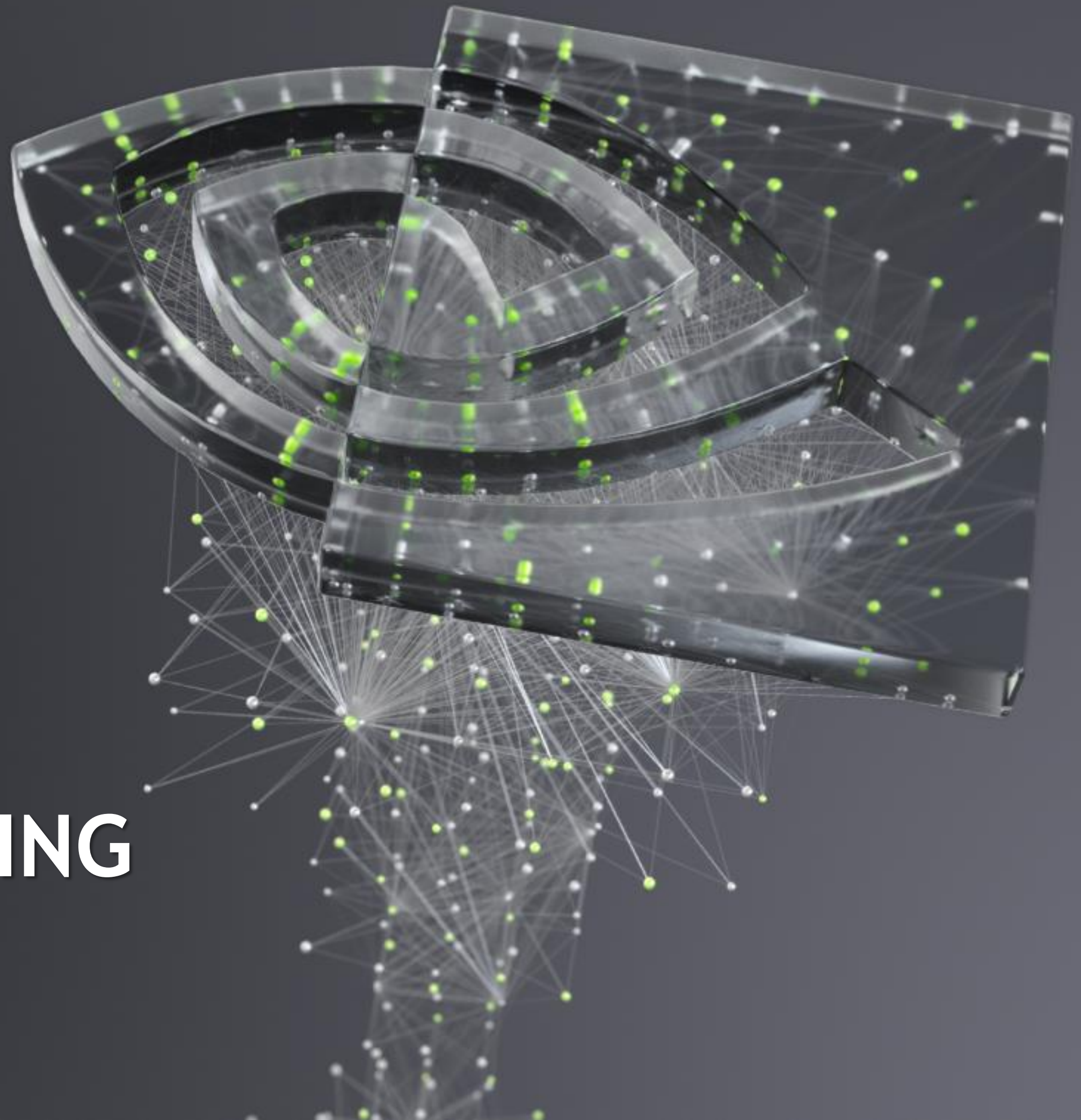


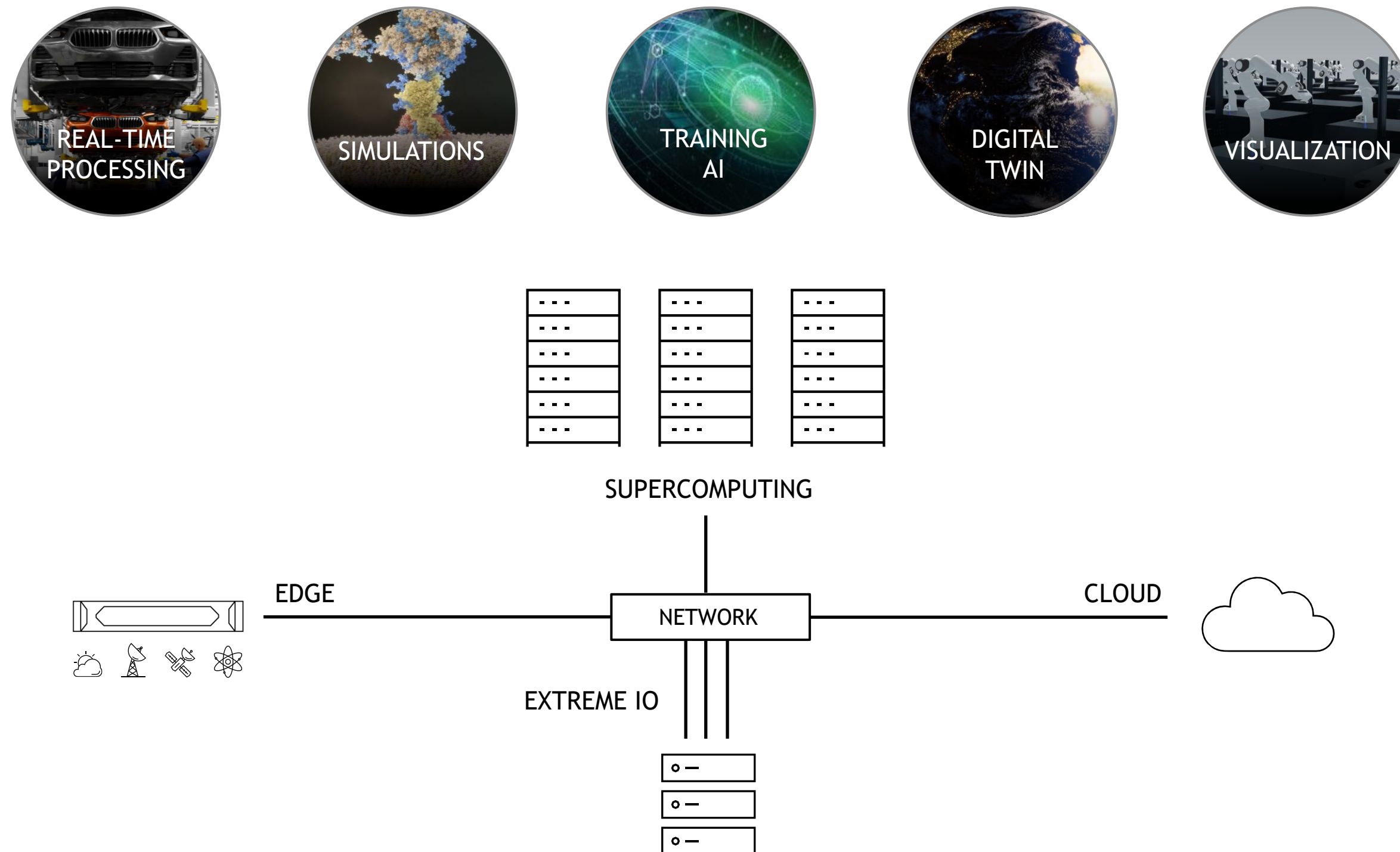


ACCELERATING SCIENTIFIC COMPUTING

NVIDIA InfiniBand Networking, MUG August 2021



EXPANDING UNIVERSE OF SCIENTIFIC COMPUTING



NVIDIA 200G HDR INFINIBAND INFRASTRUCTURE

In-Network Computing Accelerated Network for Supercomputing



Metrox-2 Long-haul



Skyway Gateway



UFM Cyber-AI



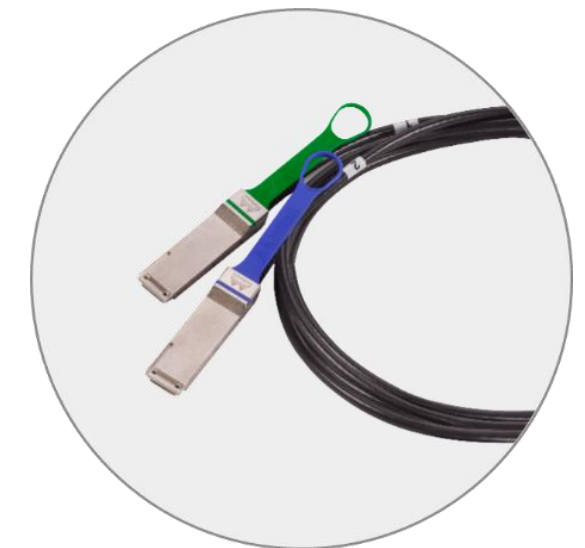
ConnectX Adapter



BlueField DPU



Quantum Switch



Linkx

IN-NETWORK COMPUTING ACCELERATED SUPERCOMPUTING

Software-Defined, Hardware-Accelerated, InfiniBand Network

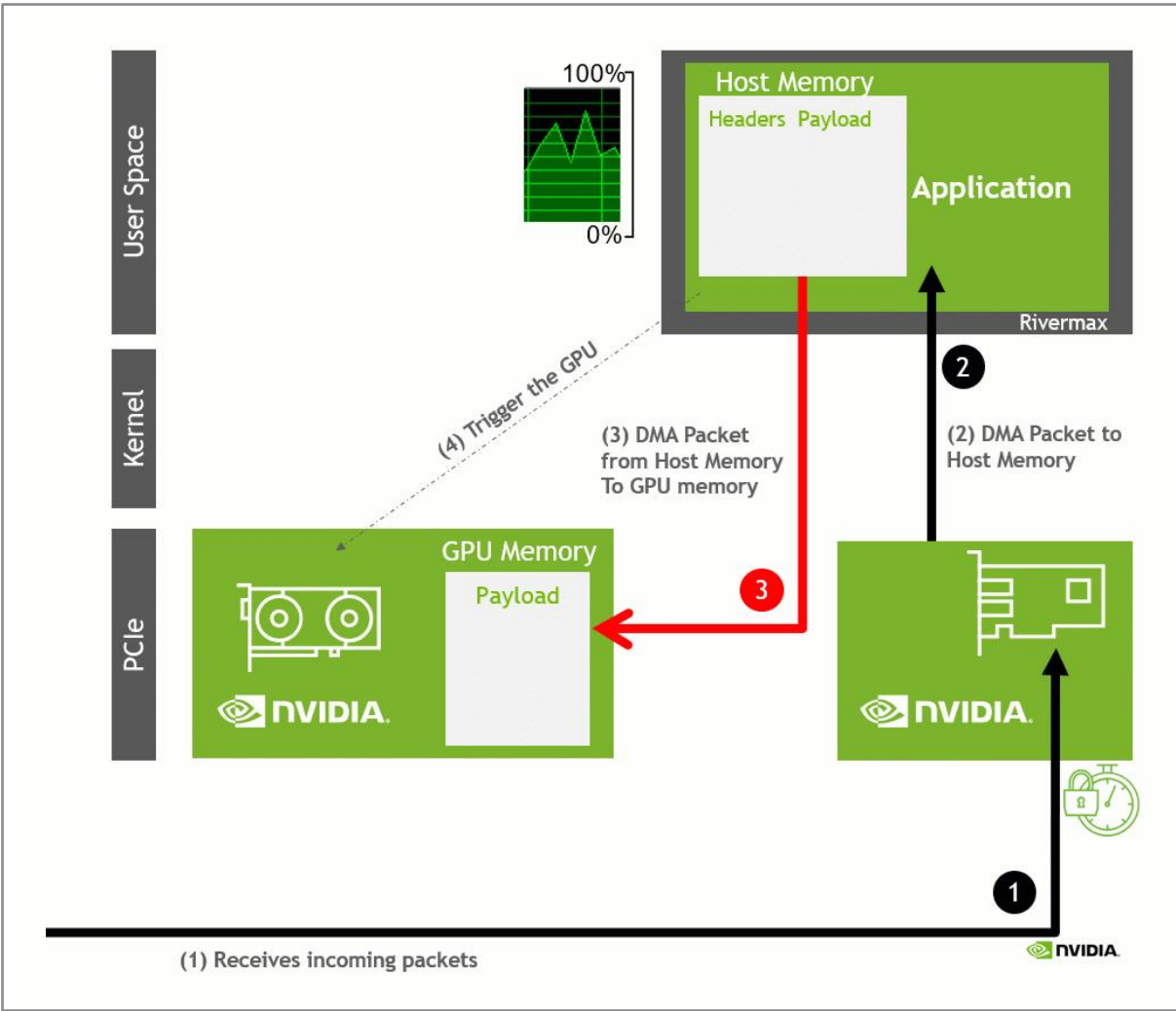
Most Advanced Networking

End-to-End	High Throughput	Extremely Low Latency	High Message Rate
	RDMA	GPUDirect RDMA	GPUDirect Storage
	Adaptive Routing	Congestion Control	Smart Topologies

In-Network Computing

Adapter/DPU	All-to-All	MPI Tag Matching	Data Reductions (SHARP)	Switch
	Programmable Datapath Accelerator	Data processing units (Arm cores)	Self Healing Network	
End-to-End	Data security / tenant isolation			End-to-End

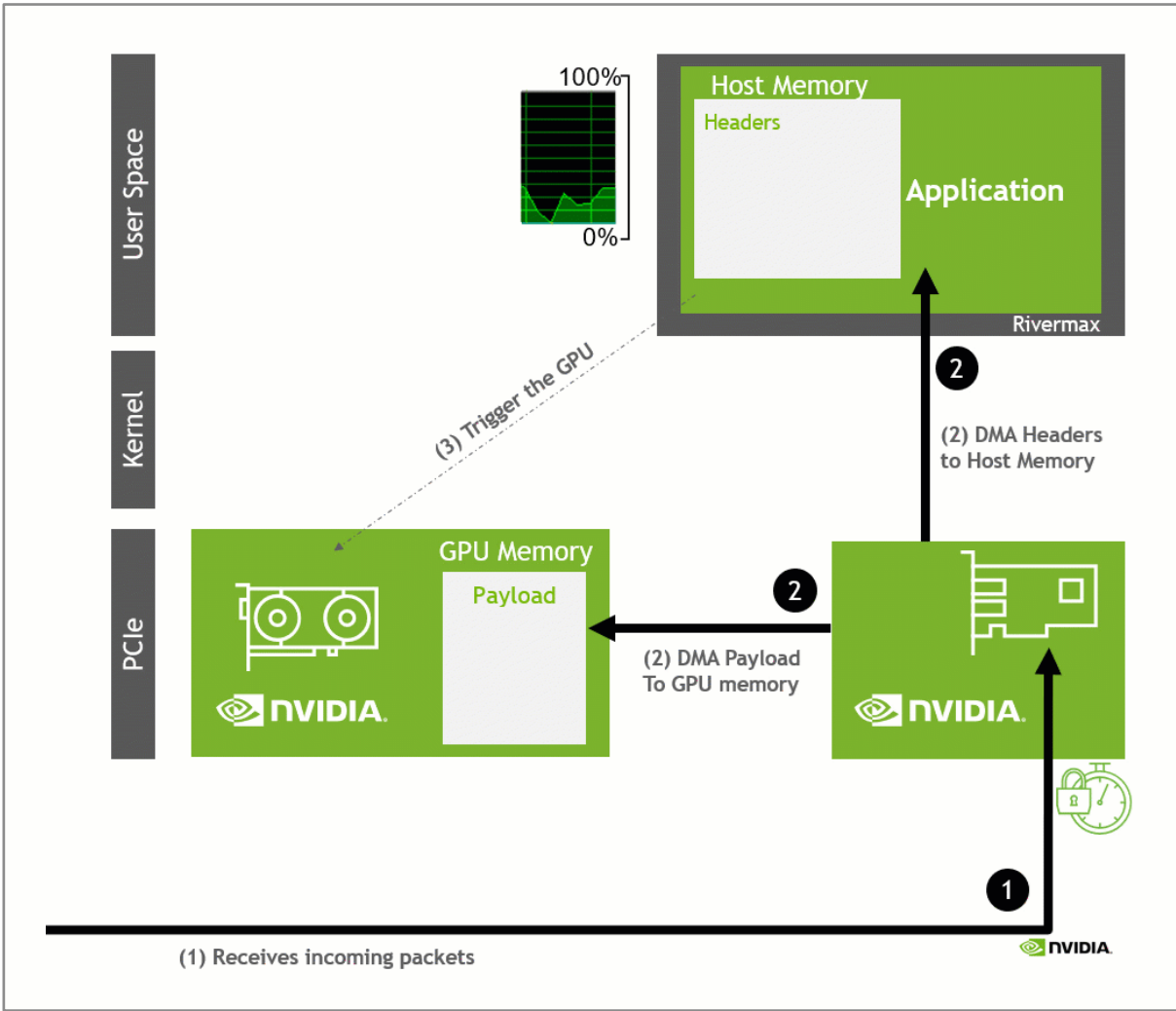
10X HIGHER PERFORMANCE WITH GPUDIRECT™ RDMA



No GPUDirect

Network Handled by CPU and CPU-Memory

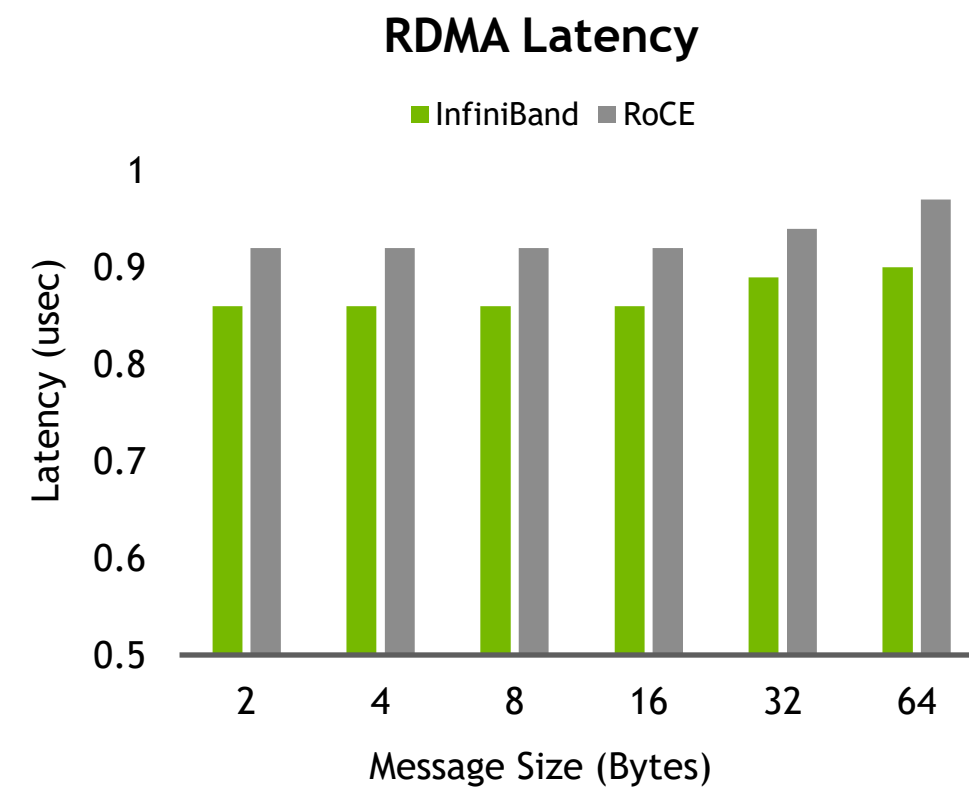
2	Full Copy Operations	0
2	PCIe Transactions	1
↓	GPU Utilization	↑
↑	CPU Usage	↓
↑	Latency	↓



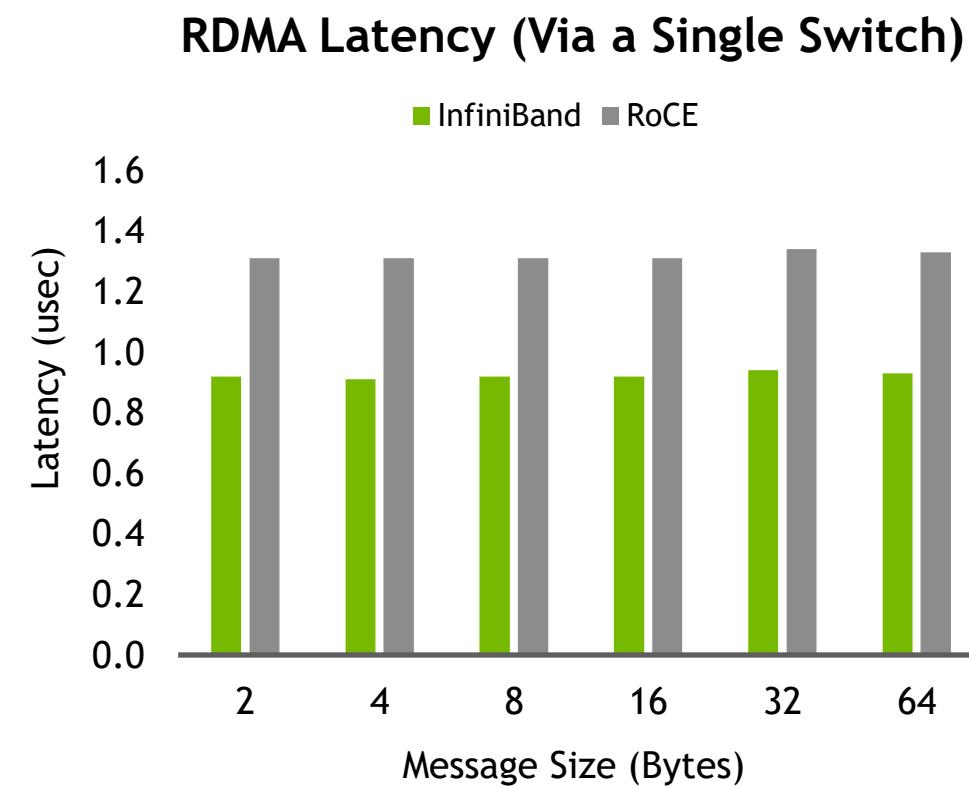
GPUDirect

Network Goes Directly to GPU Memory

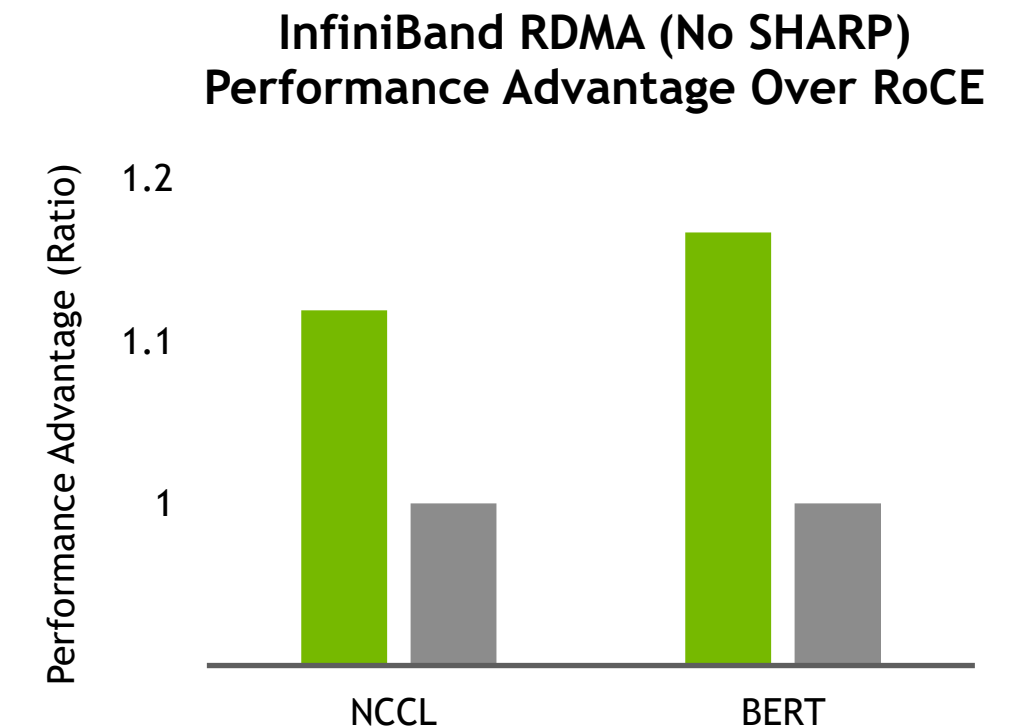
INFINIBAND RDMA ADVANTAGES



RDMA Latency Node to Node No Switch
InfiniBand is 7% Faster vs Ethernet RoCE



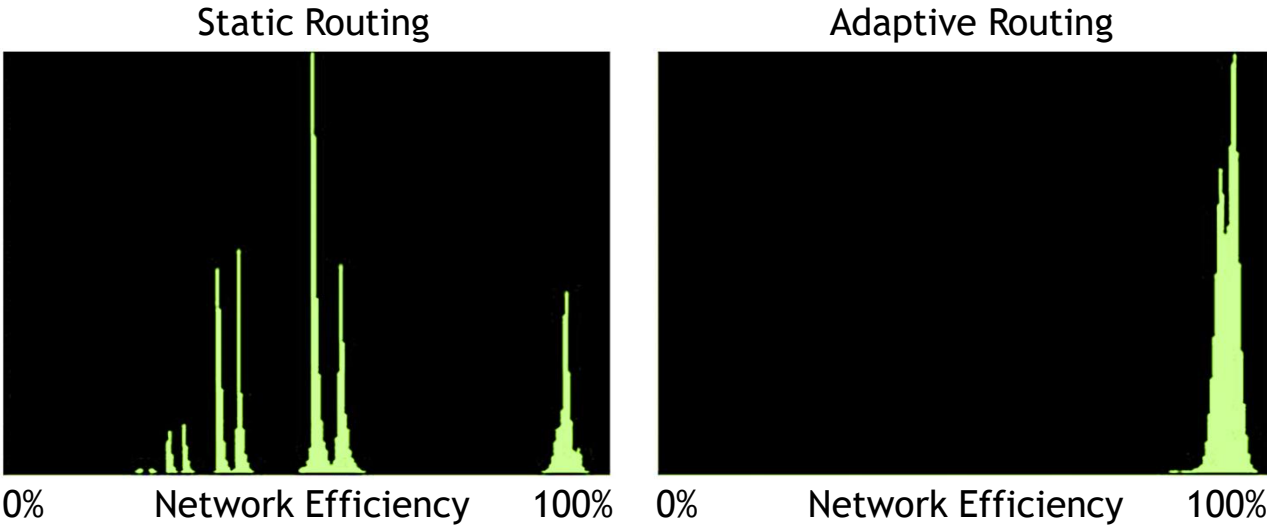
RDMA Latency Node to Node Single Switch
InfiniBand is 30% Faster vs Ethernet RoCE



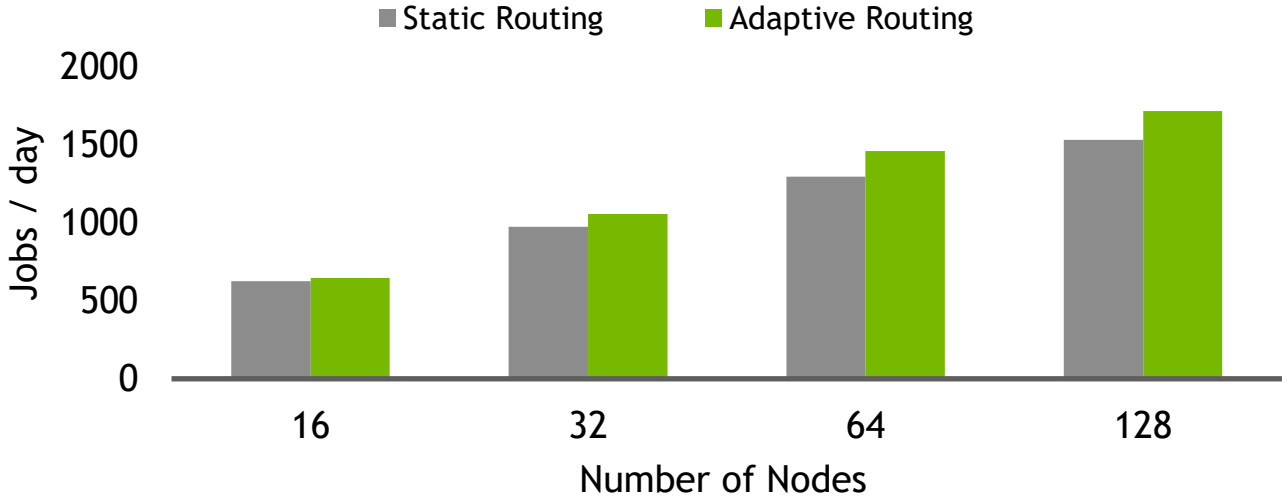
InfiniBand RDMA Enables 16% Higher
BERT AI Performance (4 Nodes)

INFINIBAND ADAPTIVE ROUTING

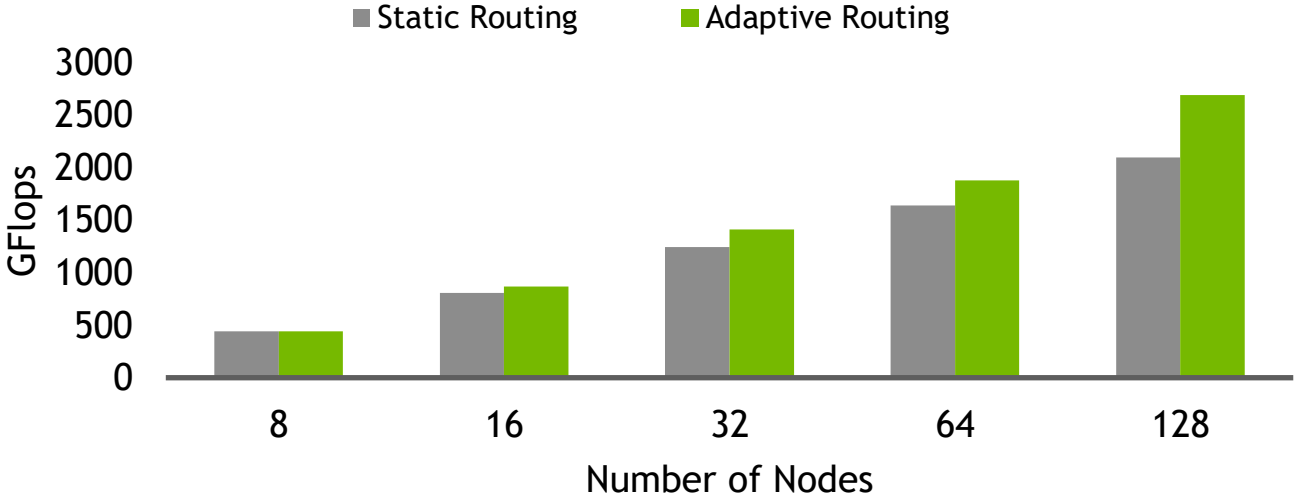
mpiGraph: Static Routing versus Adaptive Routing
(Summit)



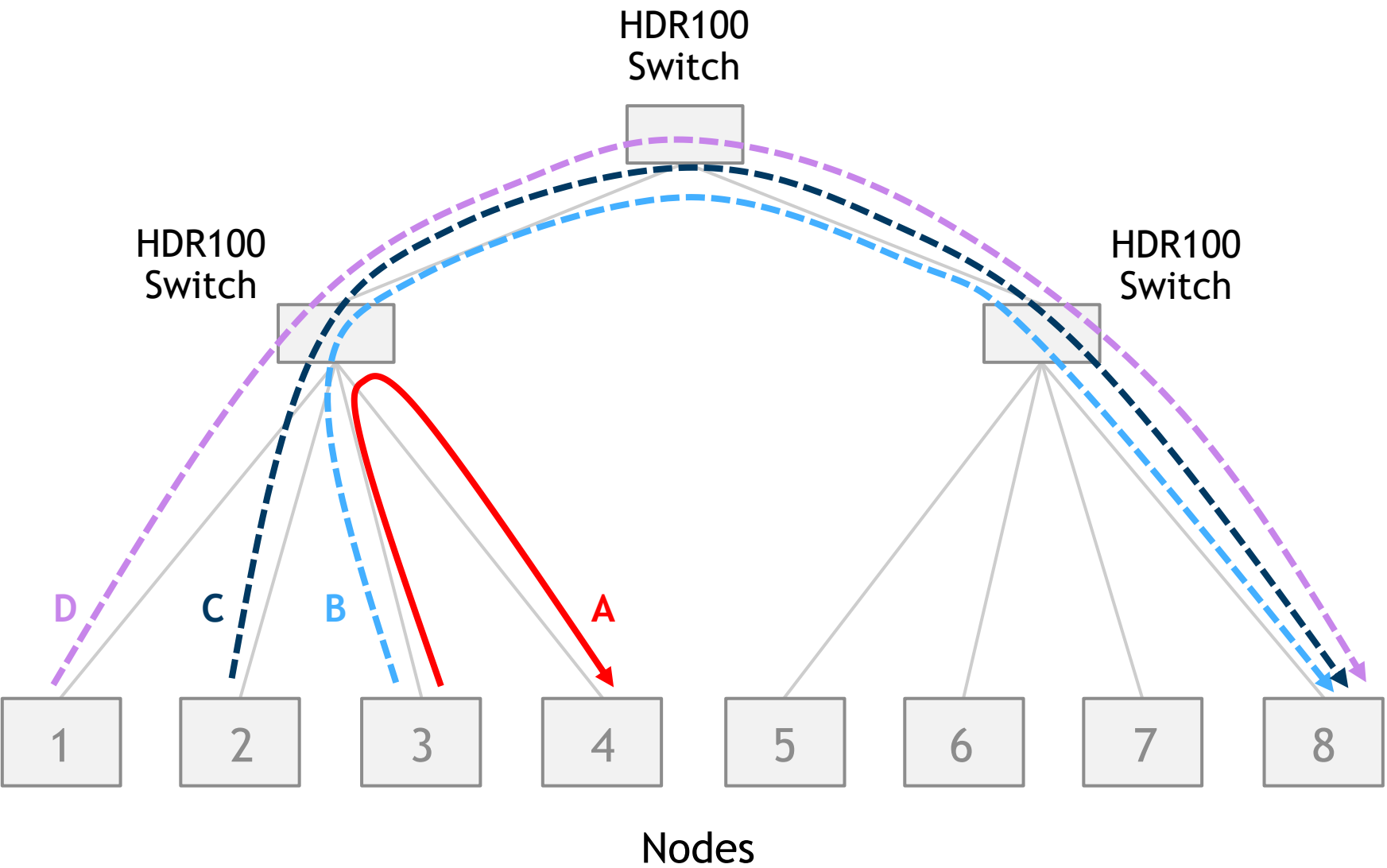
VASP
(GaAsBi512)



BSMBench
(Balance)



HDR INFINIBAND CONGESTION CONTROL



Without InfiniBand Congestion Control



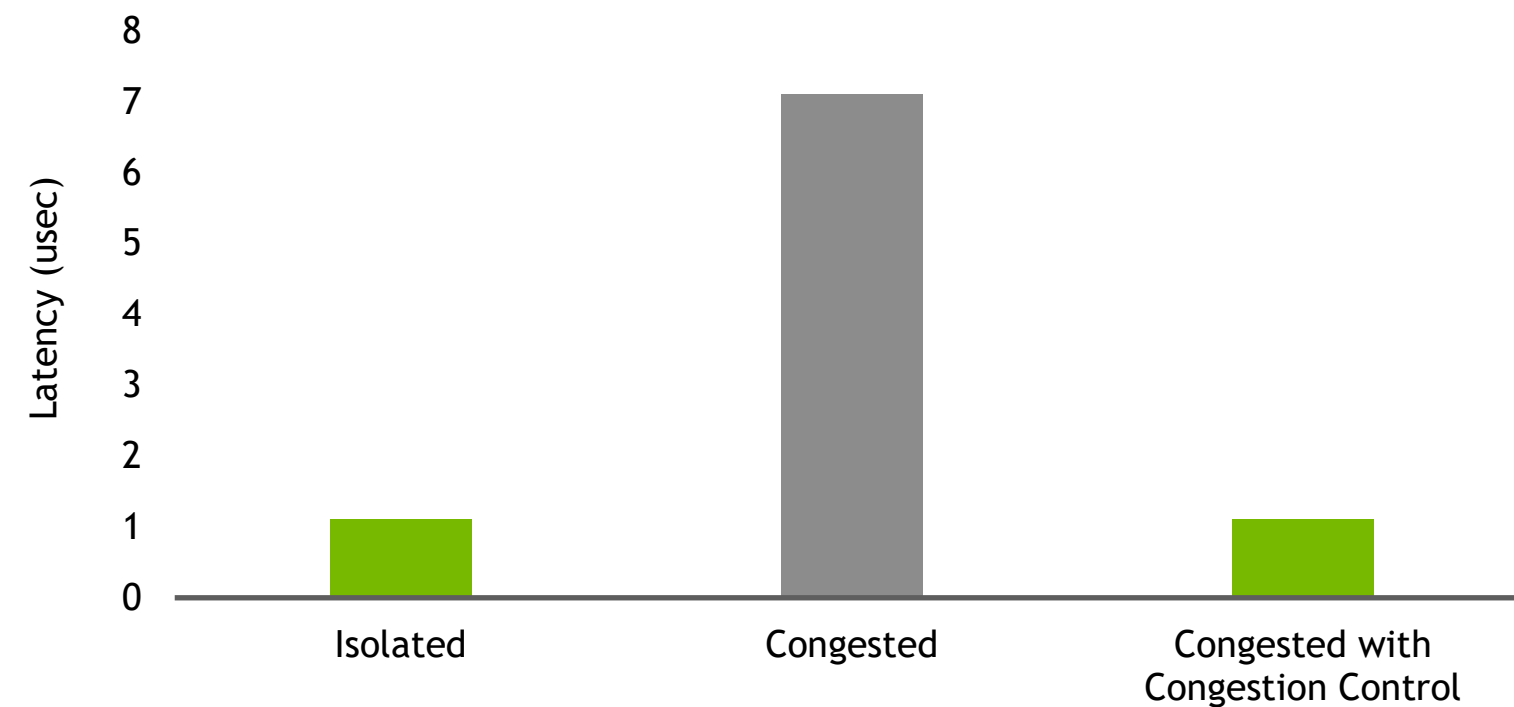
With InfiniBand Congestion Control



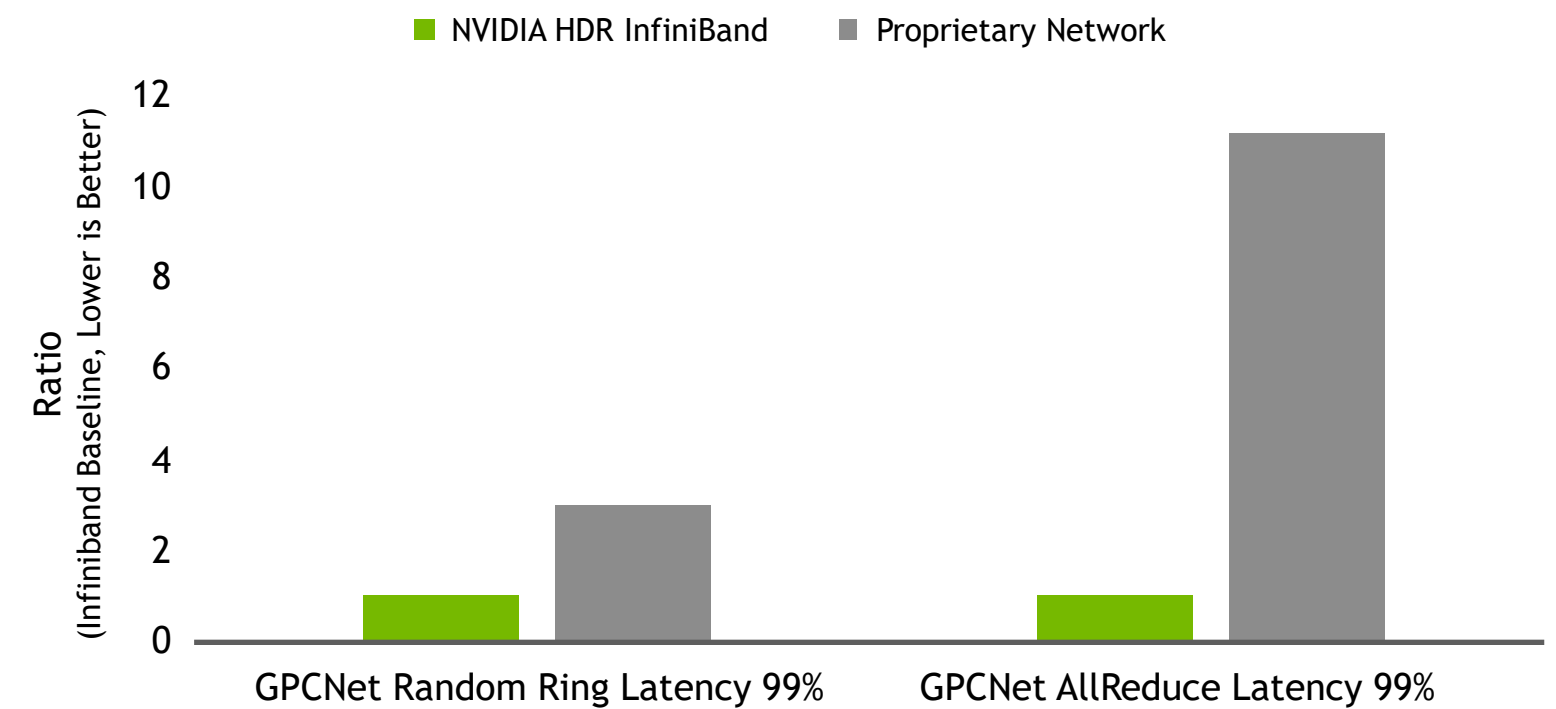
INFINIBAND CONGESTION CONTROL

Eliminates Congestion Hot Spots

Random Ring Latency



GPCNet Benchmark Performance Comparison



SCALABLE HIERARCHICAL AGGREGATION AND REDUCTION PROTOCOL (SHARP)

In-network Tree Based Aggregation Mechanism

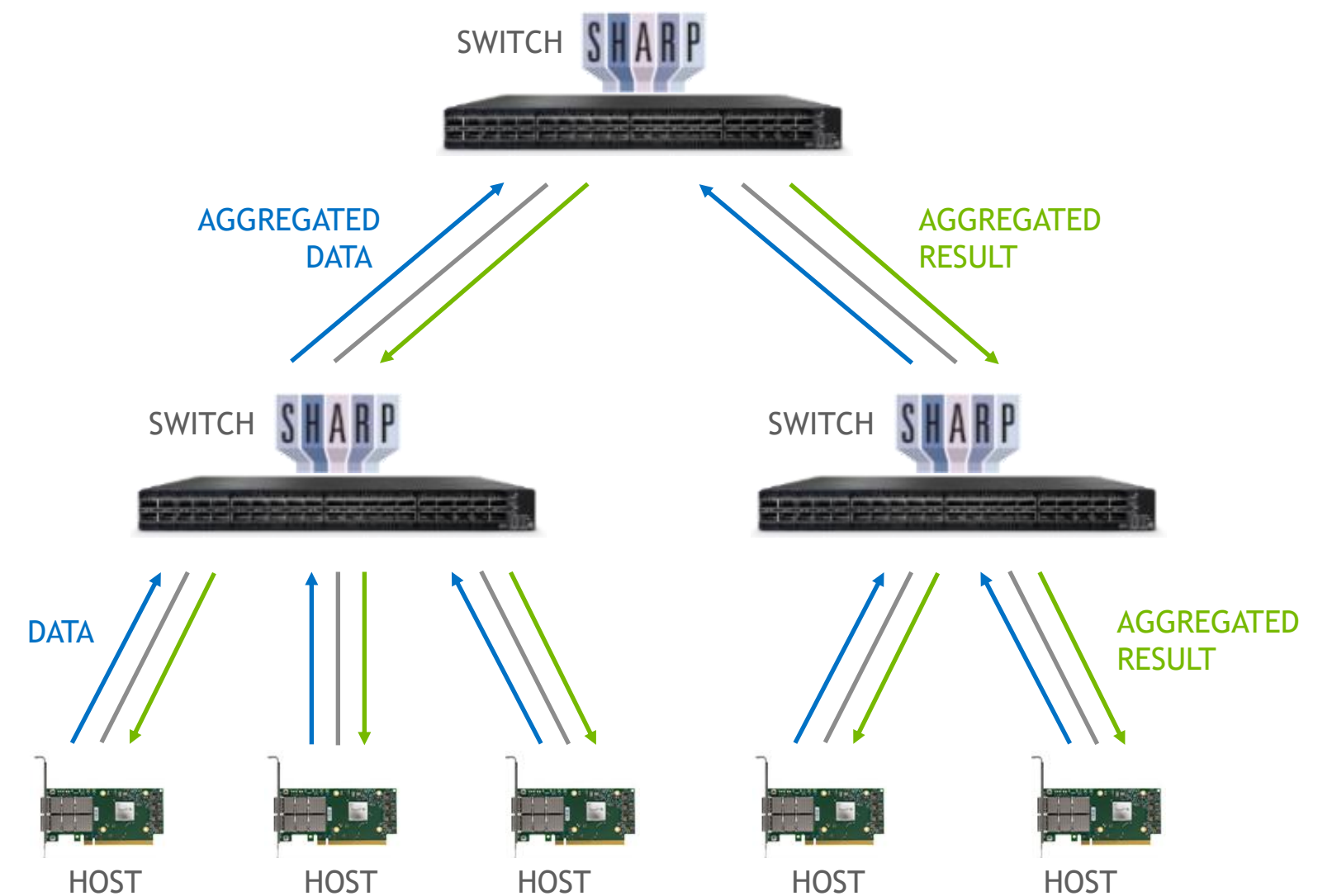
Multiple Simultaneous Outstanding Operations

Small Message and Large Message Reduction

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



SCALABLE HIERARCHICAL AGGREGATION AND REDUCTION PROTOCOL (SHARP)

In-network Tree Based Aggregation Mechanism

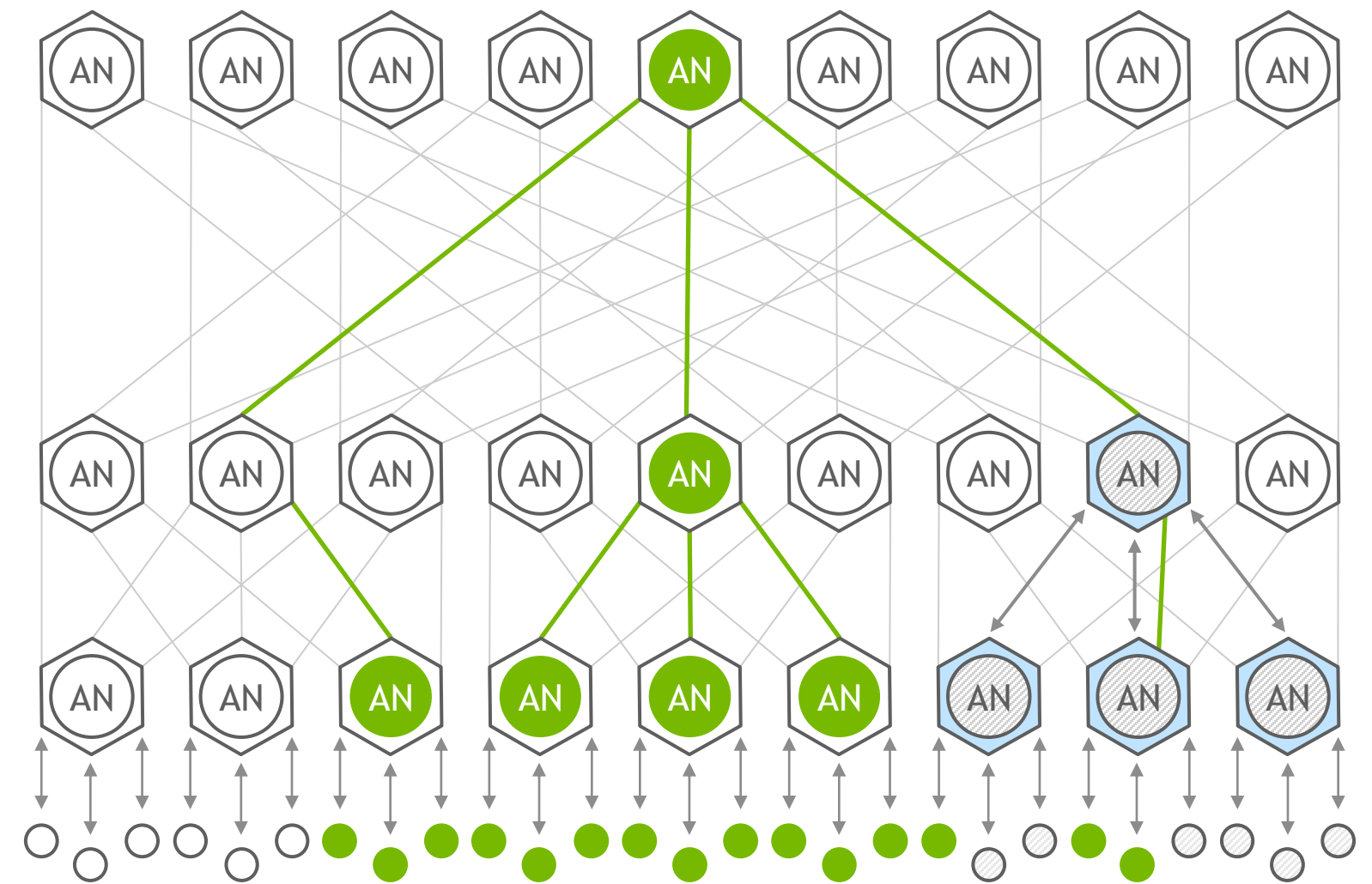
Multiple Simultaneous Outstanding Operations

Small Message and Large Message Reduction

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



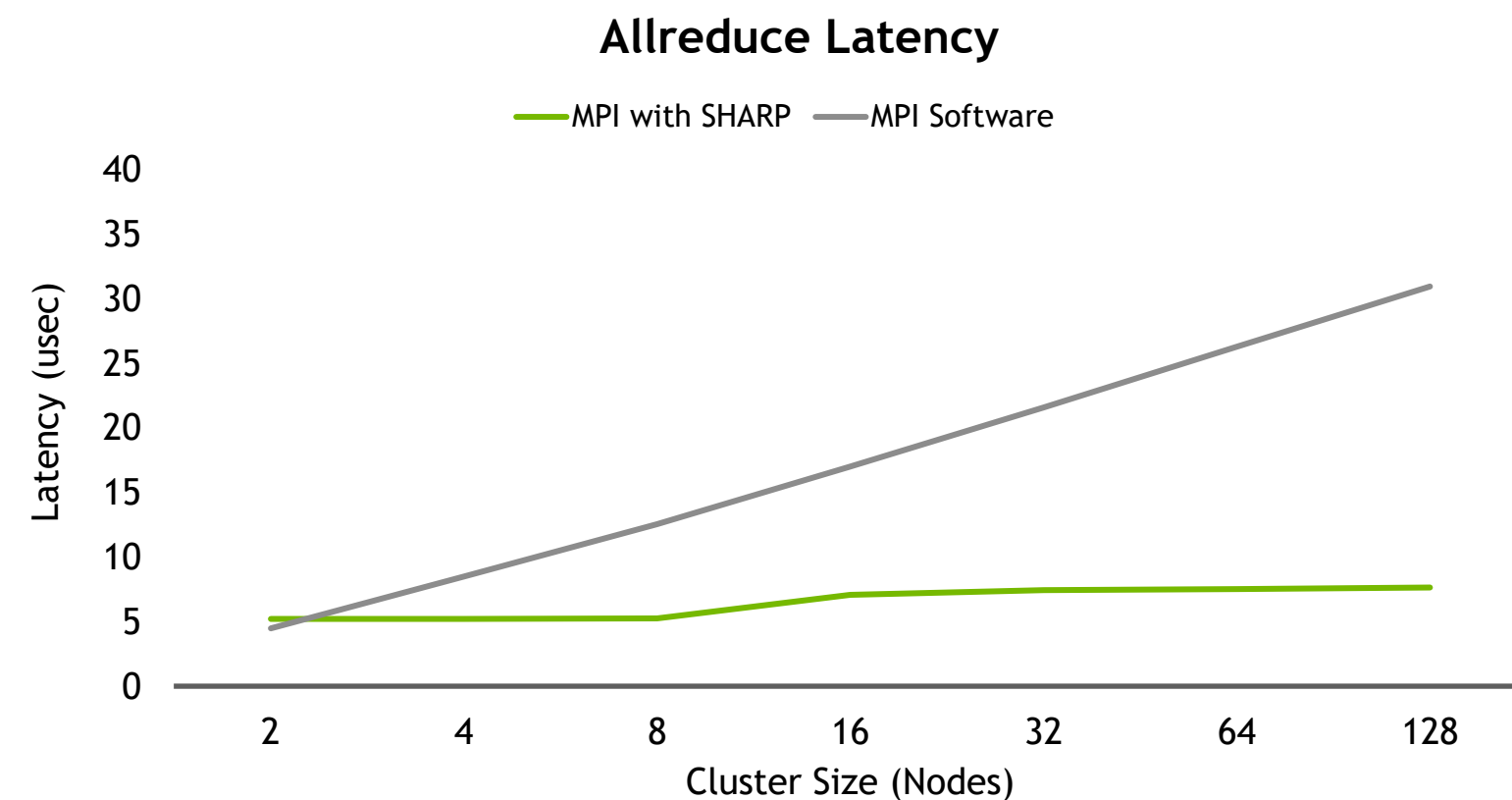
SHARP Aggregation Node: Switch Resident



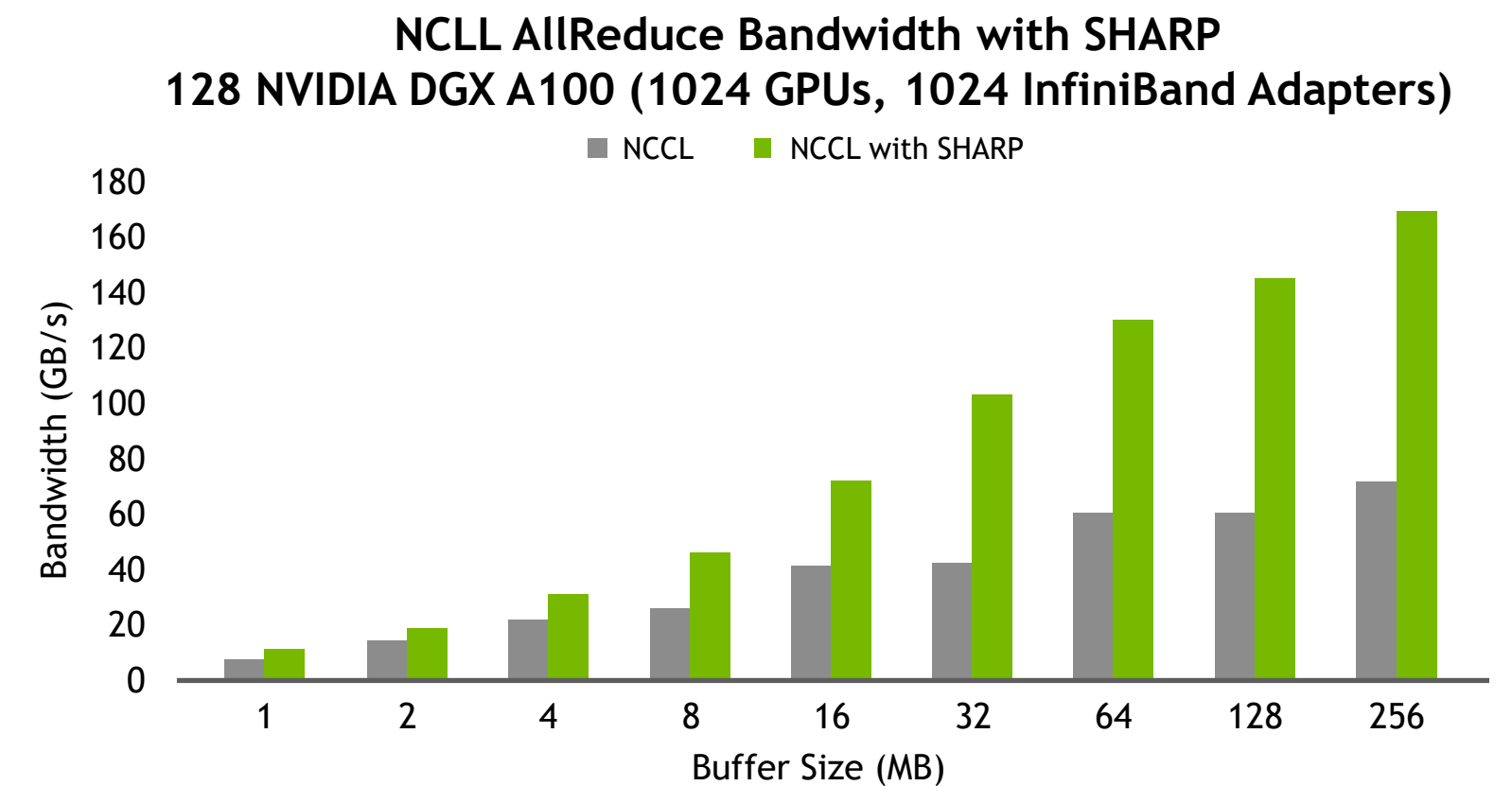
Host: Data source and Destination

SHARP ALLREDUCE PERFORMANCE ADVANTAGES

7x Higher MPI Performance, 2.5x Higher AI Performance



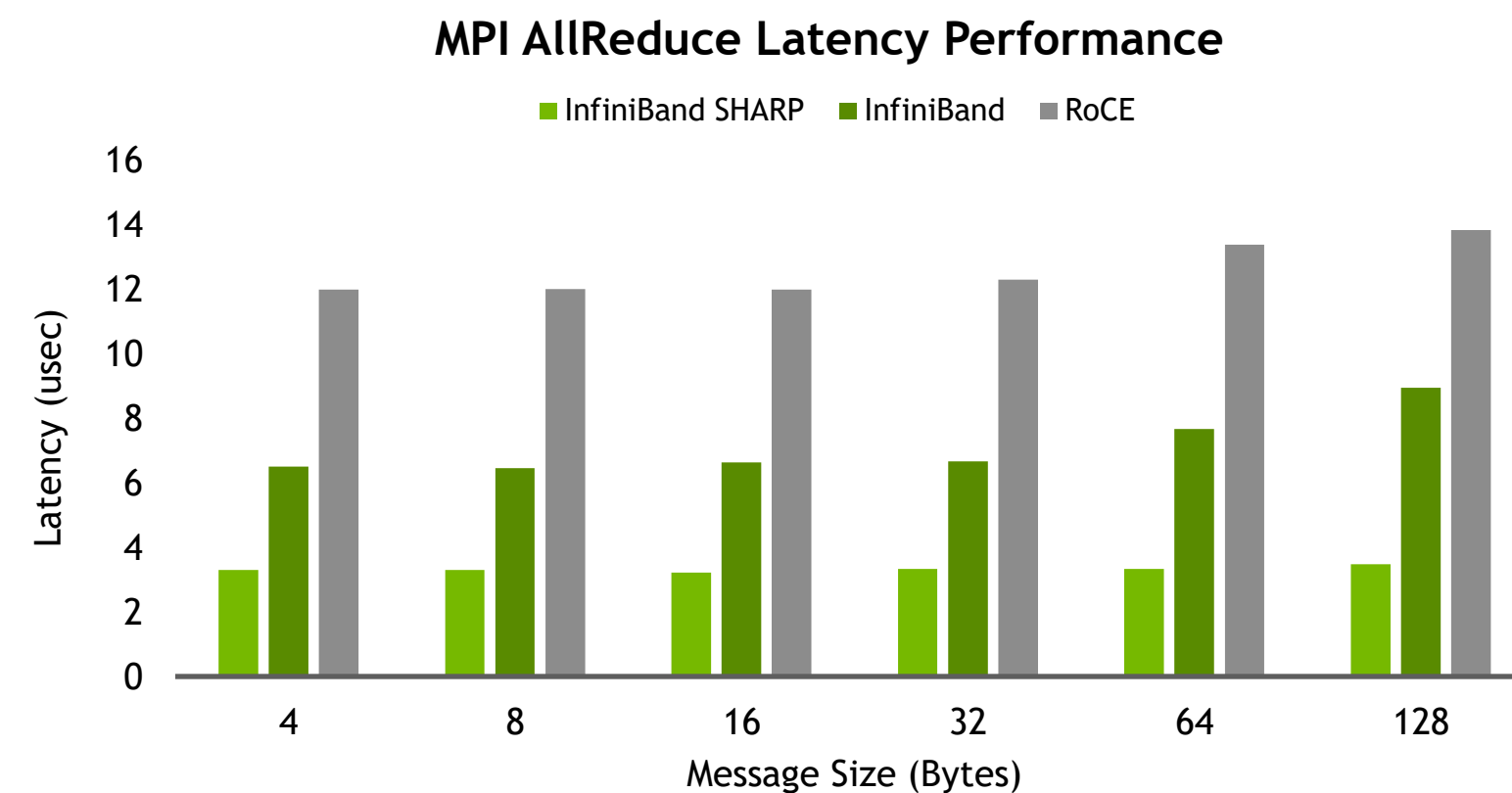
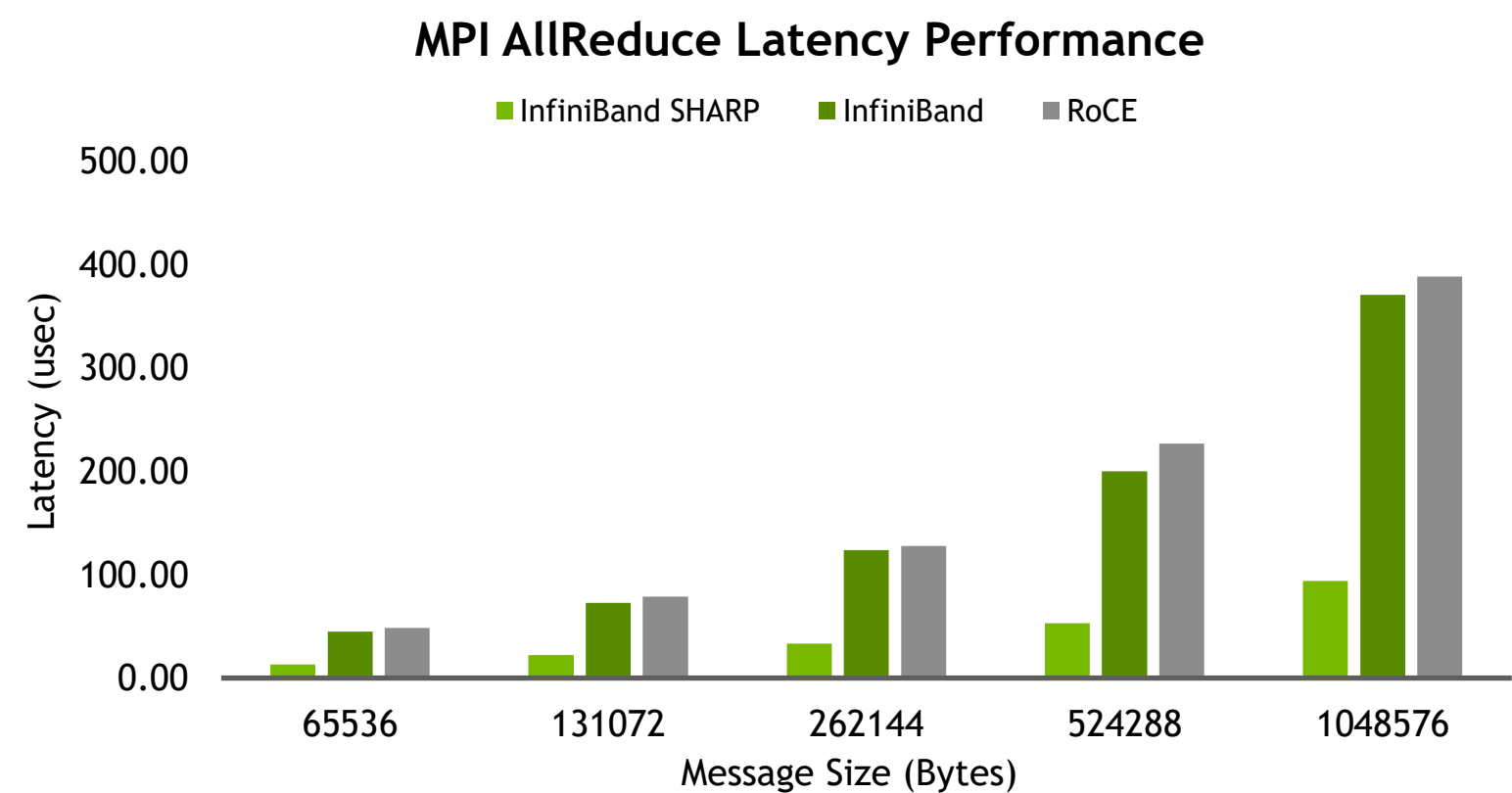
Lower is Better



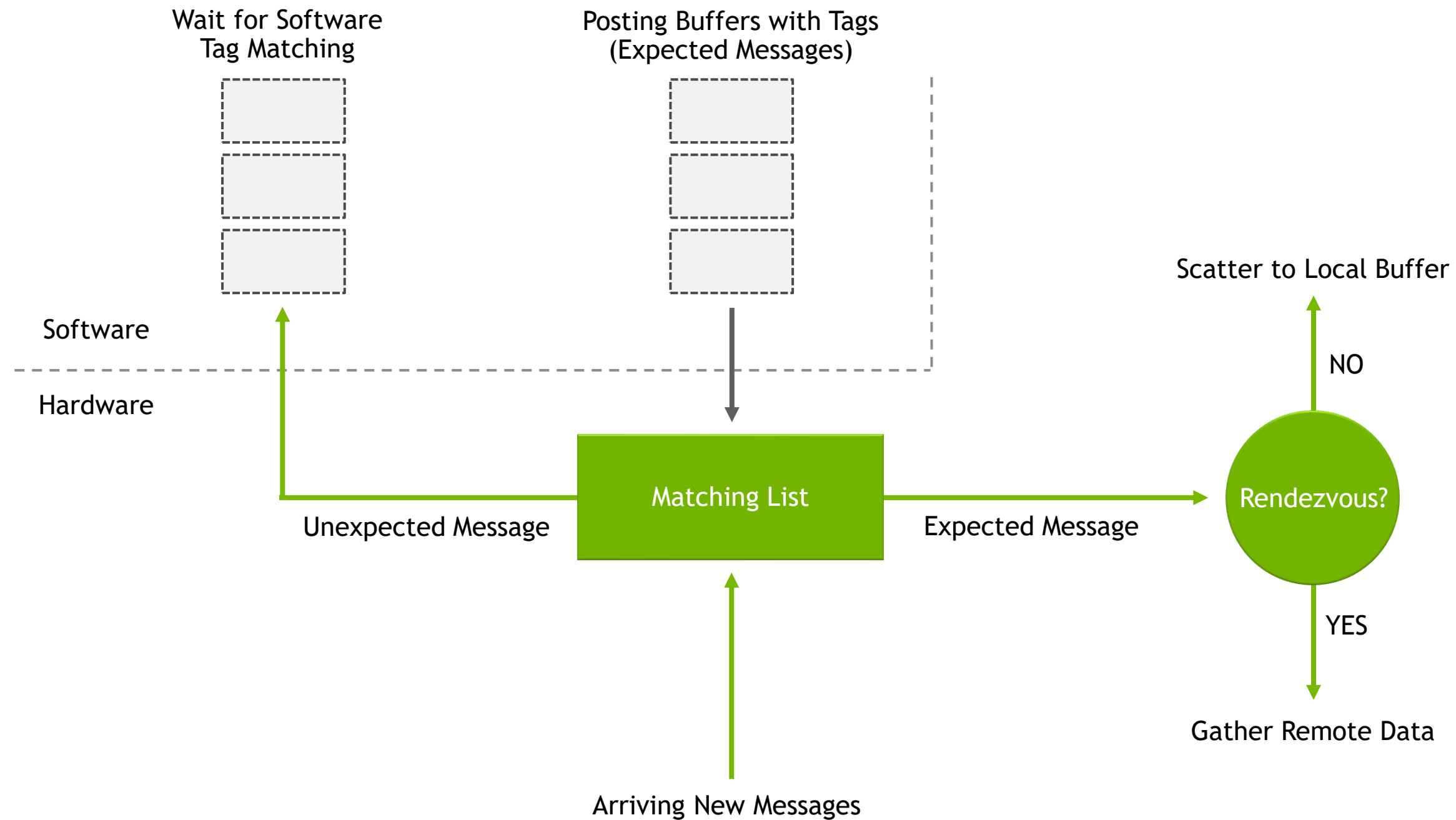
Higher is Better

INFINIBAND SHARP PERFORMANCE ADVANTAGE OVER ROCE

4X Higher Performance



INFINIBAND MPI TAG MATCHING HARDWARE ENGINE



IN-NETWORK COMPUTING ACCELERATED SUPERCOMPUTING

Software-Defined, Hardware-Accelerated, InfiniBand Network

In-Network Computing Acceleration Engines



New generations Introduce and Enhance Acceleration Technologies

NDR InfiniBand Includes SHARP v3 and All-to-all Engines

	Faster Data Communications	Higher Application Performance
Small Data Reduction SHARP v2	7x Faster All-Reduce	~15% higher Performance OpenFOAM, DL-POLY-4, Relion
Large Data Reduction SHARP v2	2.5x Faster All-Reduce	15% Faster Deep Learning Recommend. 17% Faster Natural Language Processing
MPI Tag Matching	1.8x Faster MPI Iscatterv 100% Overlapping	Up to 40% Higher Performance LAMMPS, Nekbone, 3D Stencil
All-to-All (Introduced with NDR 400G)	4x Higher Throughput	Coming Soon with NVIDIA NDR InfiniBand!

CLOUD-NATIVE SUPERCOMPUTING

Bare-metal Secured Infrastructure

Higher Application Performance

From the Edge to the Main Data Center



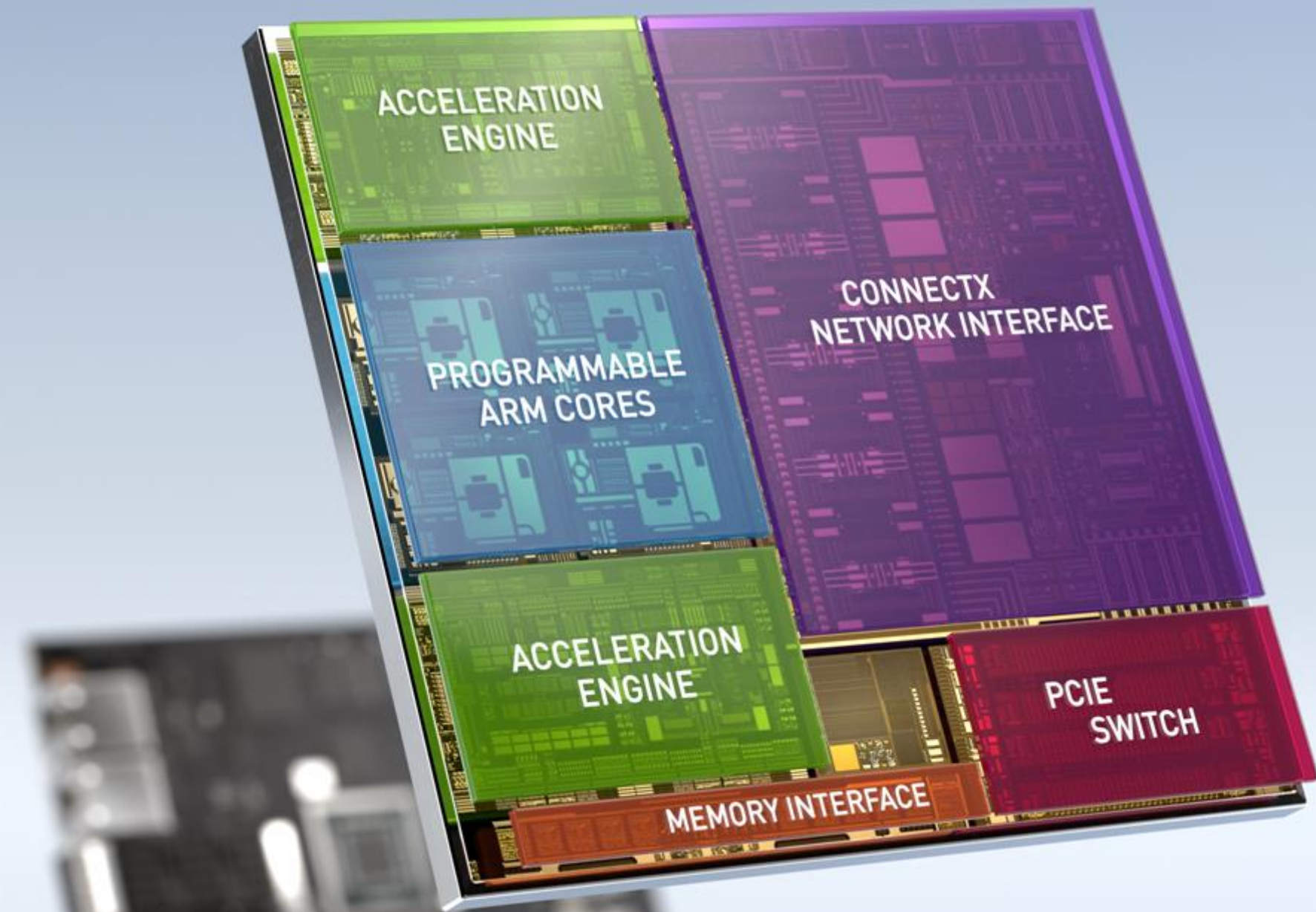
BARE-METAL
PERFORMANCE



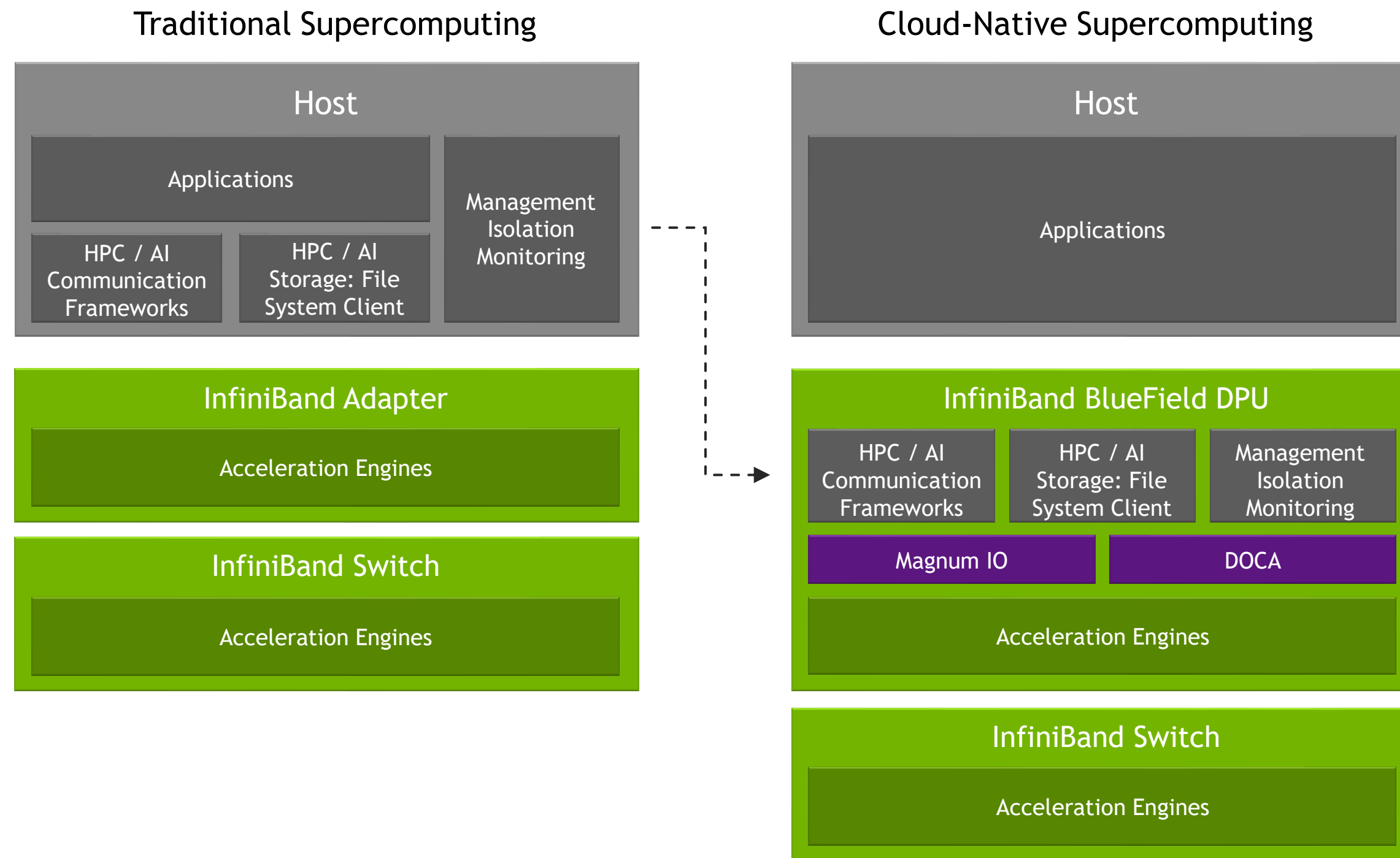
MULTI
TENANCY



EDGE
COMPUTING



CLOUD-NATIVE SUPERCOMPUTING INFRASTRUCTURE



MULTI-TENANT ISOLATION

Zero-Trust Architecture

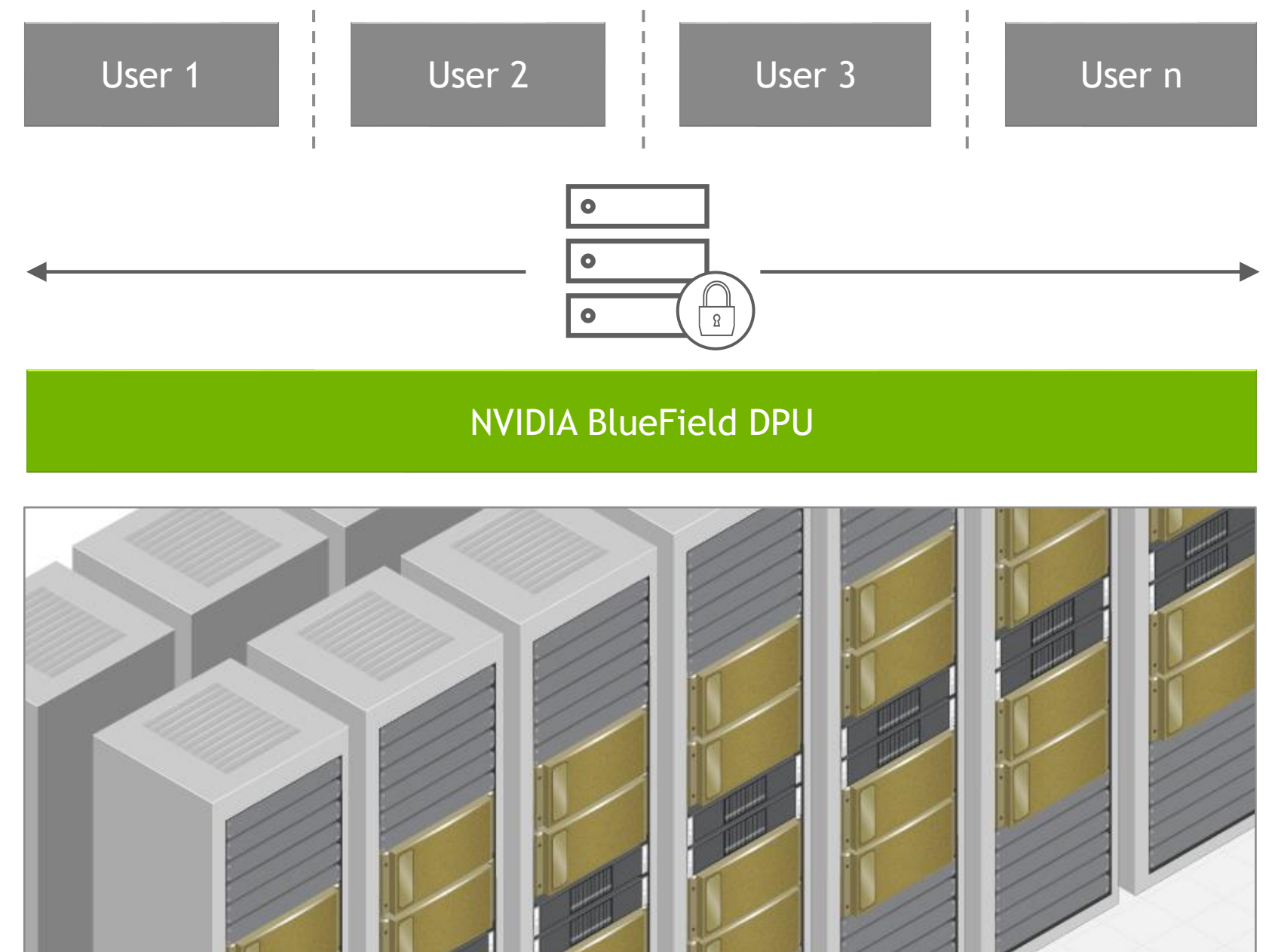
Secured Network Infrastructure and Configuration

Storage Virtualization

Tenant Service Level Agreement (SLA)

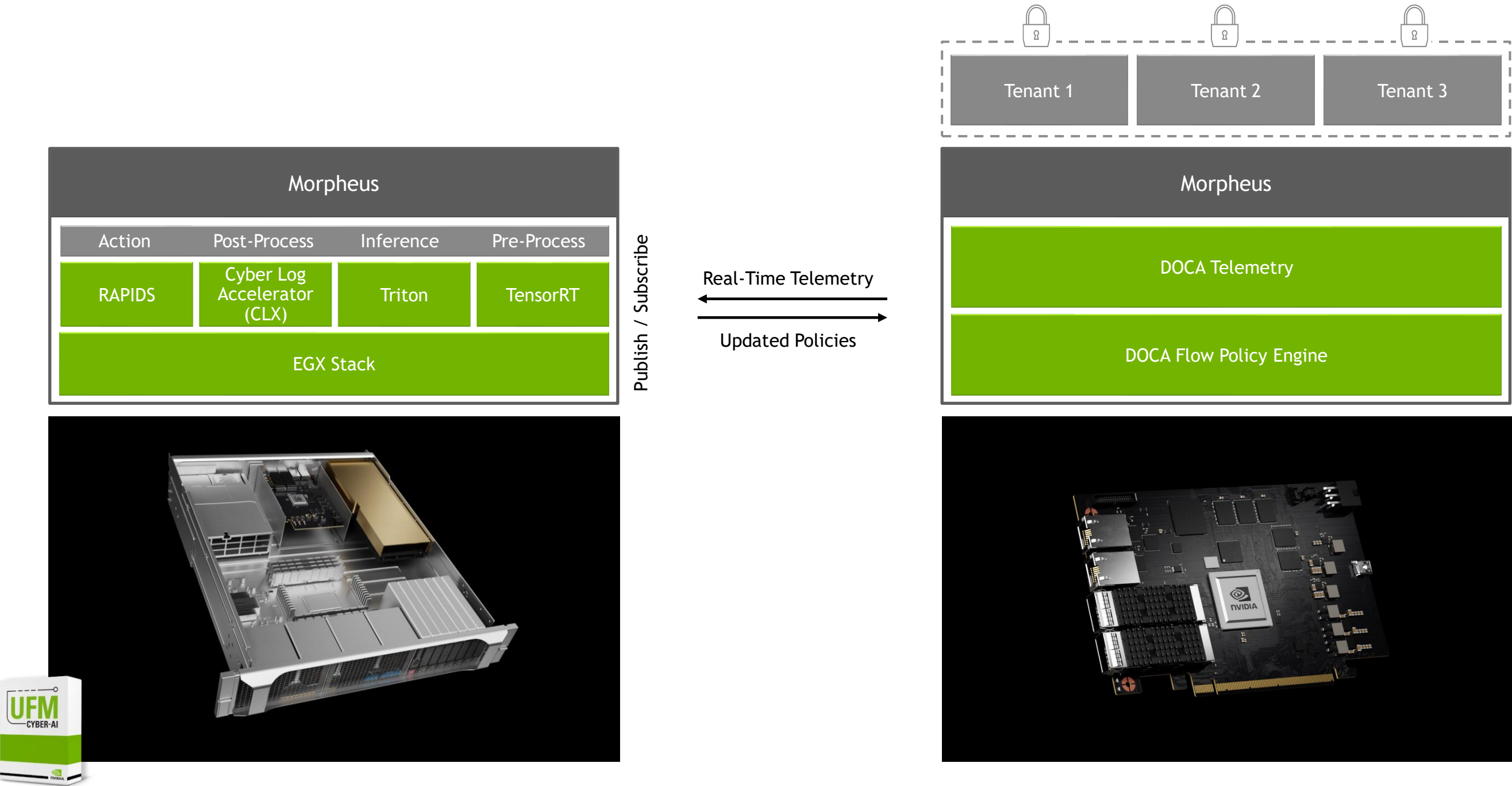
32K Concurrent Isolated Users on Single Subnet

Secure Partitioning with Bare-Metal Performance



MORPHEUS AI CYBERSECURITY FRAMEWORK

UFM Cyber-AI with Morpheus for Higher Security



HIGHER APPLICATION PERFORMANCE

DPU-Accelerated HPC Communications

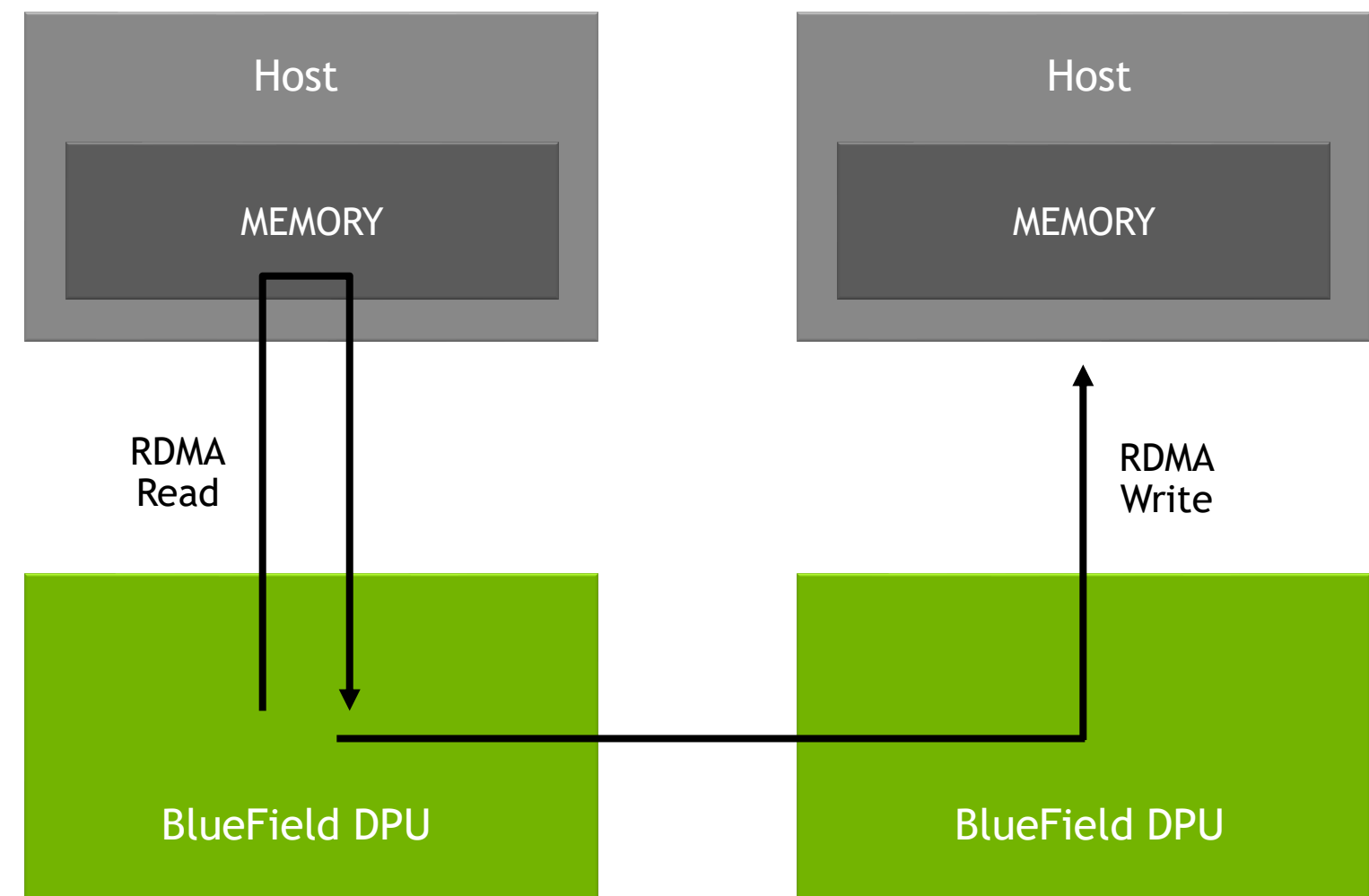
Collective Offloads

Active Messages

Smart MPI Progression

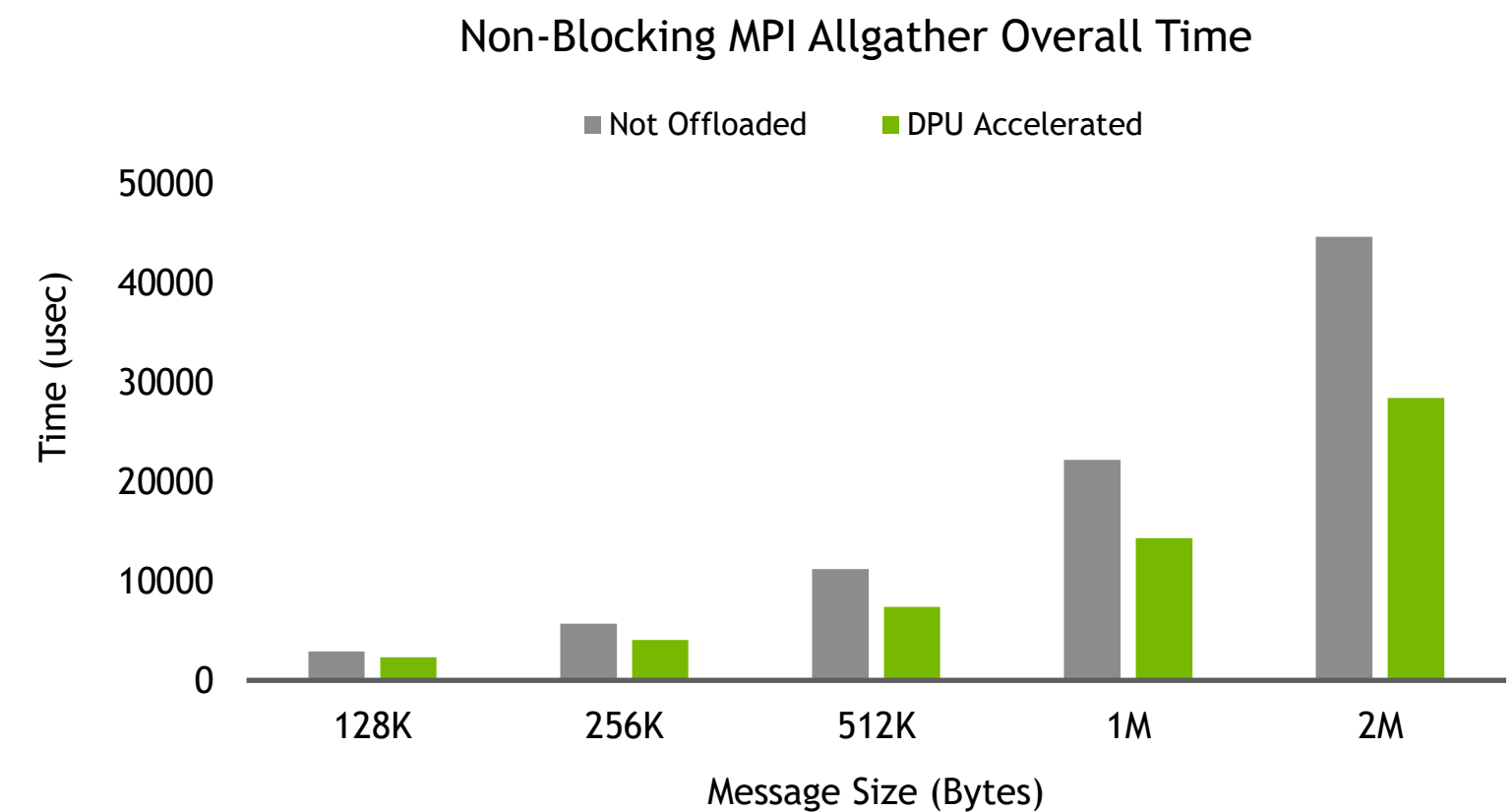
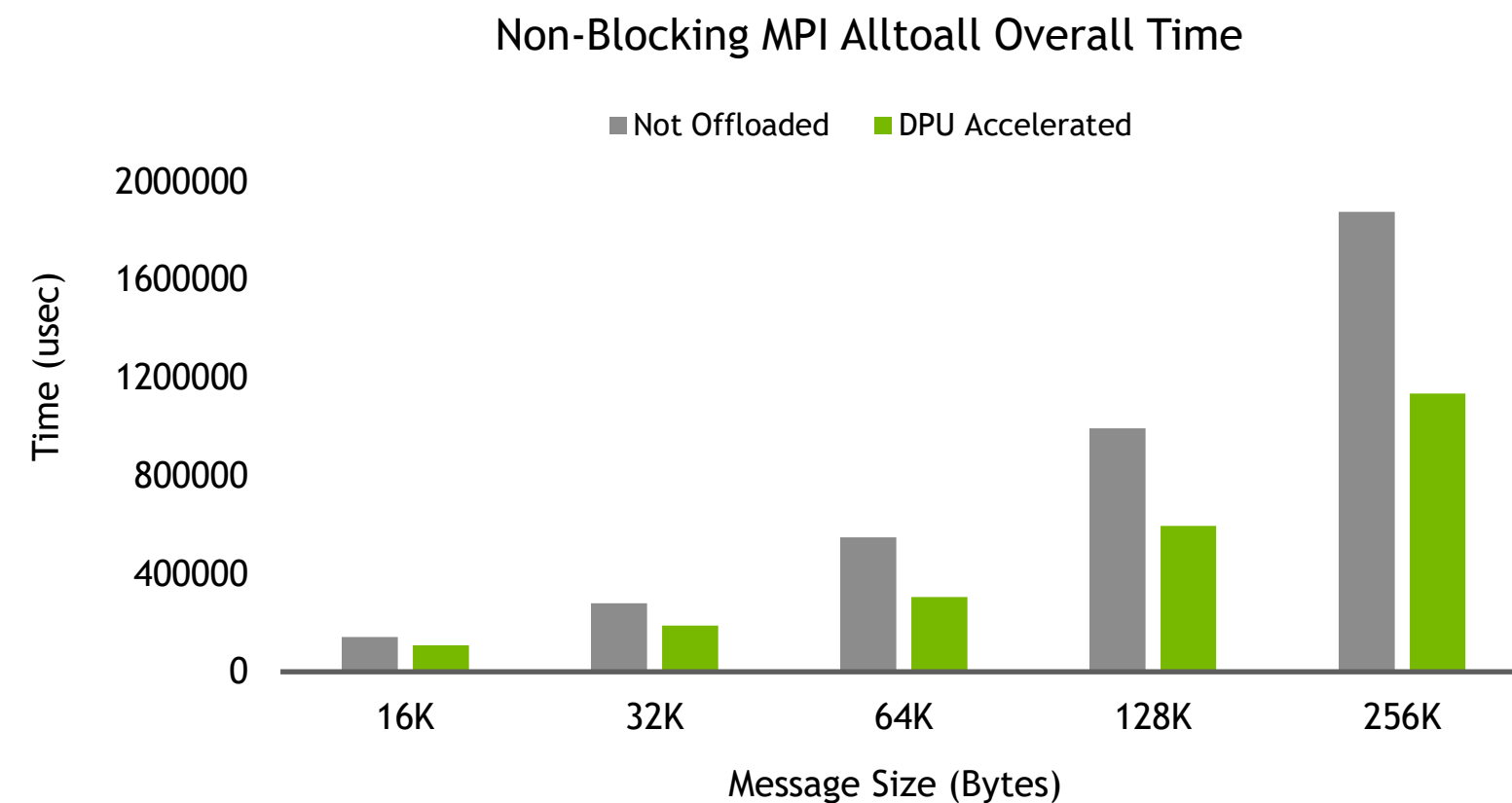
Data Compression

User-defined Algorithms



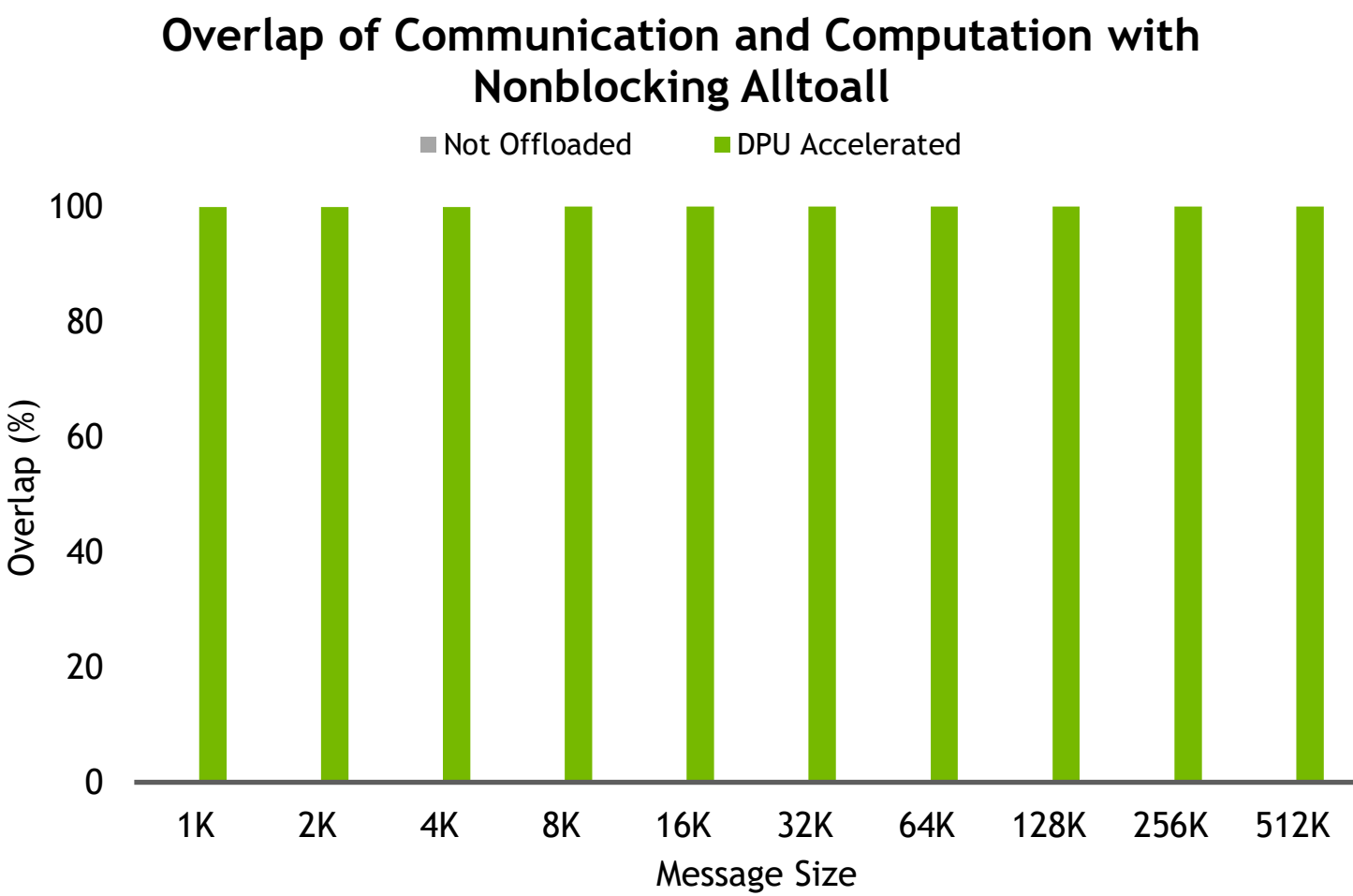
NON-BLOCKING MPI PERFORMANCE

44% Performance Increase for MPI iAlltoall, 36% Performance Increase for MPI iAllgather



HIGHER APPLICATION PERFORMANCE

100% Communication - Computation Overlap



Courtesy of: Ohio State University MVAPICH Team and X-ScaleSolutions



MVAPICH



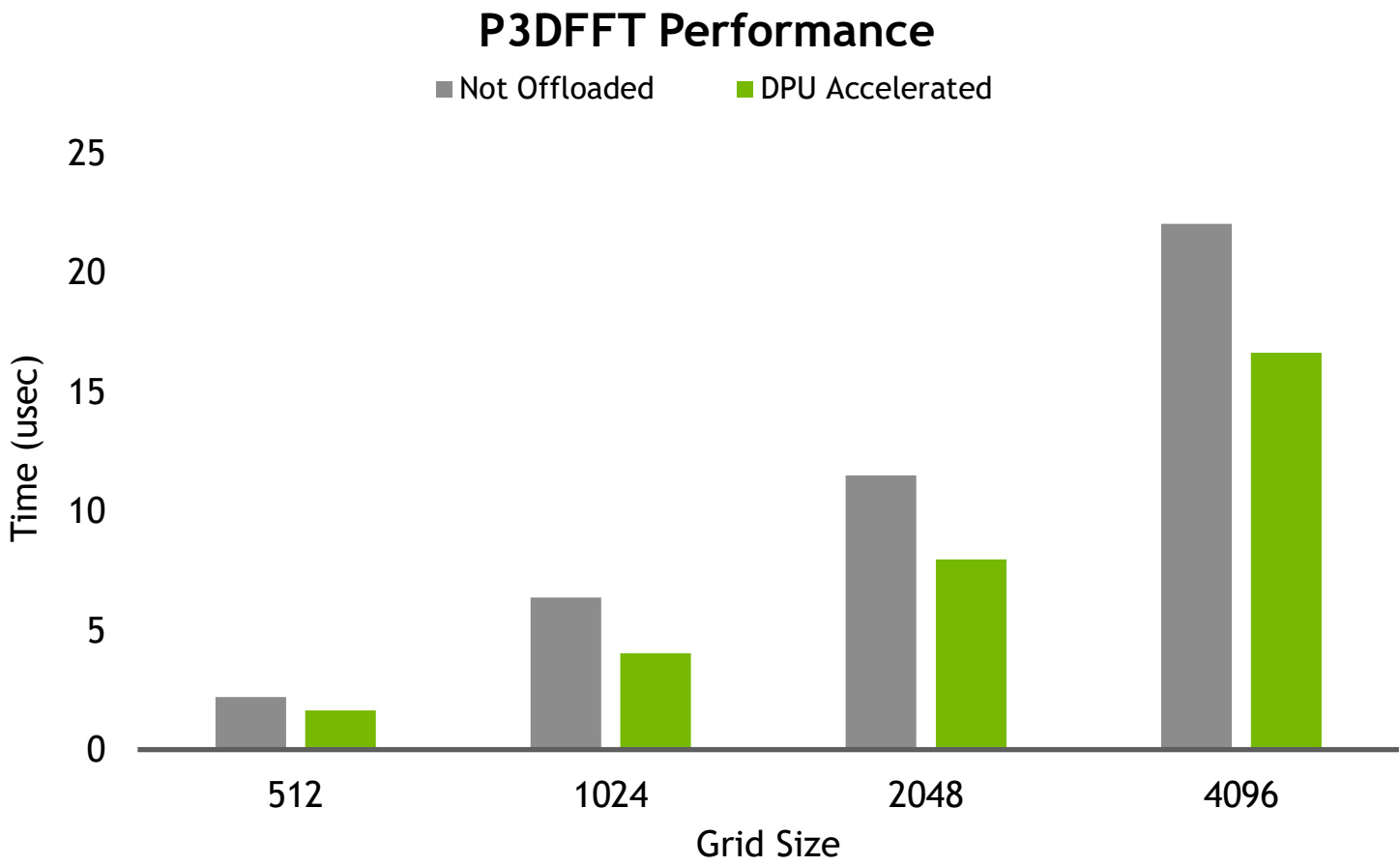
X-ScaleSolutions

32 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 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.



HIGHER APPLICATION PERFORMANCE

Higher App Performance,
MPI Collectives Offload



Courtesy of: Ohio State University MVAPICH Team and X-ScaleSolutions



MVAPICH



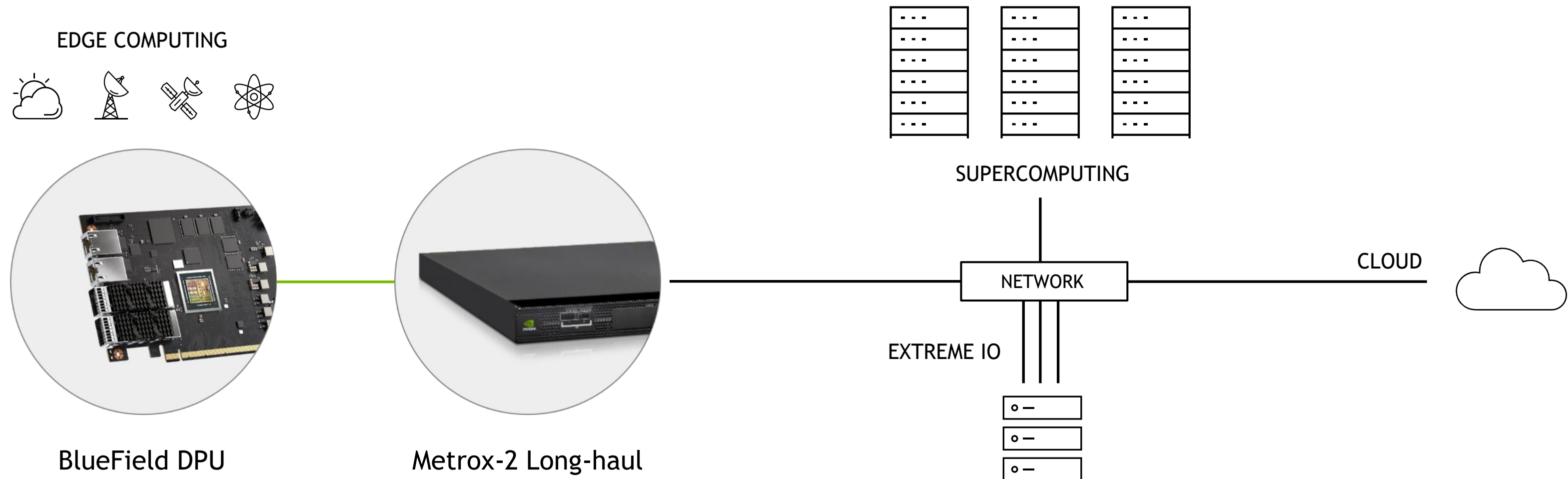
X-ScaleSolutions

32 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 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.



FROM THE EDGE TO THE DATA CENTER

Pre-Processing at the Edge, Native InfiniBand RDMA to the Main Supercomputer



NVIDIA DOCA

Enabling Broad DPU Partner Ecosystem

Software Application Framework for BlueField DPUs

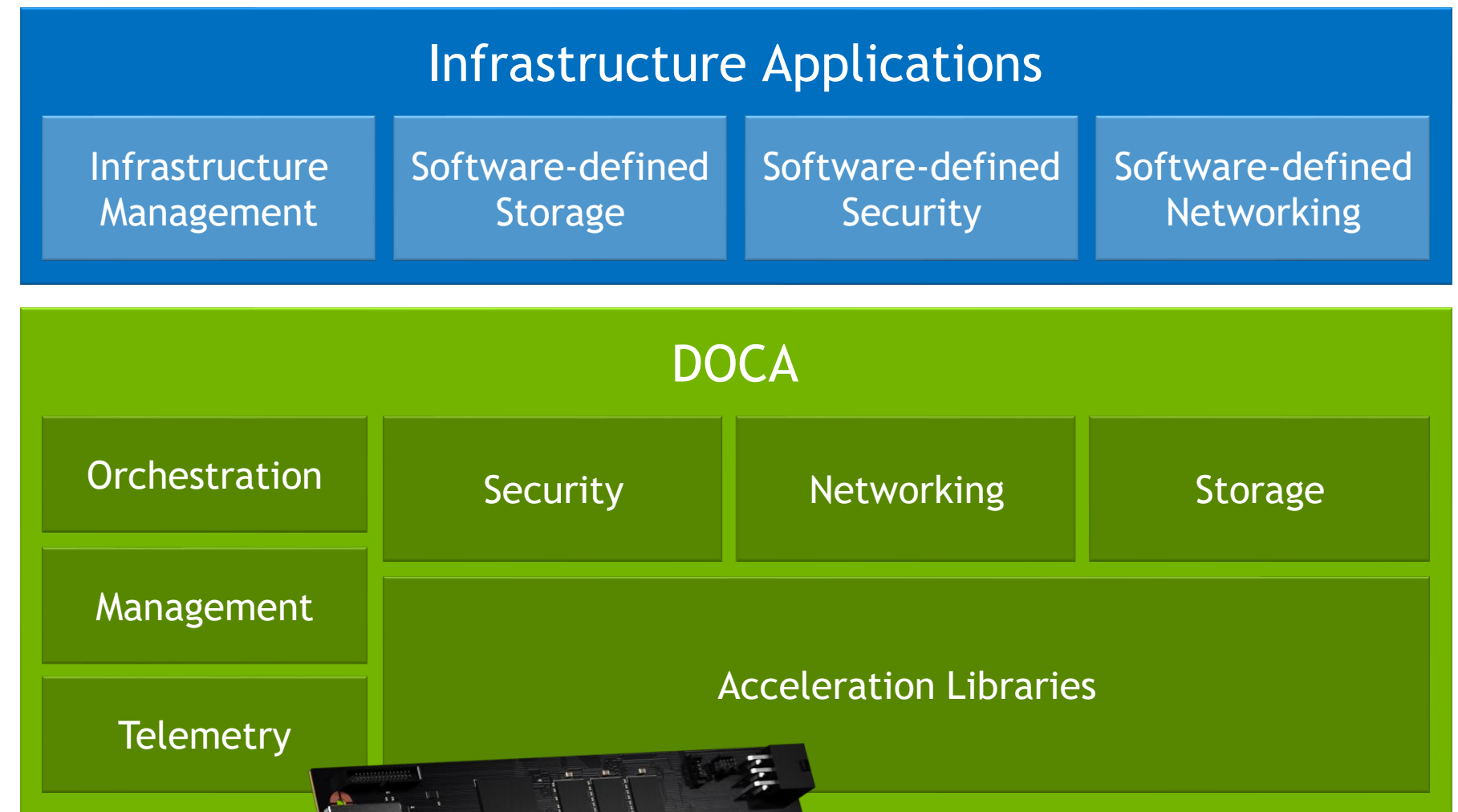
DOCA is for DPUs What CUDA is for GPUs

Protects Developer Investment for Future DPUs

Certified Reference Applications, APIs & Partner Solutions

Rich Partner Ecosystem Across Industries and Workloads

<https://developer.nvidia.com/networking/doca>

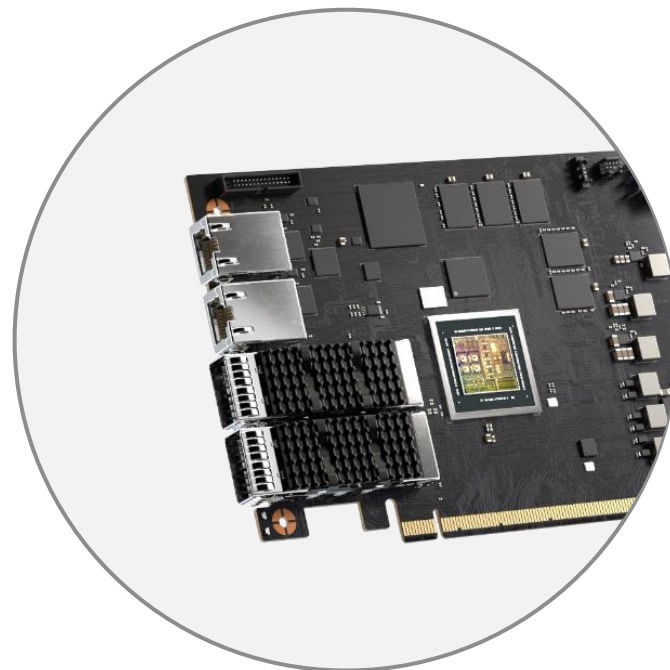


MELLANOX NDR 400G INFINIBAND: NEXT-GENERATION INFINIBAND ARCHITECTURE



ConnectX-7 Adapter

NDR 400G InfiniBand
PCIe Gen5
Programmable Datapath
In-Network Computing



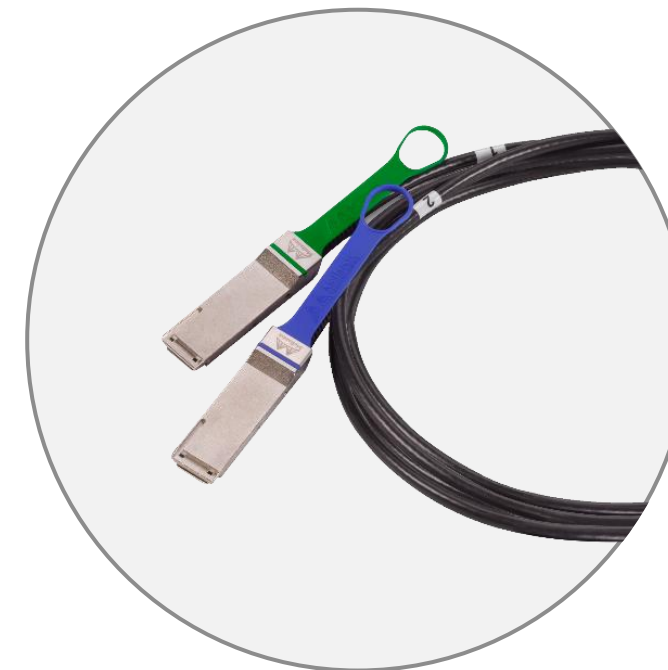
BlueField-3 DPU

NDR 400G InfiniBand with Arm Cores
PCIe Gen5, DDR5
AI Application Accelerators
Programmable Datapath
In-Network Computing



Quantum-2 Switch

64-ports NDR 400G InfiniBand
128-ports 200G NDR200
In-Network Computing



Cable

Copper Cables
Active Copper Cables
Optical Transceivers

ANNOUNCING NVIDIA NDR 400G INFINIBAND SYSTEMS

In-Network Computing Accelerated
Network for Cloud-Native
Supercomputing at Any Scale

2x

Data Throughput
400 Gigabits per Second

32x

More AI Acceleration
SHARP In-Network Computing

6.5x

Higher Scalability
>1M nodes with DF+ 3 hops

5x

Switch System Capacity
>1.6 Petabit per Second



ANNOUNCING NVIDIA NDR 400G INFINIBAND SYSTEMS

In-Network Computing Accelerated
Network for Cloud-Native
Supercomputing at Any Scale

15%

Faster Deep Learning
Recommendations

17%

Faster Natural
Language Processing

15%

Faster Computational
Fluid Dynamics Simulations

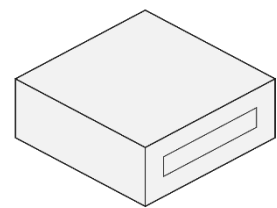
36%

Lower
Power Consumption

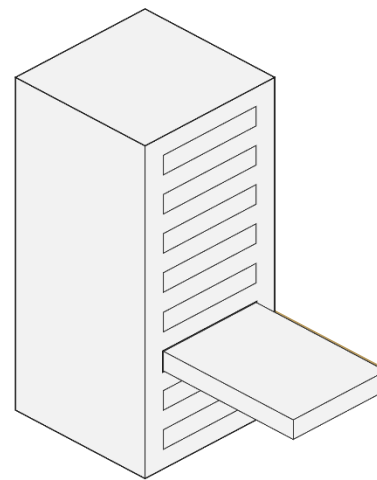


ANNOUNCING NVIDIA NDR 400G INFINIBAND SYSTEMS

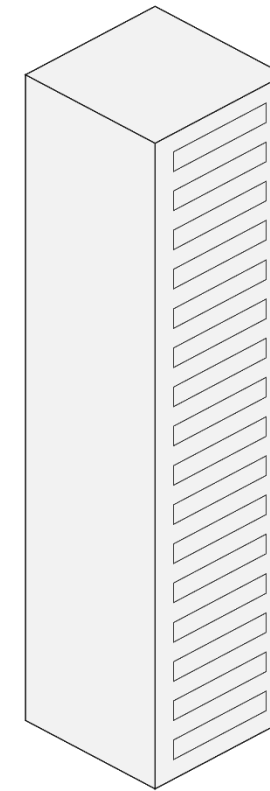
In-Network Computing Accelerates Cloud-Native Supercomputing at Any Scale



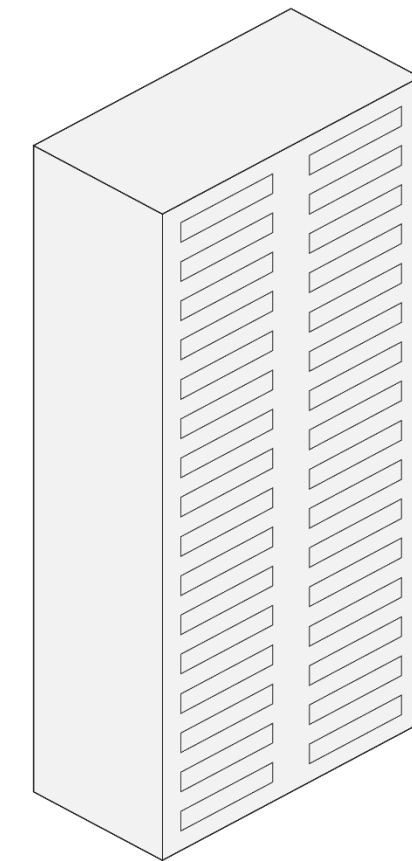
64 NDR Ports
128 NDR200 Ports



512 NDR Ports
1024 NDR200 Ports

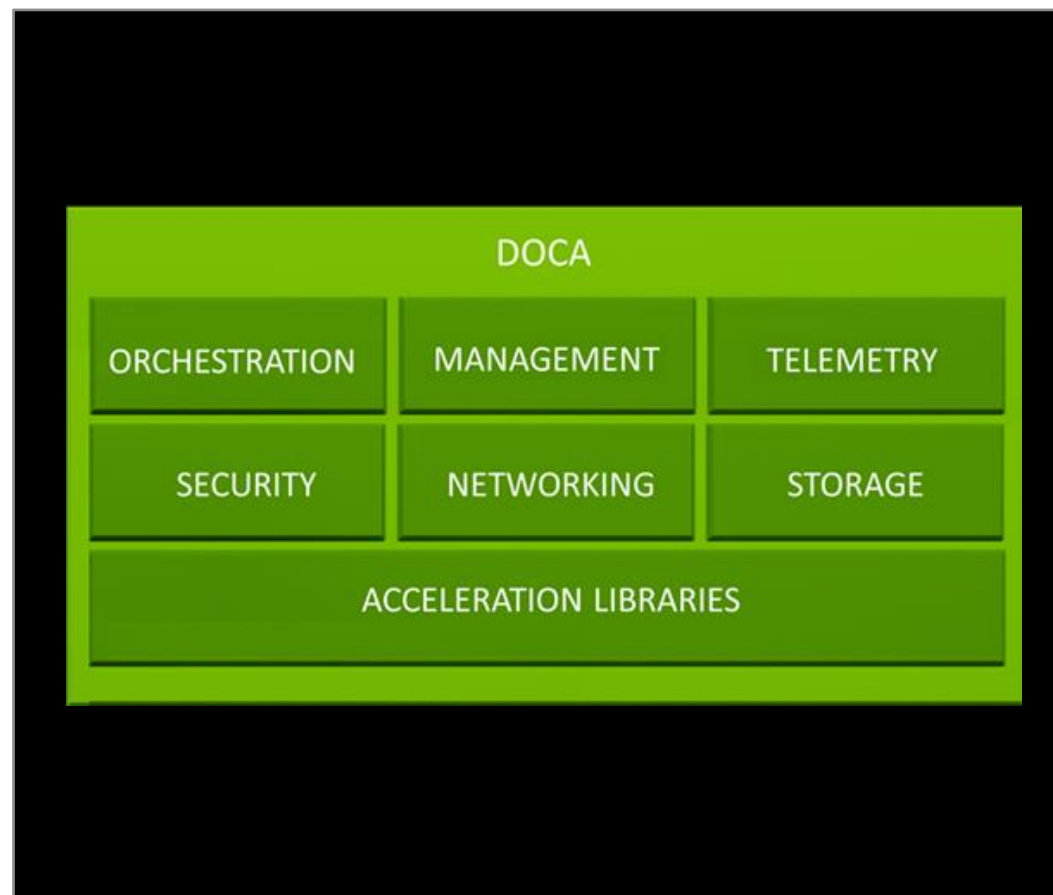


1024 NDR Ports
2048 NDR200 Ports

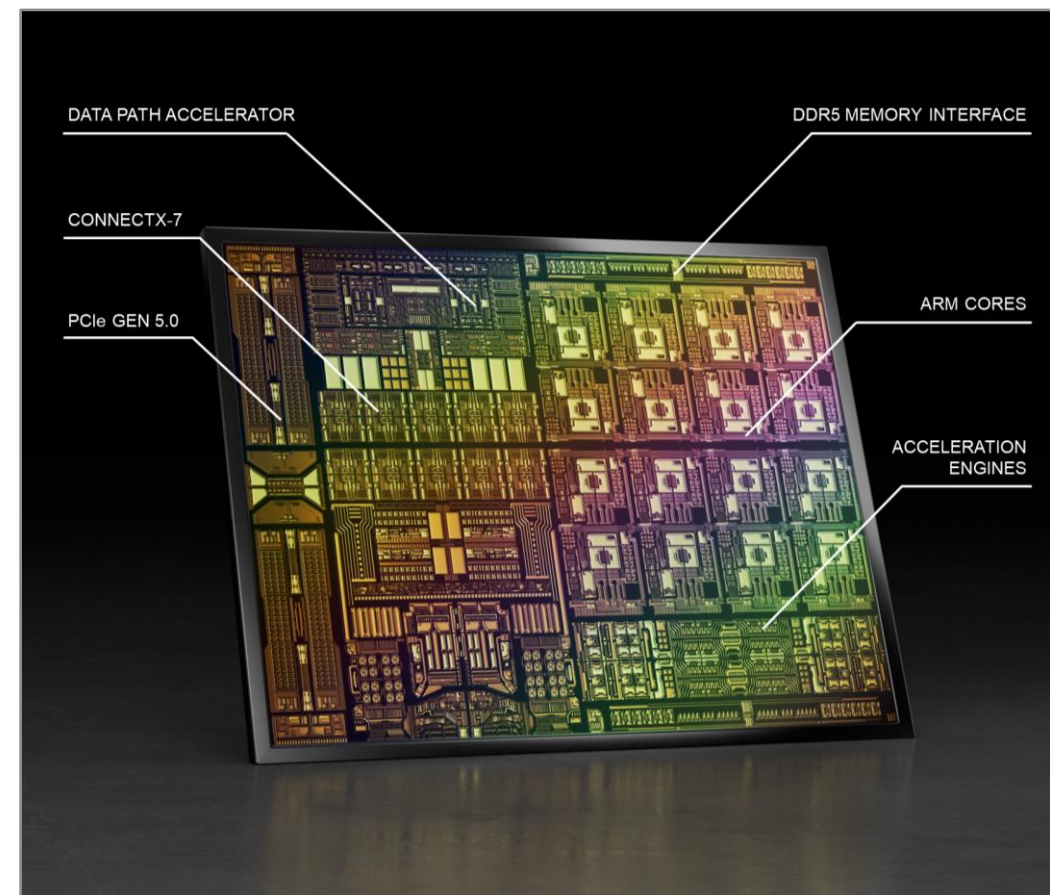


2048 NDR Ports
4096 NDR200 Ports

FROM SUPERCOMPUTERS TO SUPERCLOUDS: CLOUD-NATIVE SUPERCOMPUTERS



DOCA Enabling Growing
Partner Ecosystem



Bluefield-3 Next Generation 400G
Data Center Infra Processor



NVIDIA Quantum NDR 400G InfiniBand
In-network Computing Interconnect

