (2:1)
Efficient for capacity clusters – when many jobs are running together
Better performance than Fat Tree 2:1 as most traffic patterns are evenly distributed
Grow at zero cost (no need to reserve capacity or re-cable) – this is unique value of Dragonfly+

3D,4D,5D,6D Torus
Most efficient for 3D,4D,5D,6D neighbor traffic – depends on the problem types run on the cluster
Low bandwidth and much higher latency for capability clusters with arbitrary application type

HyperCube
Nicely support some specific algorithm, but higher latency and lower bandwidth with higher cost

HyperX - Generalized Hypercube
Less flexible and scalable than Dragonfly+
Similar performance for 3D worse cost and performance for 4D and above

DRAGONFLY+ TOPOLOGY

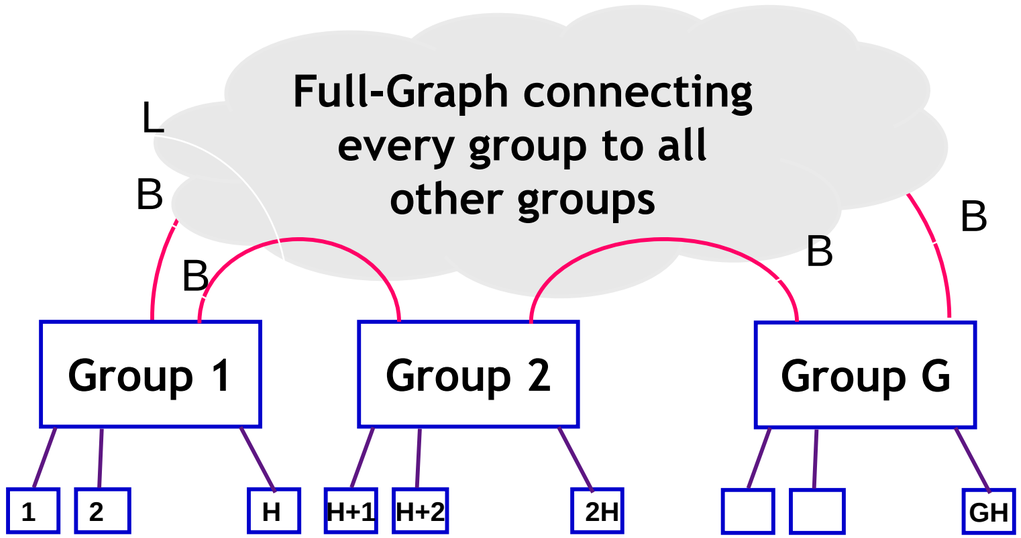
Several “groups”, connected using all to all links

The topology inside each group can be any topology

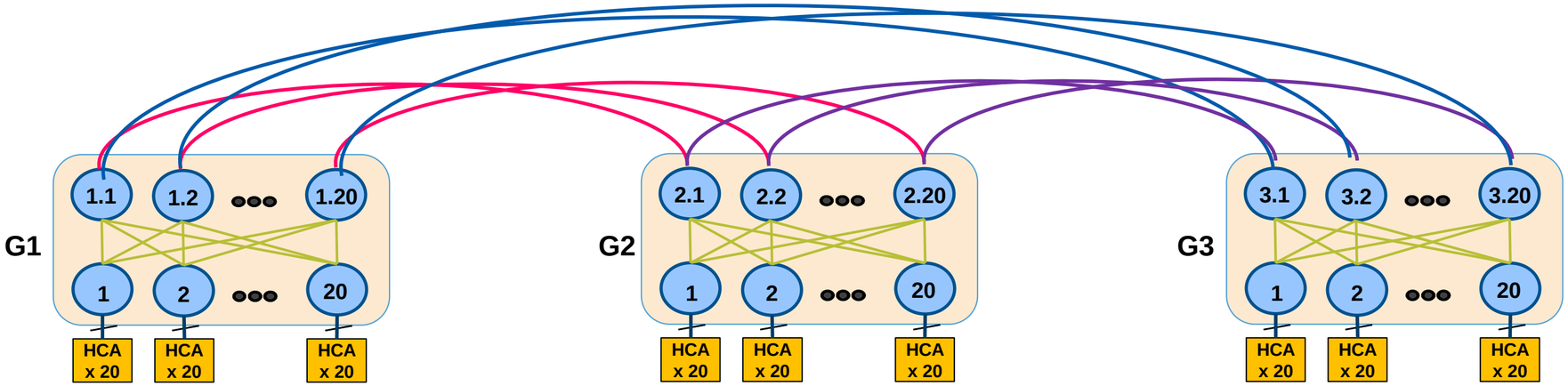
Reduce total cost of network (fewer long cables)

Utilizes Adaptive Routing to for efficient operations

Simplifies future system expansion



1200-Nodes Dragonfly+ Systems Example



DRAGONFLY+ TOPOLOGY

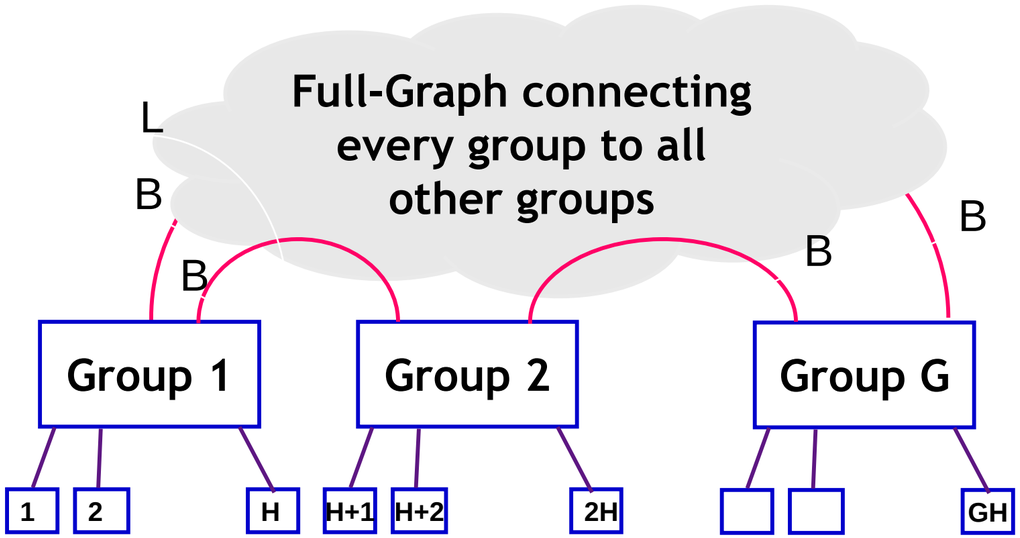
Several “groups”, connected using all to all links

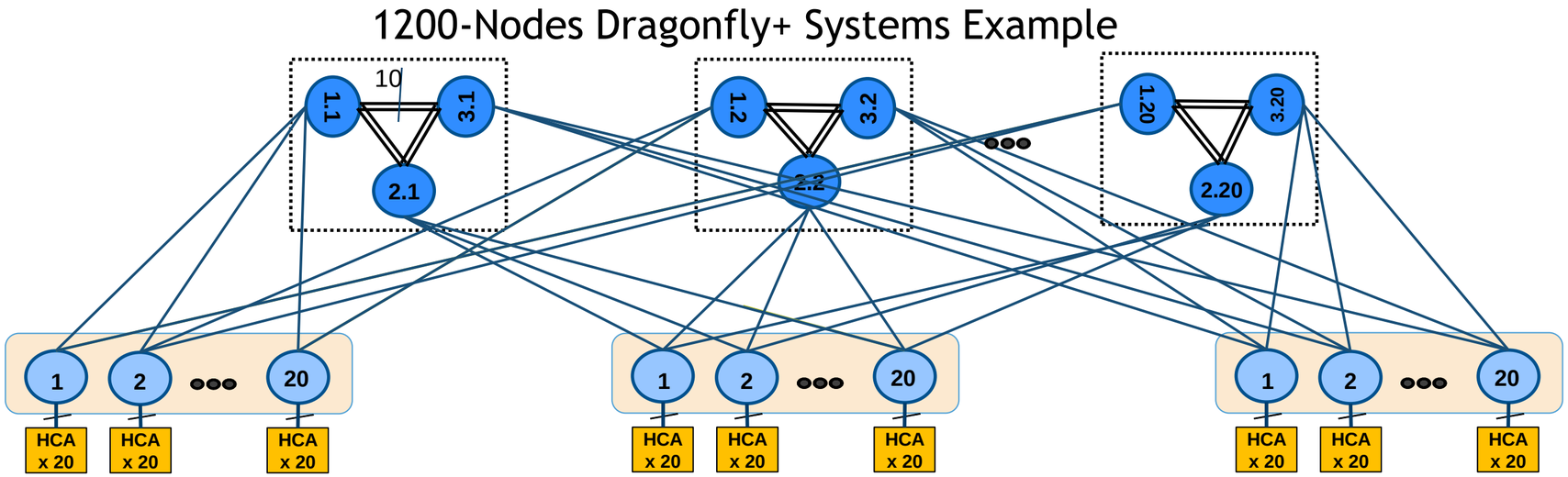
The topology inside each group can be any topology

Reduce total cost of network (fewer long cables)

Utilizes Adaptive Routing to for efficient operations

Simplifies future system expansion



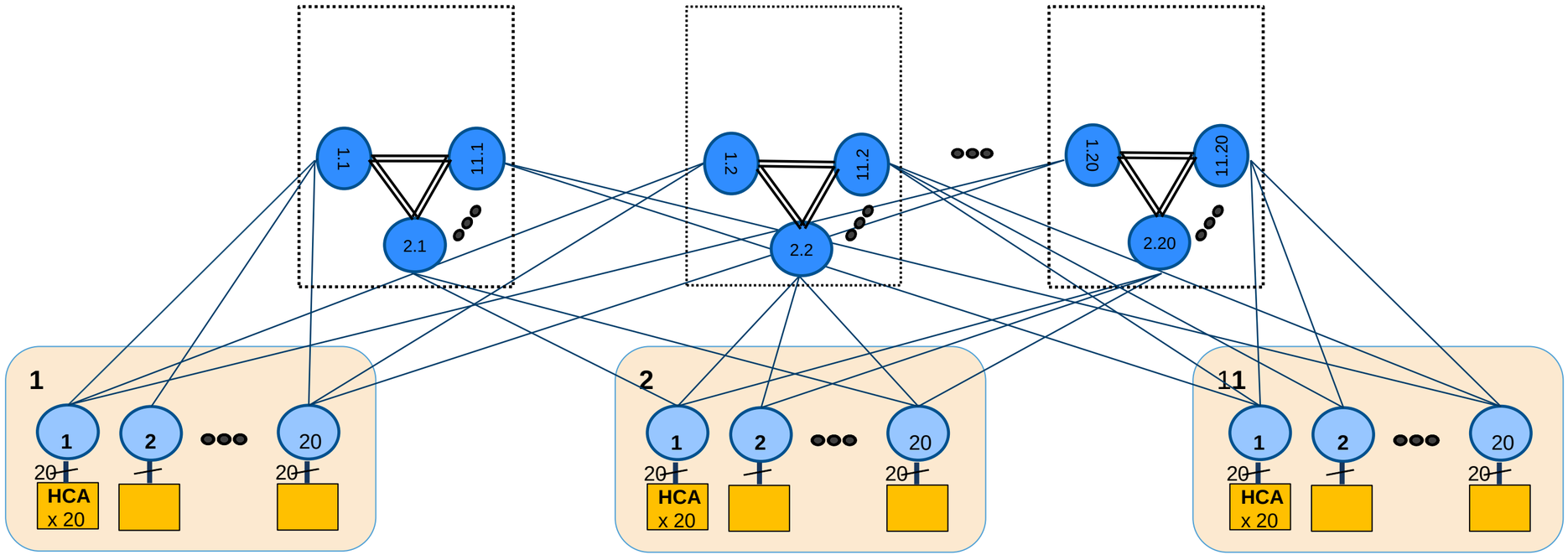


FUTURE EXPANSION OF DRAGONFLY+ BASED SYSTEM

Dragonfly+ is the only topology that allows system expansion at zero cost

While maintaining bisection bandwidth, no port reservation, no re-cabling

Phase 1:
11x400 =
4400 hosts

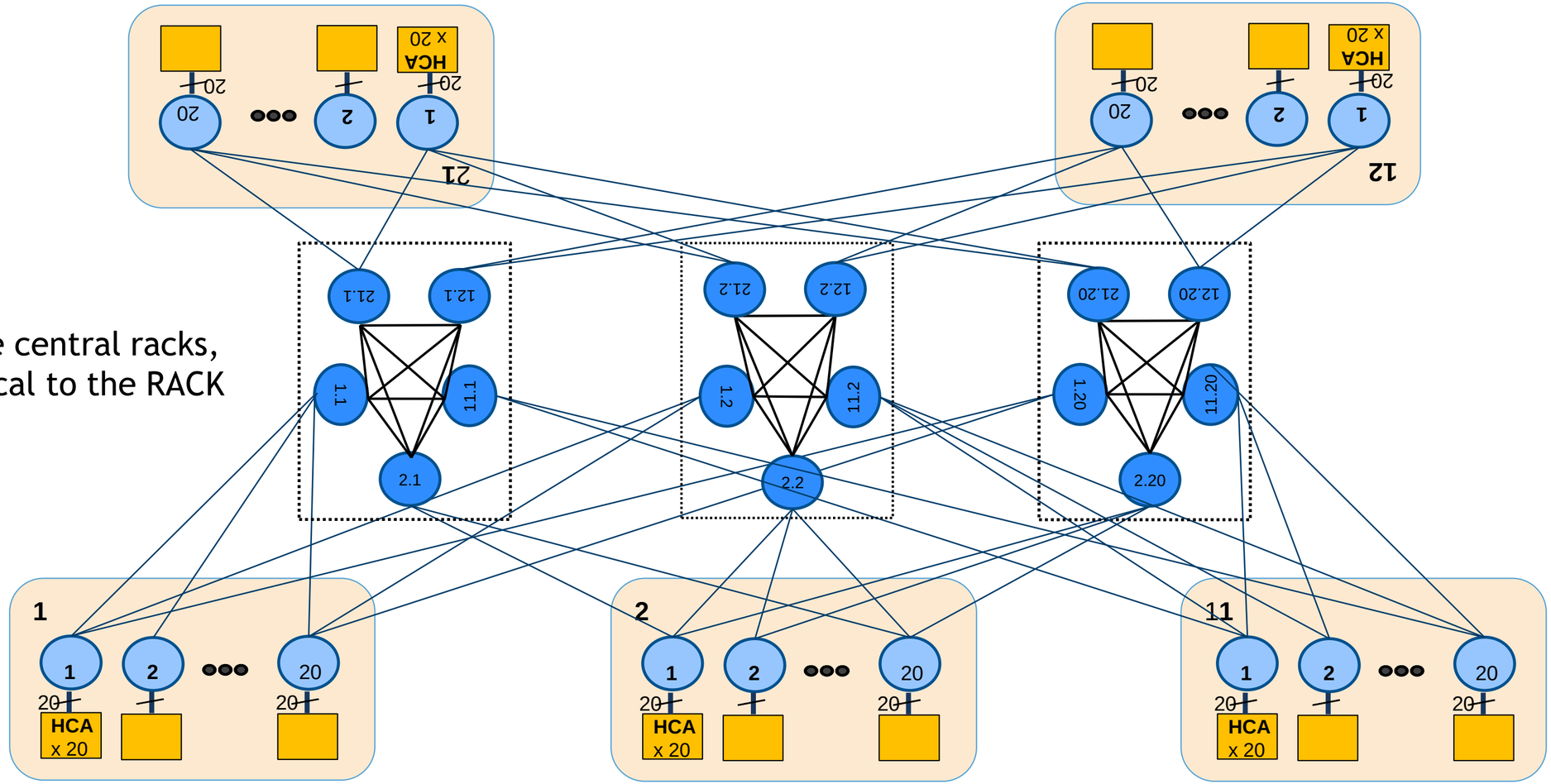


FUTURE EXPANSION OF DRAGONFLY+ BASED SYSTEM

Phase 2:
+10x400 =
8400 hosts

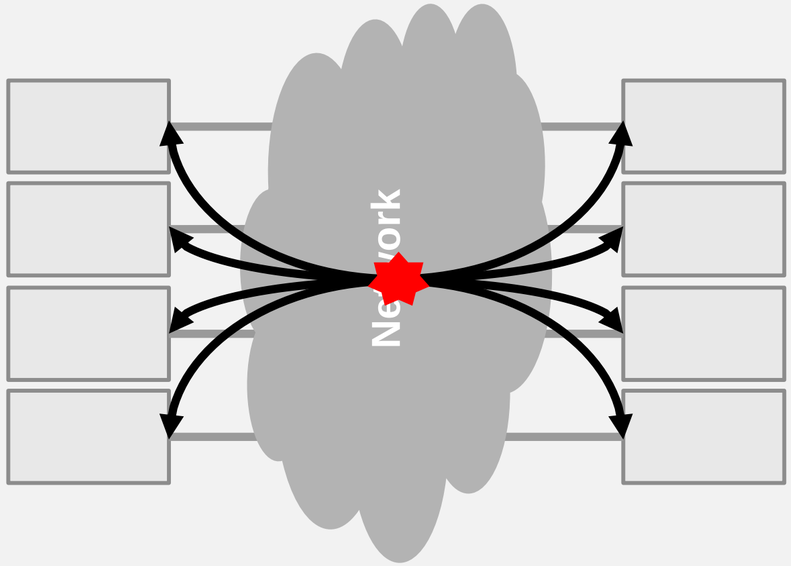
Re-cable the central racks,
a change local to the RACK

Phase 1:
11x400 =
4400 hosts



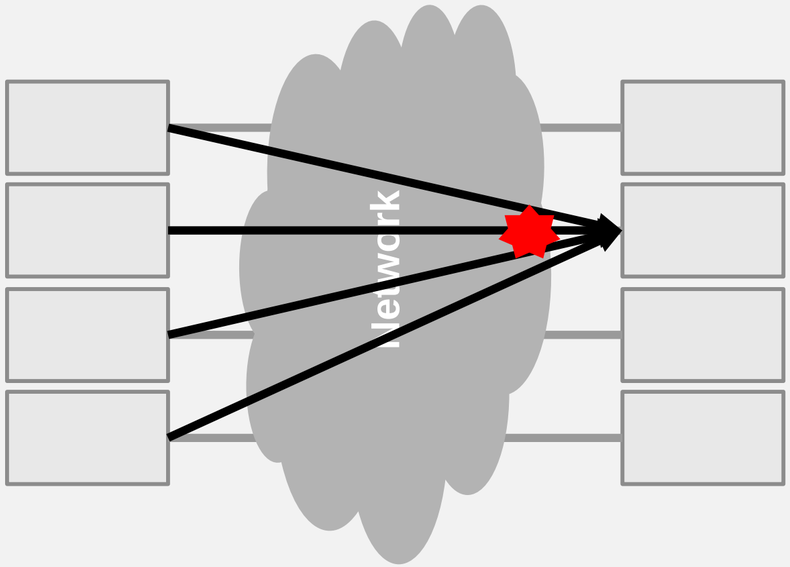
NETWORK CONGESTION TYPES

In-network Congestion



Solution: Adaptive Routing

In-cast Congestion

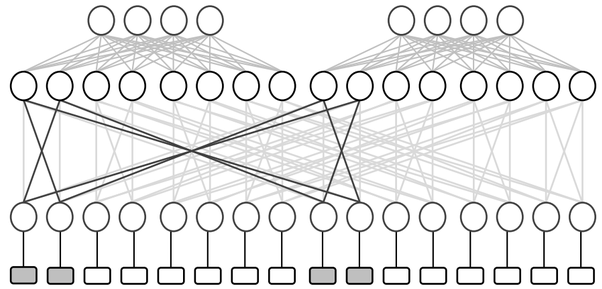


Solution: Congestion Control

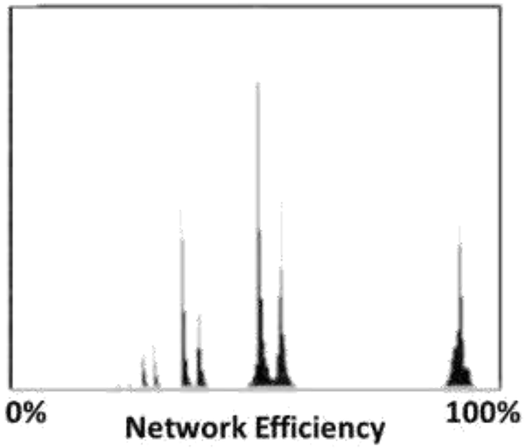
IN-NETWORK CONGESTION: ADAPTIVE ROUTING

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

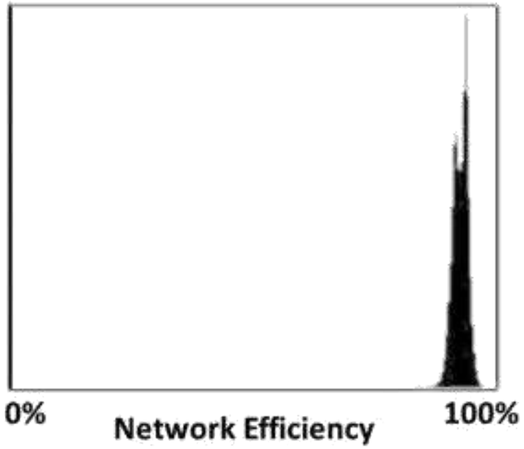
Sudharshan S. Vazhkudai†, Bronis R. de Supinski‡, Arthur S. Bland†, Al Geist†, James Sexton*, Jim Kahle*, Christopher J. Zimmer†, Scott Atchley†, Sarp Oral†, Don E. Maxwell†, Veronica G. Vergara Larrea†, Adam Bertsch‡, Robin Goldstone‡, Wayne Joubert†, Chris Chambreau‡, David Appelhans*, Robert Blackmore*, 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

IN-CAST CONGESTION

Desired behavior

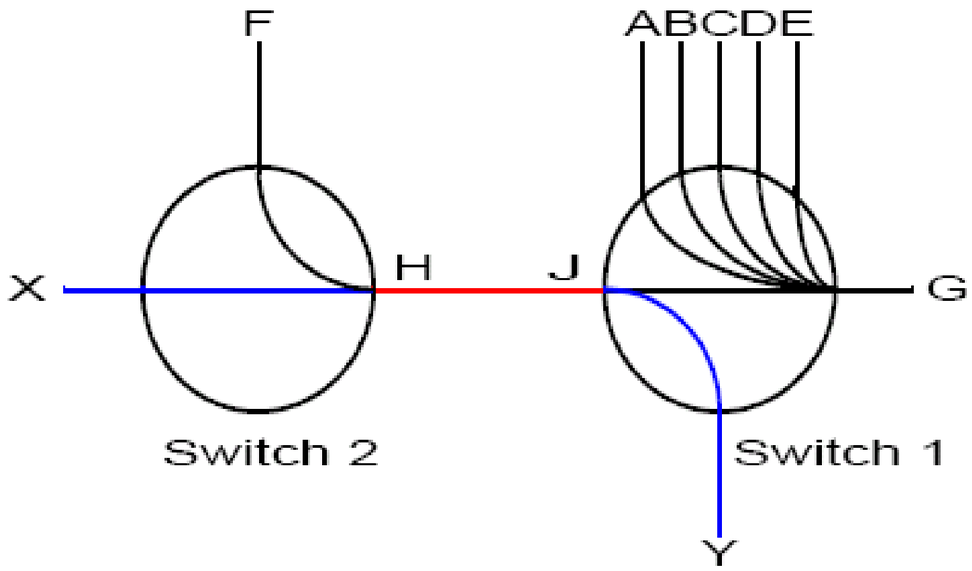
- A-G…F to G – 1/6 link BW
- X to Y – 5/6 link BW

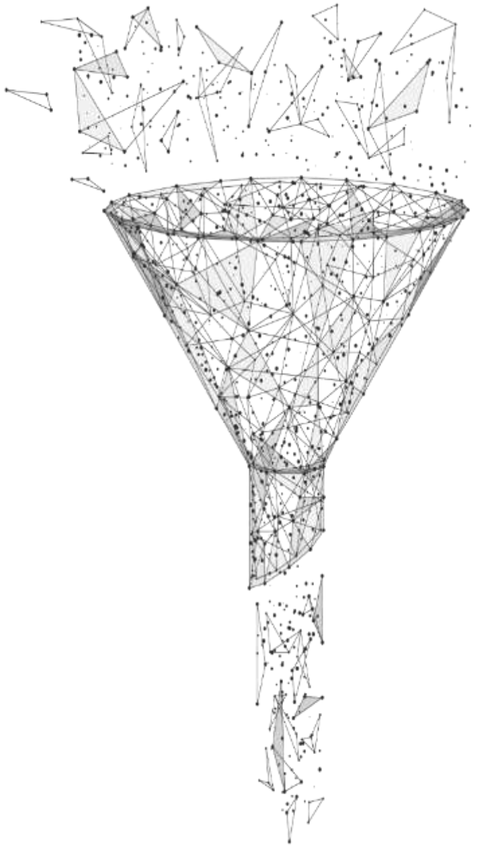
Congestion effect – lossless network:

- A-G…E to G – 1/6 link BW
- F to G – 1/12 link BW
- X to Y – 1/12 link BW (Victim flow)

Congestion effect – lossy network:

- Massive packets' drop
- Application-visible impact



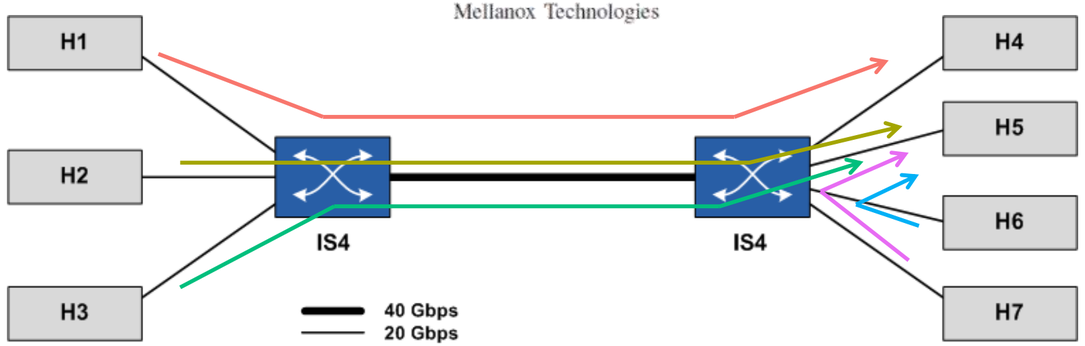


Solution – Slow Down Injection Rate

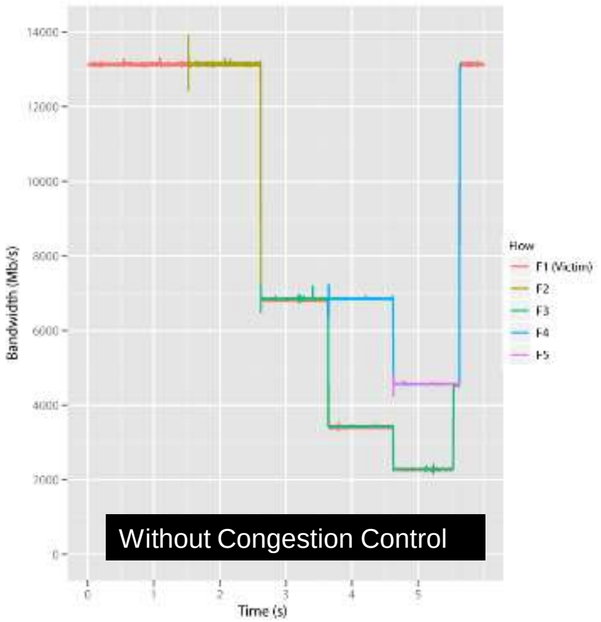
INFINIBAND CONGESTION CONTROL

First Experiences with Congestion Control in InfiniBand Hardware

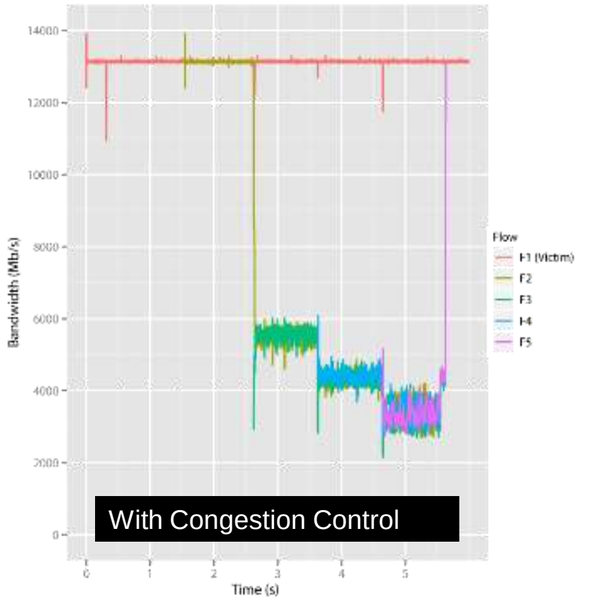
Ernst Gunnar Gran, Magne Eimot, Sven-Arne Reinemo, Tor Skeie, Olav Lysne *Member, IEEE*
Simula Research Laboratory
and
Gilad Shainer - Shainer@Mellanox.com
Mellanox Technologies



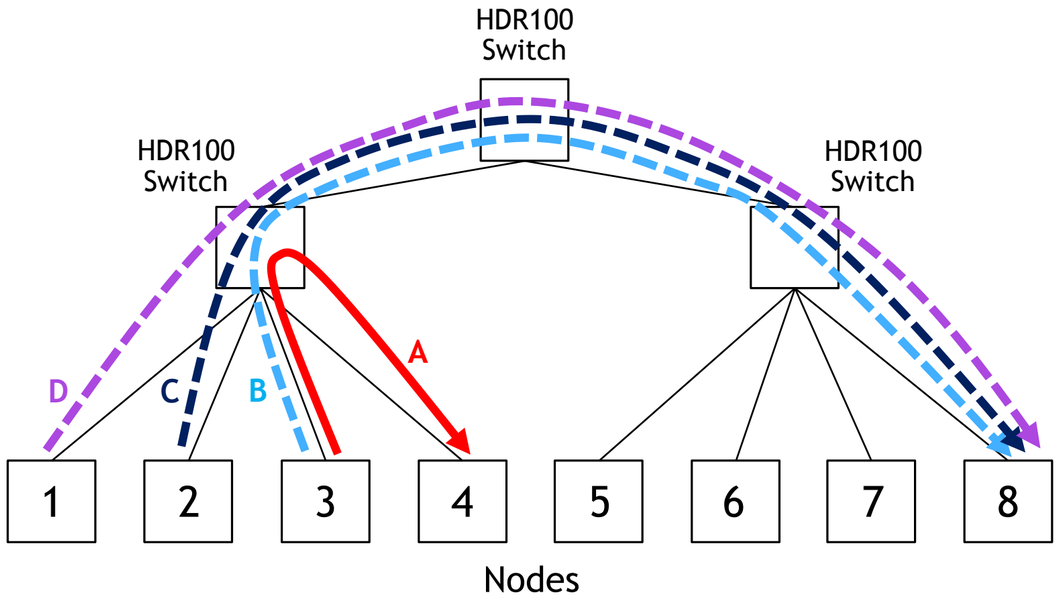
Congestion – Throughput loss



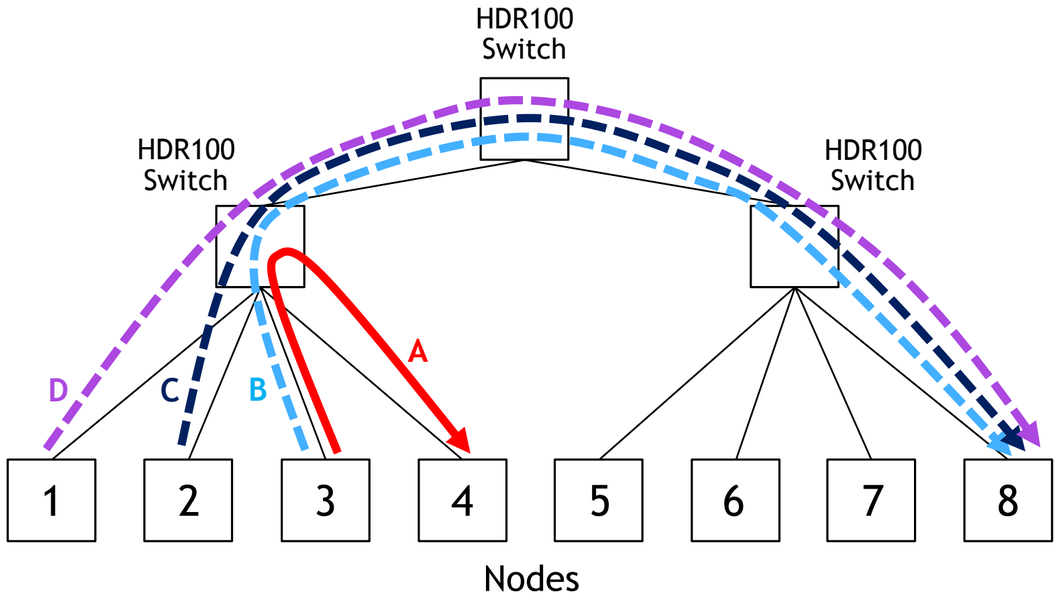
No congestion – highest throughput!

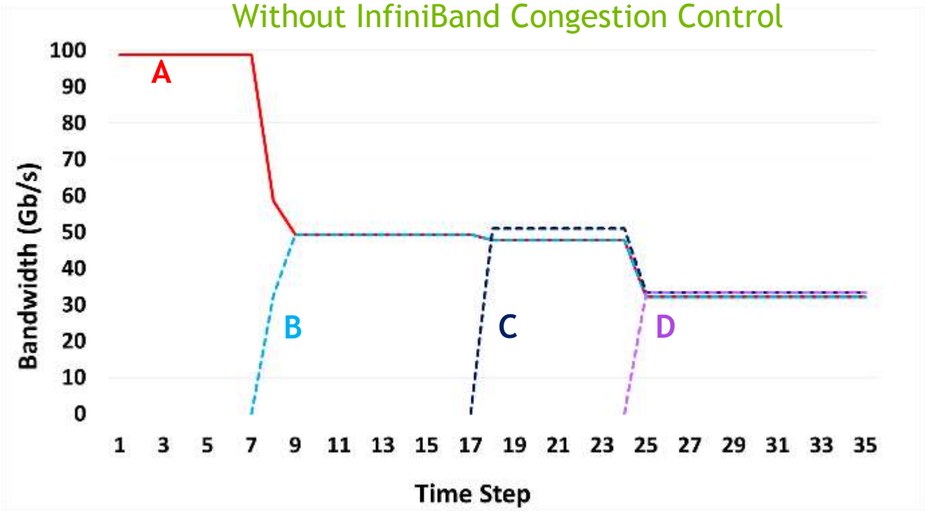


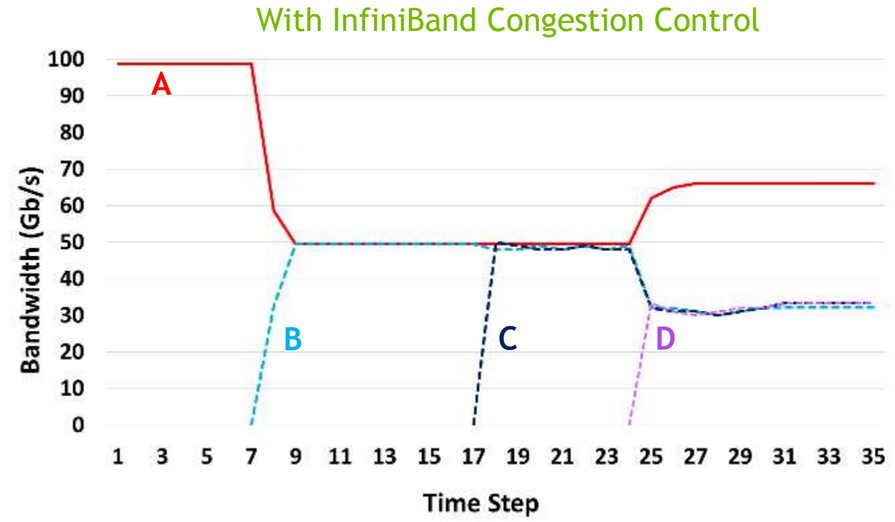
HDR INFINIBAND CONGESTION CONTROL



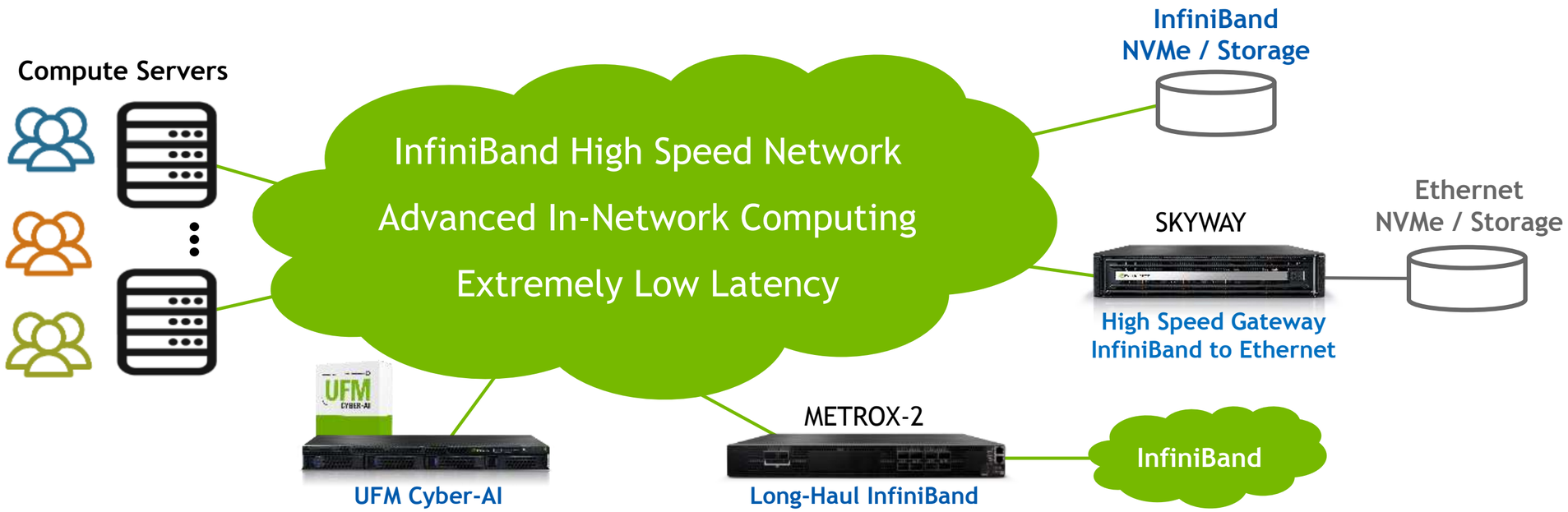
HDR INFINIBAND CONGESTION CONTROL







INFINIBAND ACCELERATED DATA CENTER



High data throughput, extremely low latency, high message rate, RDMA, GPU Direct RDMA, GPU Direct Storage

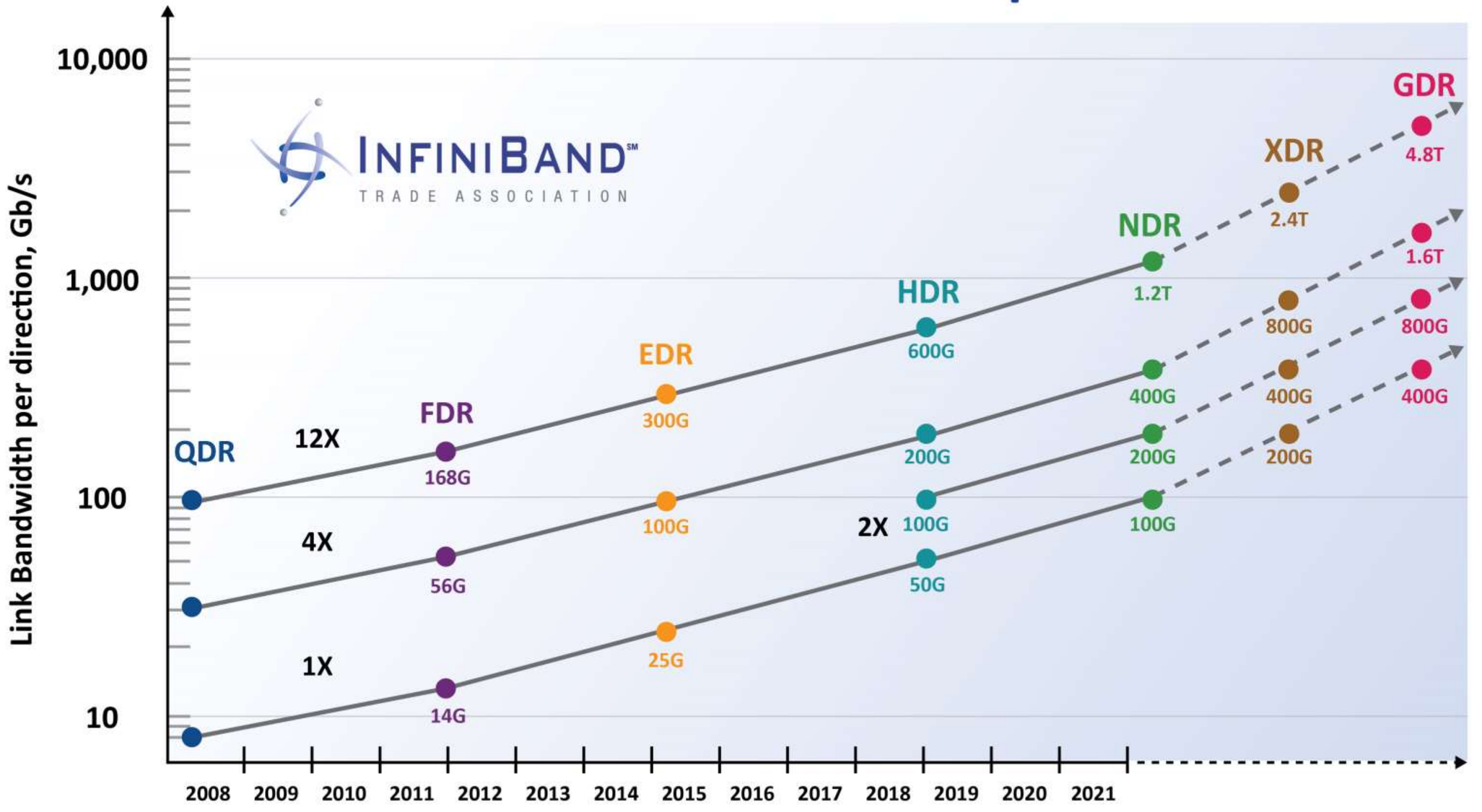
Advanced adaptive routing, congestion control and quality of service for highest network efficiency

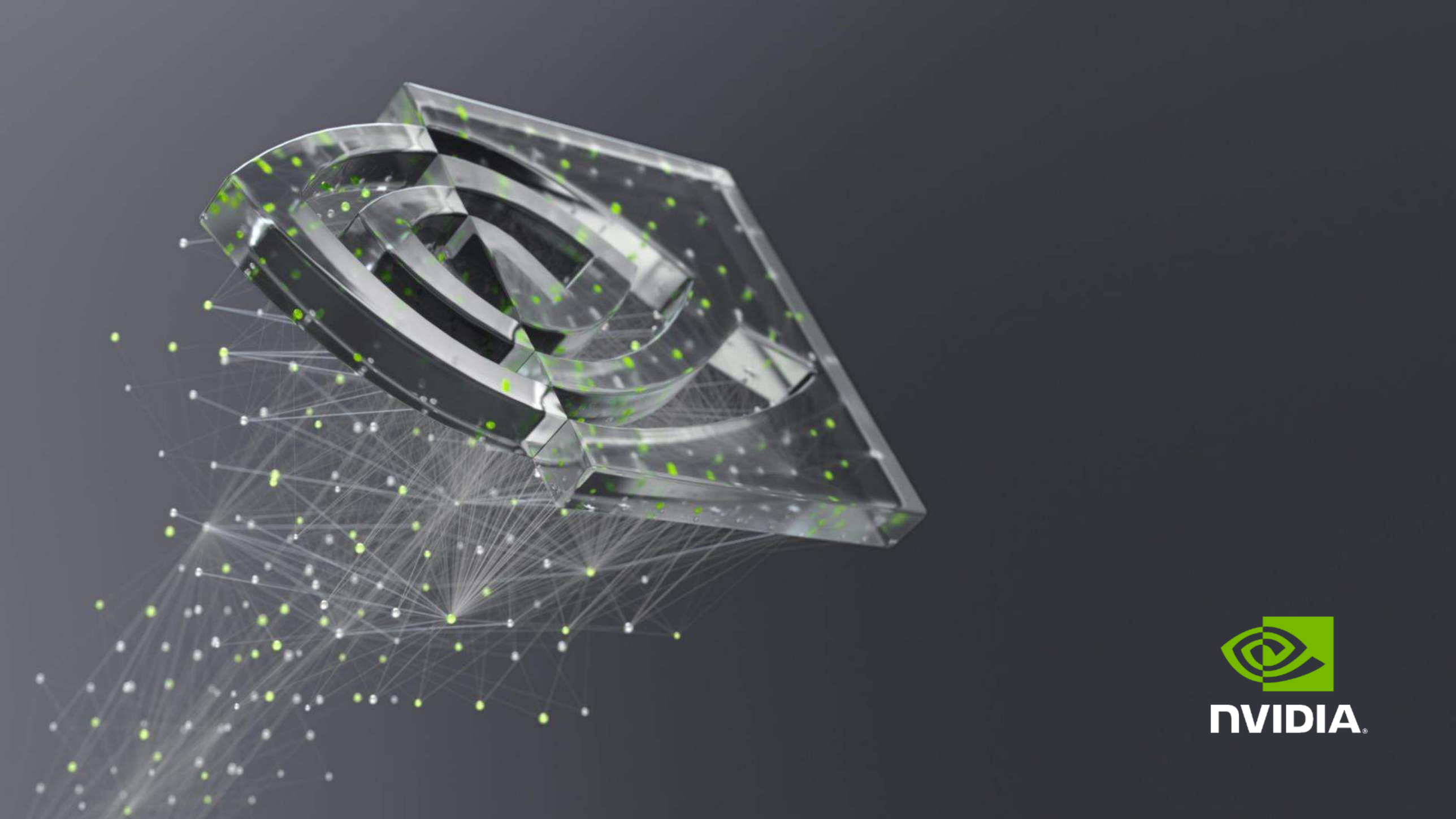
In-Network Computing engines for accelerating applications performance and scalability

Self Healing Network for highest network resiliency

Standard - backward and forward compatibility – protecting datacenter investments

InfiniBand Roadmap




NVIDIA