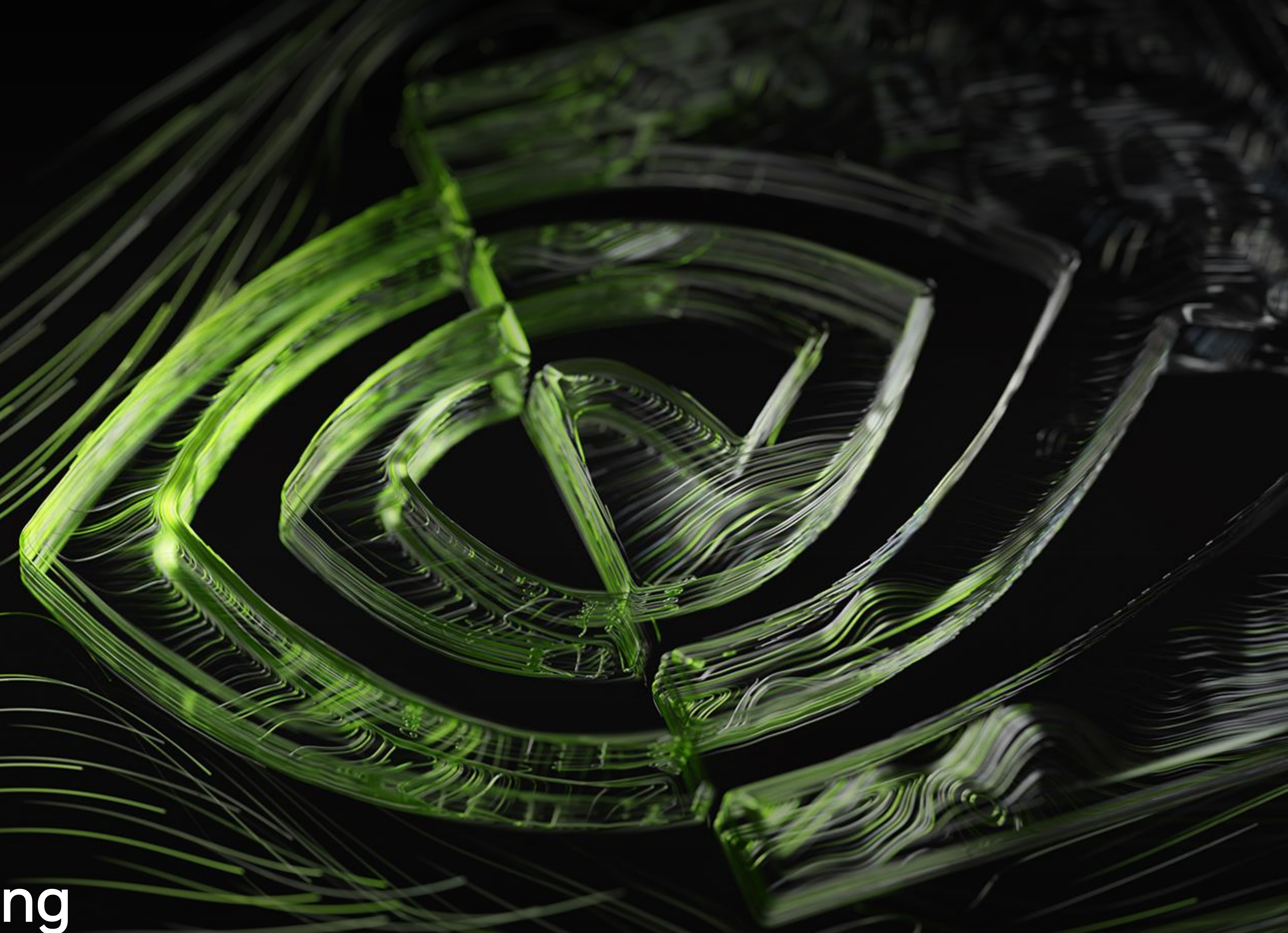


In-Network Computing

InfiniBand Quantum-2 Platform and DPU

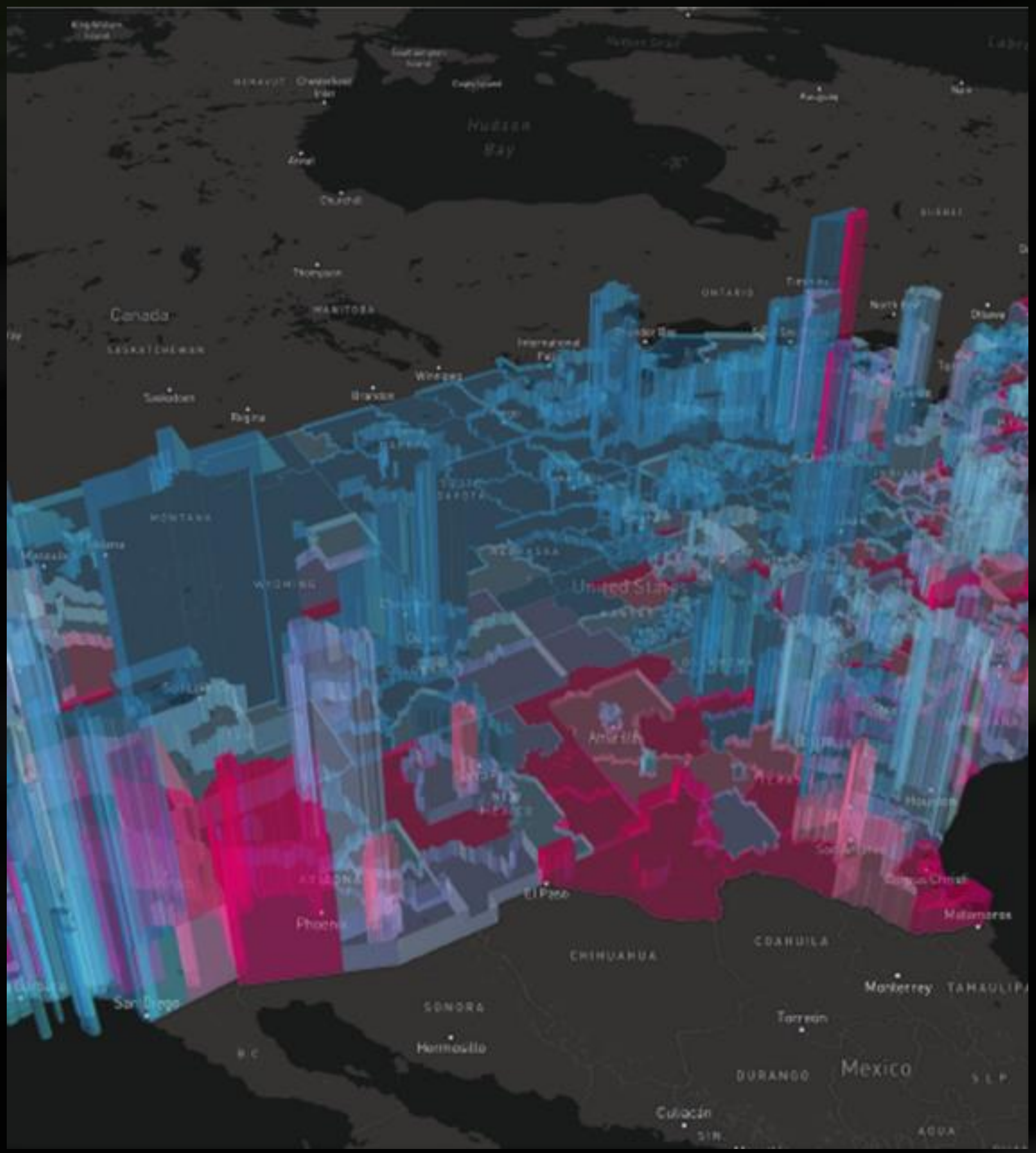
Sungta Tsai | April 2023

Next Wave of Applications

Transformative technologies opening new frontiers for thousands of new companies

Generative AI

Create new novel and exciting content



Data Science

Turn massive data sets into actionable insights

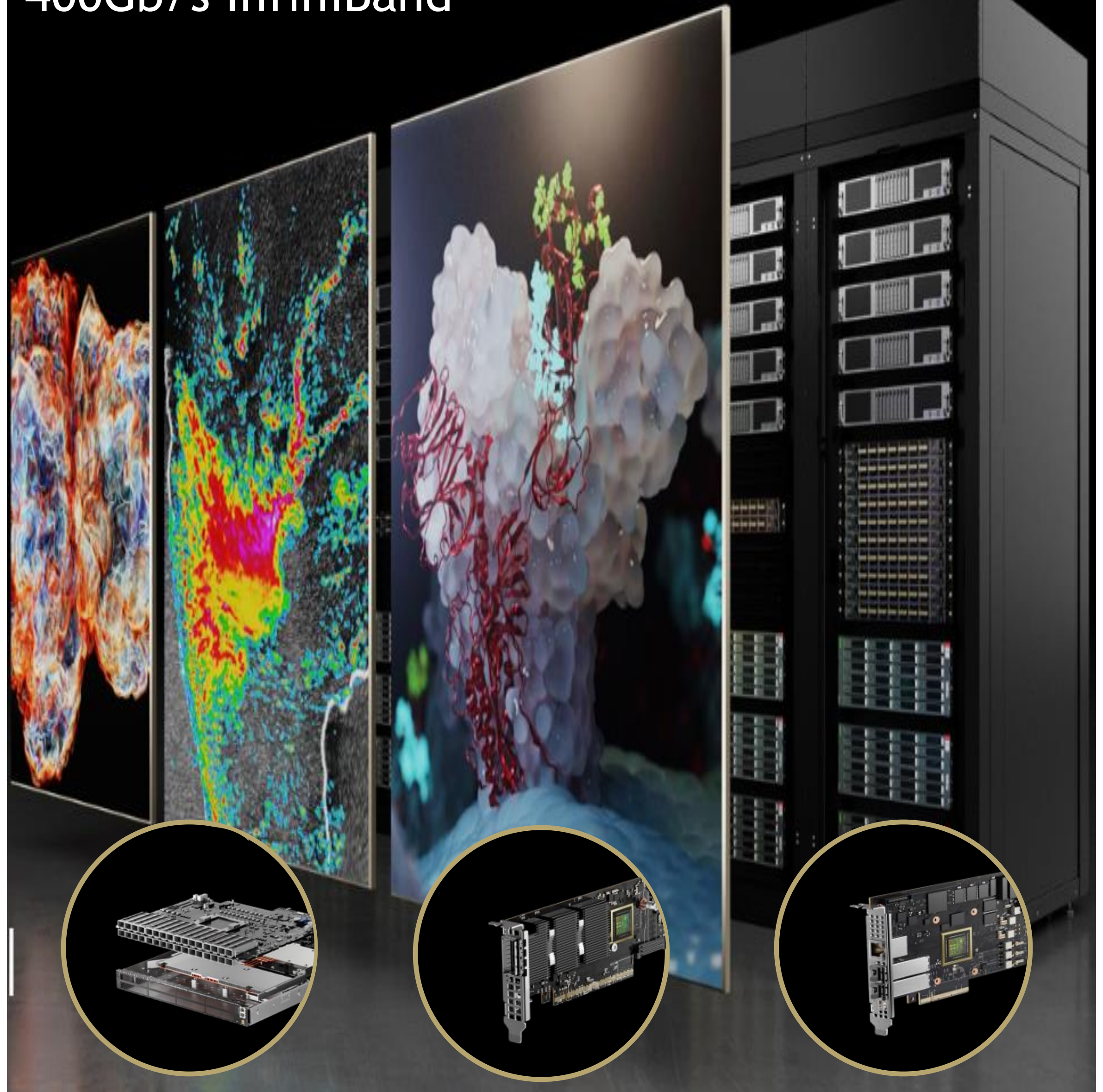
The Metaverse

Design, build, and operate virtual worlds and digital twins

Quantum-2 InfiniBand Platform

Unprecedented Performance, Scalability, and Security for HPC-AI

Bare-Metal Secured Multi-Tenant Infrastructure
Performance Isolation with Congestion Control
Advanced Adaptive Routing
In-Network Computing
400Gb/s InfiniBand



NVIDIA Quantum-2 Switch

BlueField-3 DPU

ConnectX-7

Most Advanced Networking

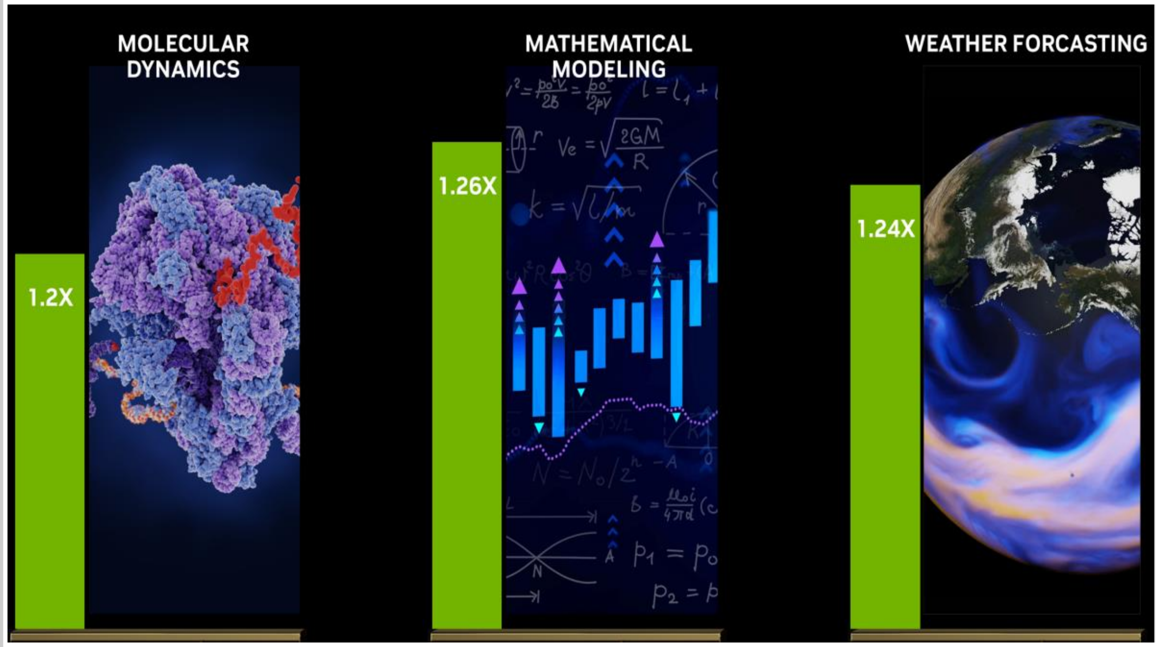
End-to-End	High Throughput	Extremely Low Latency	High Message Rate
	RDMA	GPUDirect RDMA	GPUDirect Storage
	Adaptive Routing	Congestion Control	Smart Topologies

In-Network Computing

Adapter or/ 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

Los Alamos NATIONAL LABORATORY
Durham University
Microsoft Azure
TACC
UNIVERSITY OF CAMBRIDGE
OAK RIDGE National Laboratory

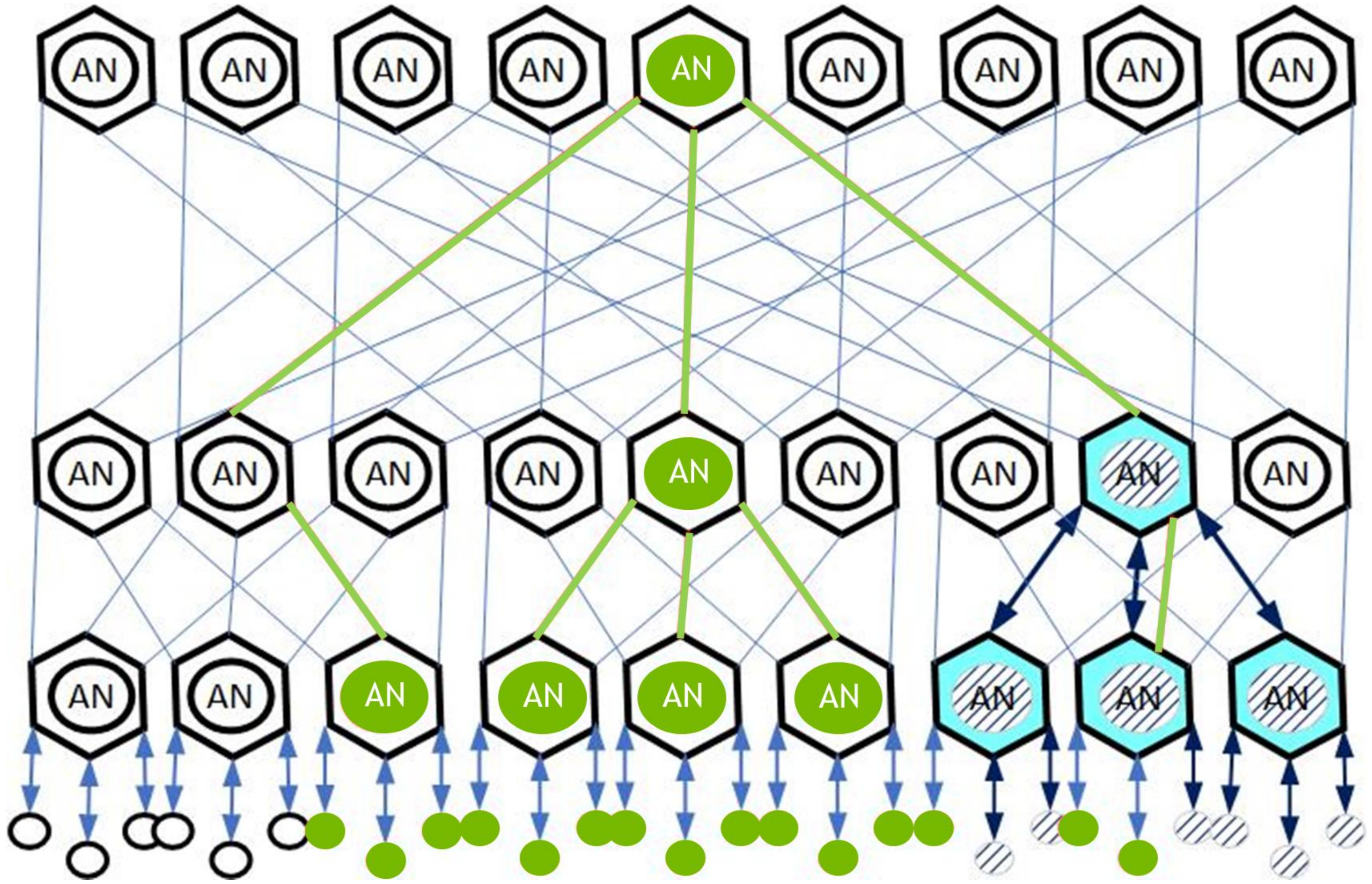
1.2x Higher Application Performance with BlueField DPU and Quantum InfiniBand In-Network Computing



Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

Offloads collective operations from the host to the network switch

- 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

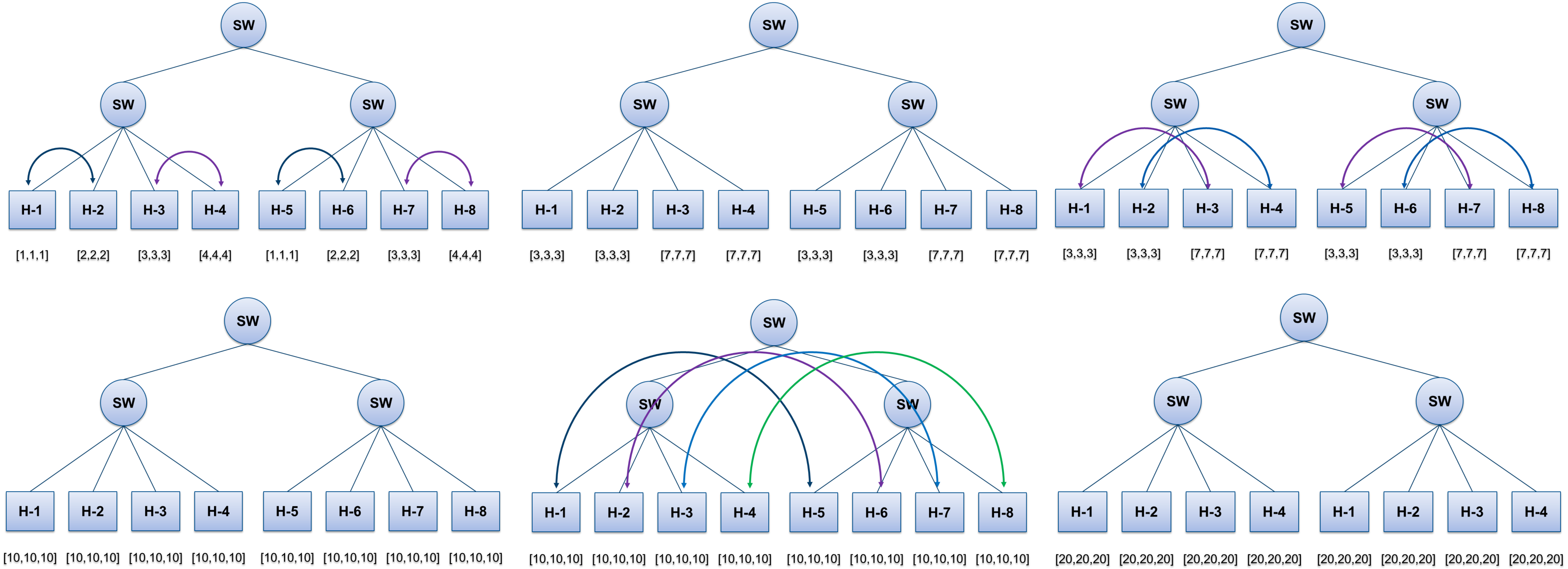

AN AN AN AN AN AN AN AN AN
AN AN AN AN AN AN AN AN AN
AN AN AN AN AN AN AN AN AN

AN SHARP Aggregation Node: Switch Resident

○ Host: Data source and Destination

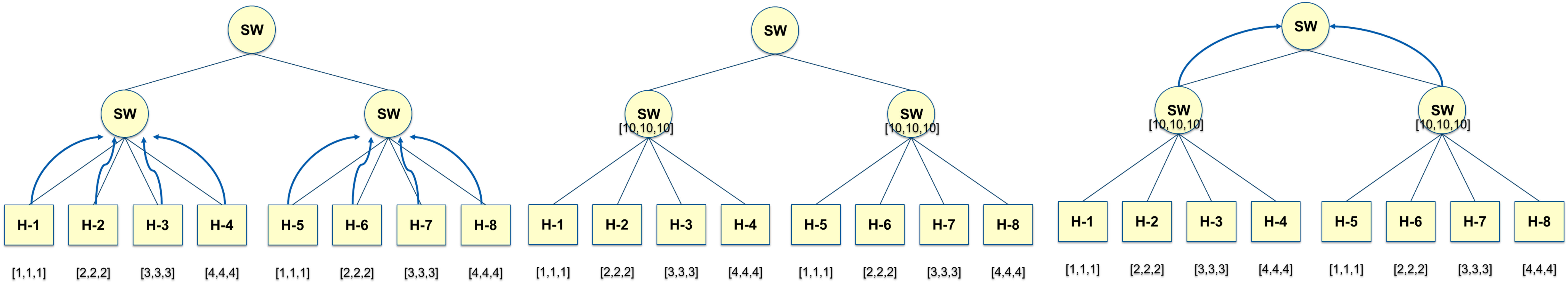
Scalable Hierarchical Aggregation and Reduction Protocol

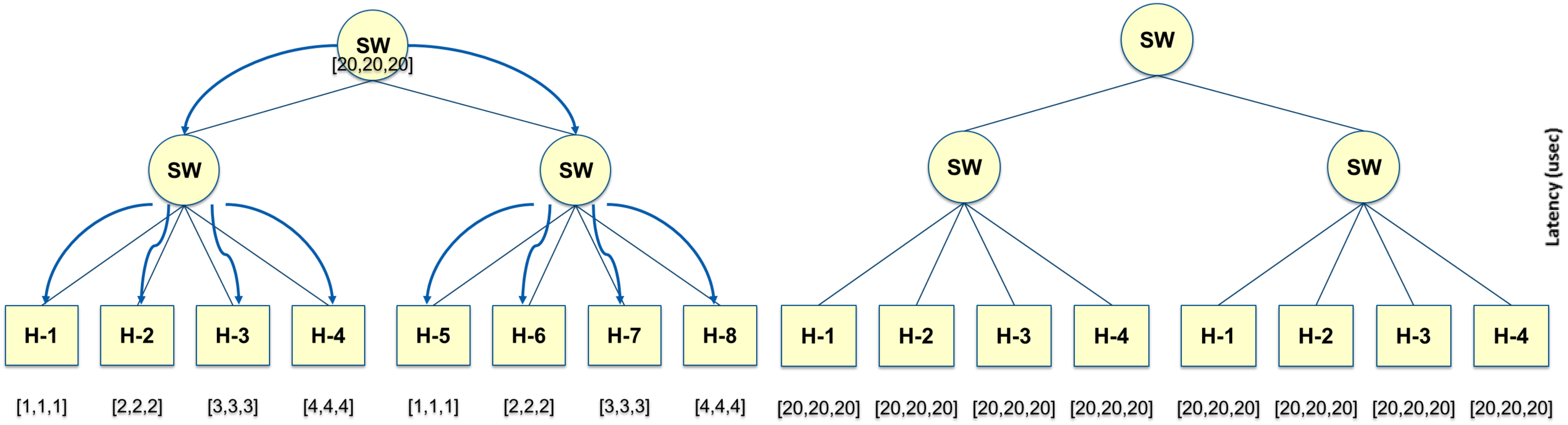
Recursive doubling Algorithm for AllReduce Operation

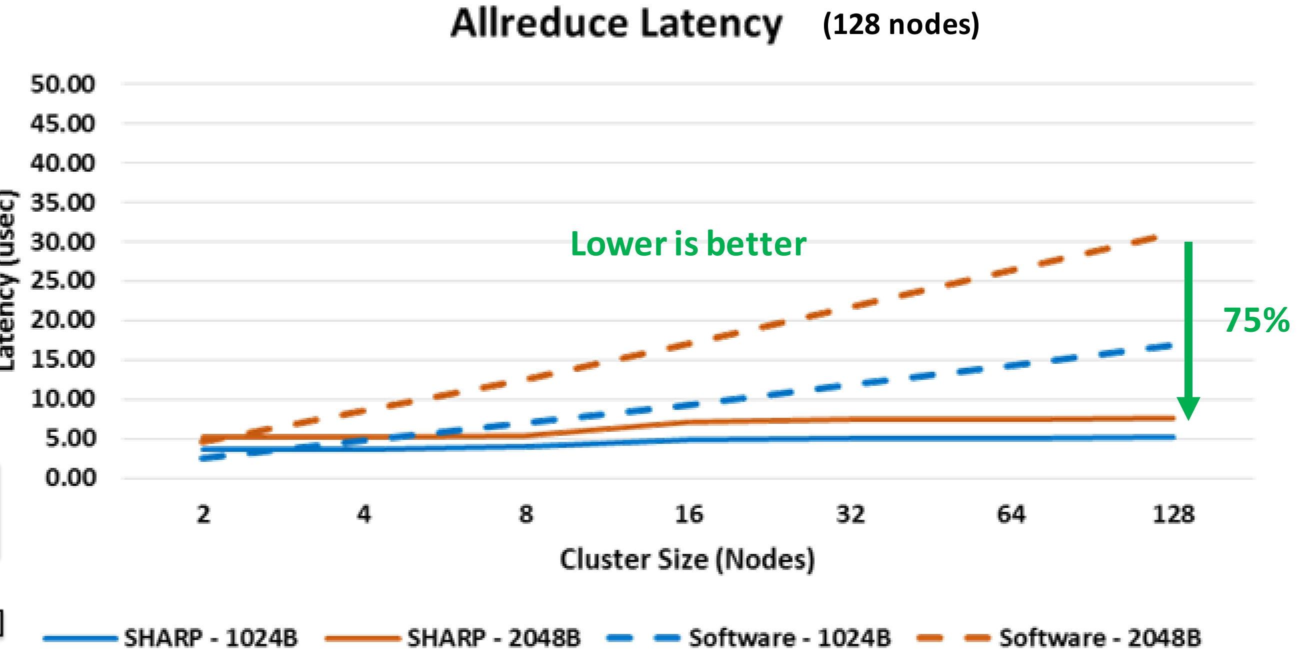


Scalable Hierarchical Aggregation and Reduction Protocol

SHARP for AllReduce Operation







Comparison

Discussion

Decoupling latency from node count*

Without the capabilities of SHARP, Allreduce requires at least O(log(N)) phases.

SHARP operates level-by-level in the topology, proportional to the tree height.

Offload

Standard algorithms require the compute nodes to perform the reduction's compute.

SHARP boosts the effective utilization of the endpoints.

Network performance and consistency

SHARP is data-movement optimal.

Consistent, predictable network performance.

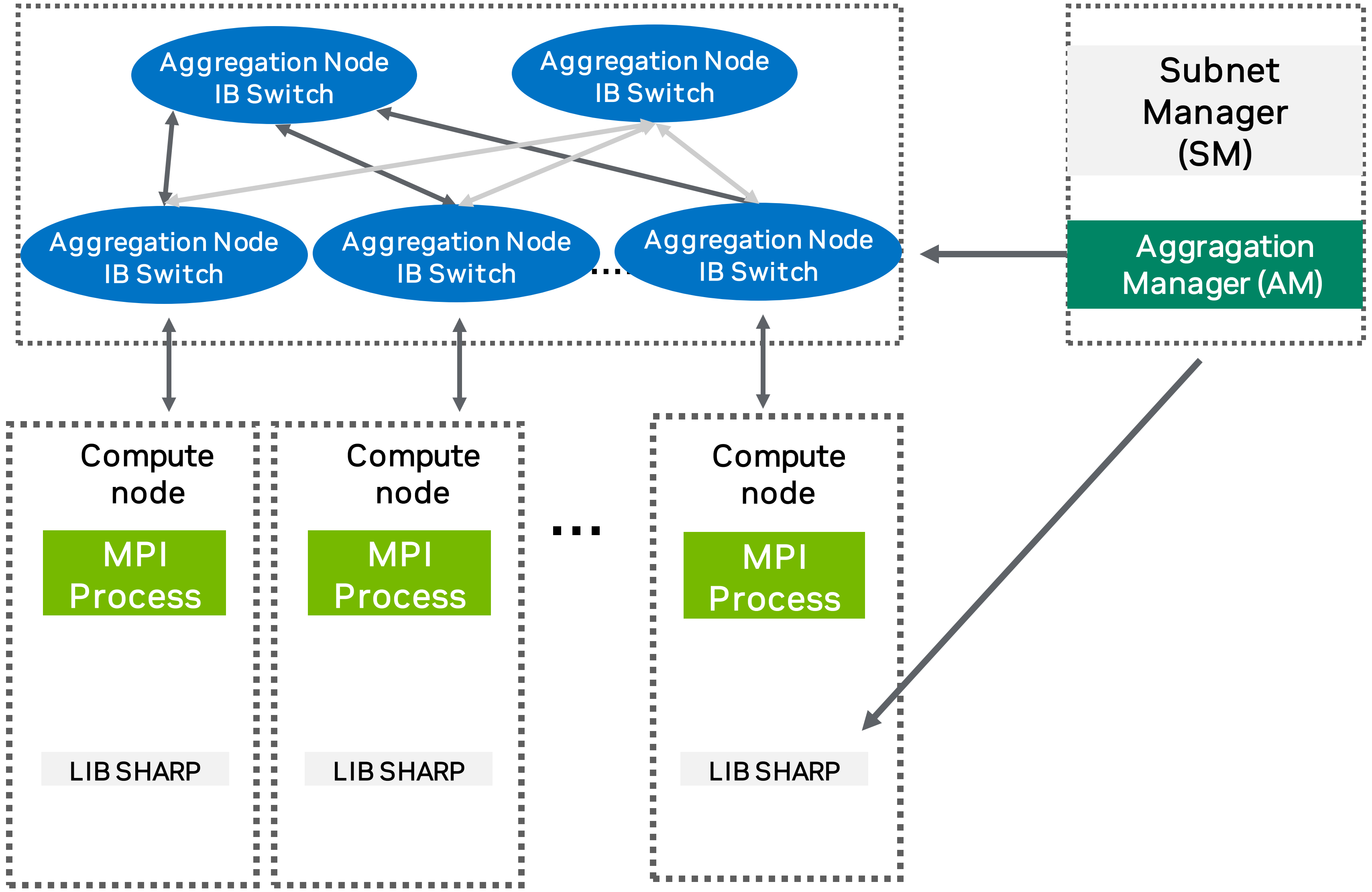
Algorithm	Latency	Bandwidth Req.
Recursive doubling	$\log2(\mathrm{N}) \cdot \alpha$	$\log2(N) \cdot m\beta$
SHARP	$2 \cdot \alpha$	$m \cdot \beta$

Description of parameters

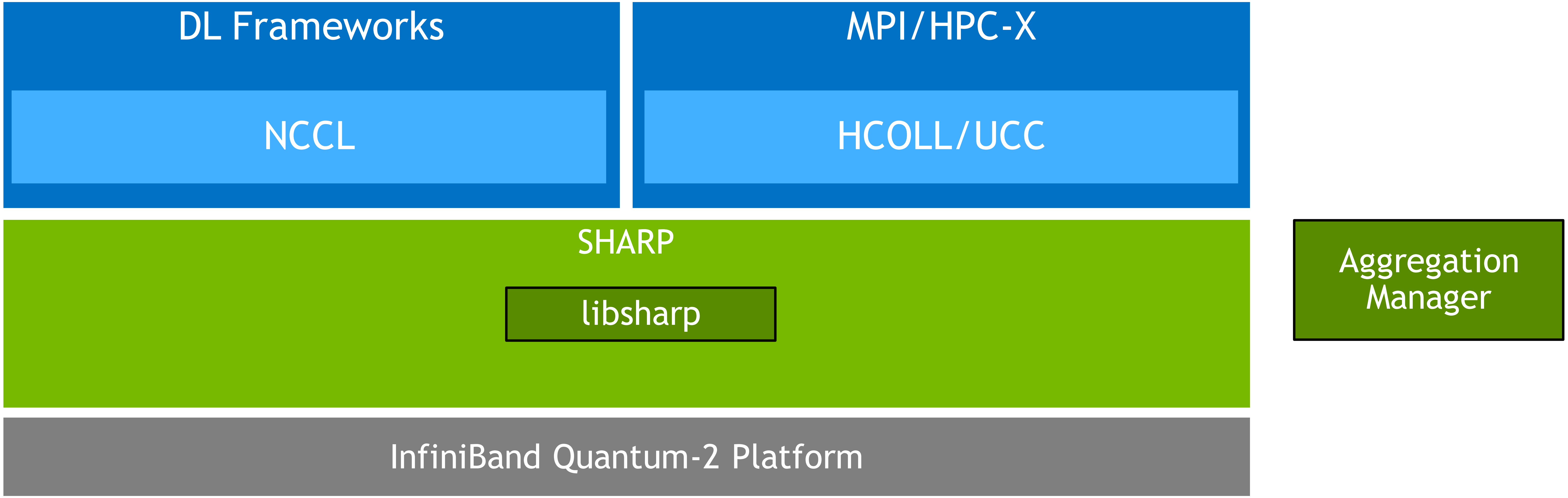
α – latency; cost of sending a single message
β – Inverse bandwidth (1/200 Gbps)
N – number of endpoints/GPUs
m – memory size of a single GPU's data (total memory: Nm)

SHARP Implementation Details

- InfiniBand networks have a Subnet Manager (“SM”) entity, which is independent of SHARP.
  - The SM populates the SMDB which contains network topology data used by SHARP.
- SHARP is managed by the Aggregation Manager (“AM”), a global software entity.
  - The AM is responsible for all SHARP resources – construction of trees, allocations, locks, and so forth.
- Once a SHARP tree has been allocated to a user process, its compute nodes can push data into the corresponding Aggregation Nodes, addressable as virtual nodes.



SHARP Software Architecture


DL Frameworks

NCCL

MPI/HPC-X

HCOLL/UCC

SHARP

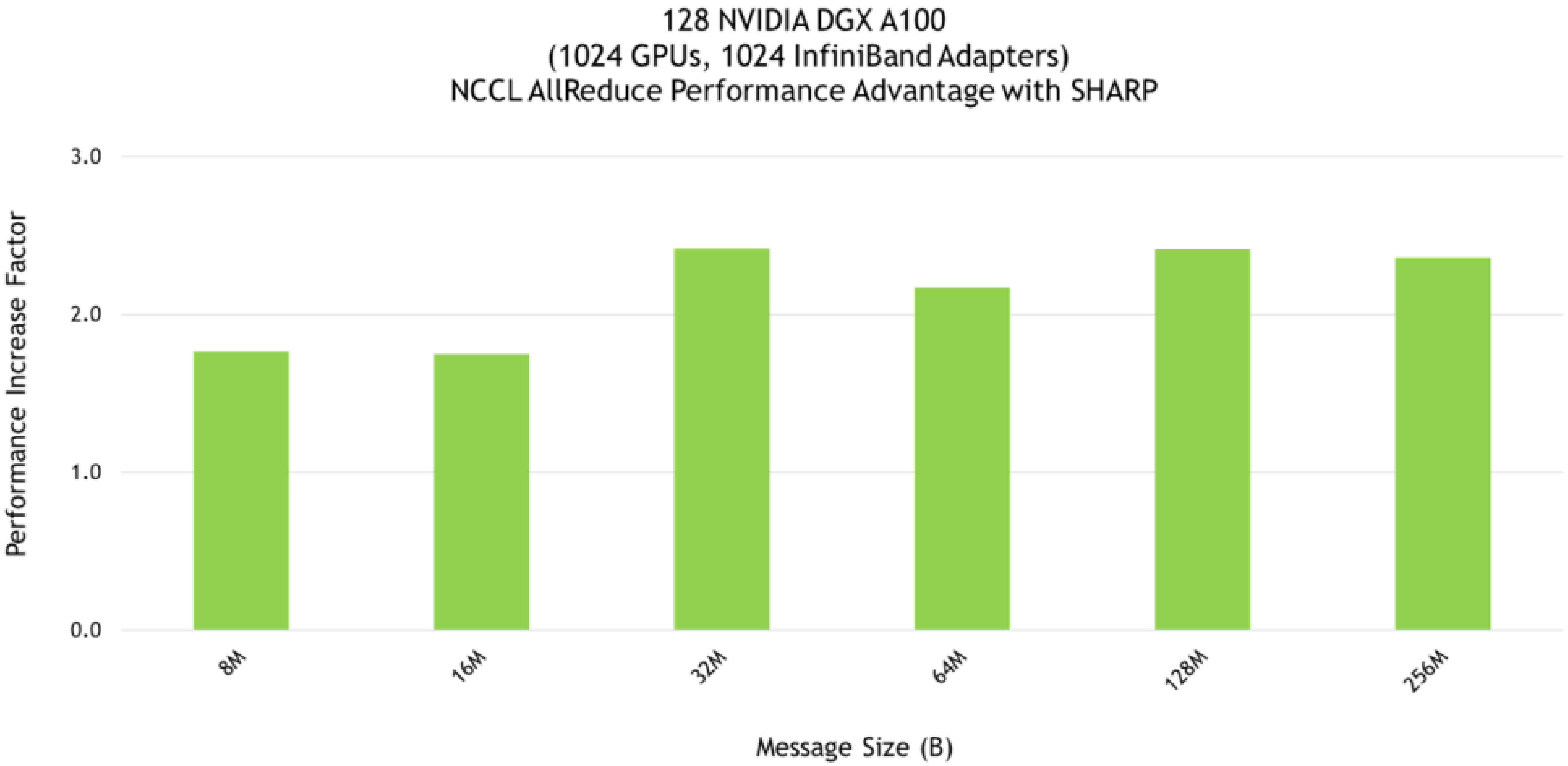
libsharp

Aggregation Manager

InfiniBand Quantum-2 Platform

SHARP AI Performance Advantages

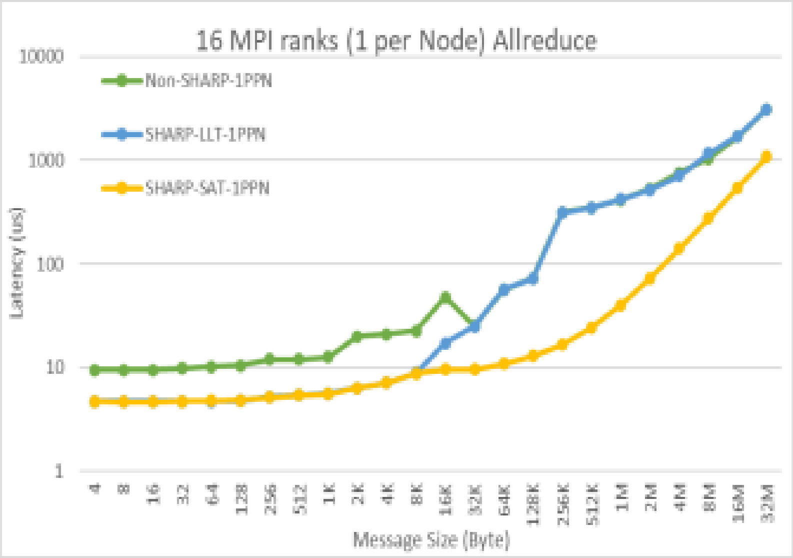
2.5X Higher Performance

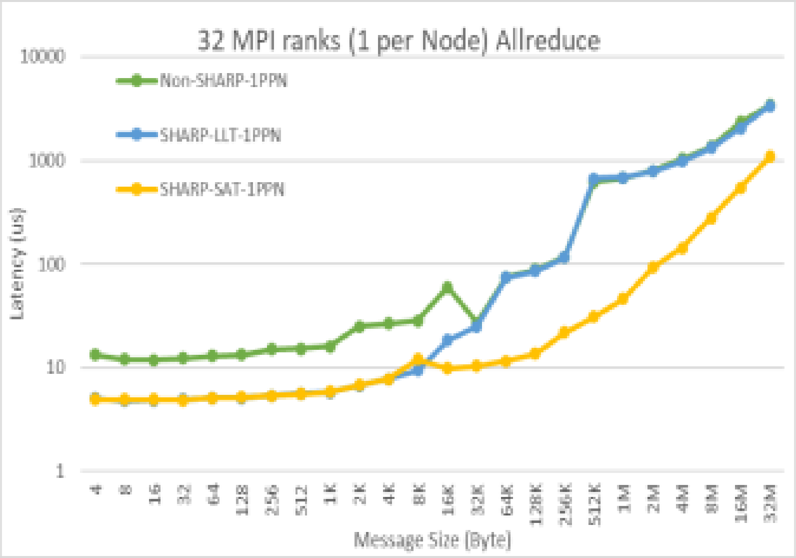


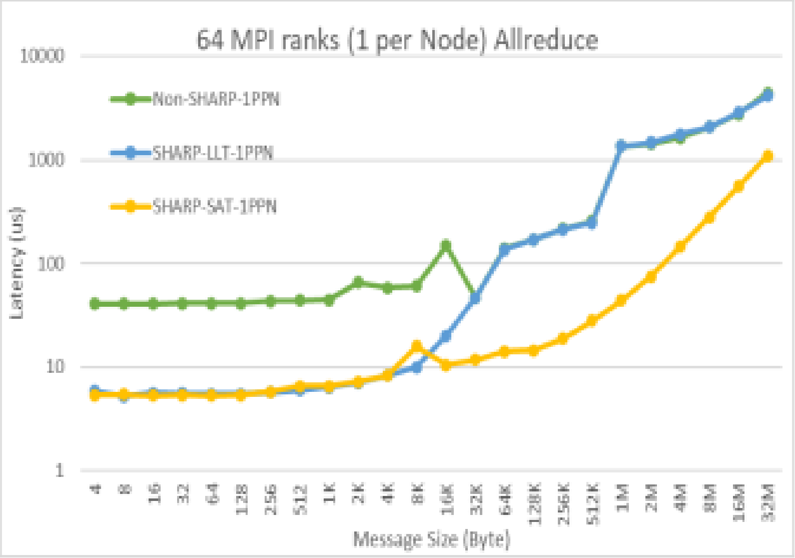
https://www.nvidia.com/en-us/on-demand/session/gtcspring21-s32067/

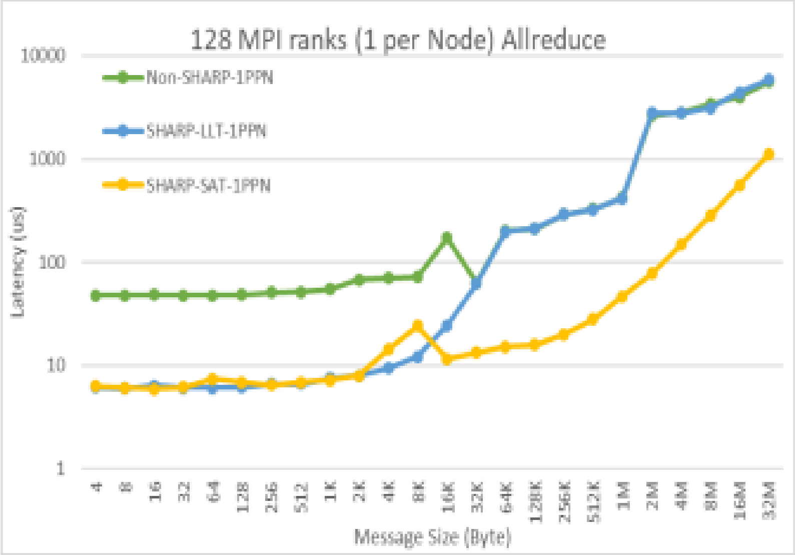
InfiniBand NDR SHARP Performance

NDR: SHARP Performance









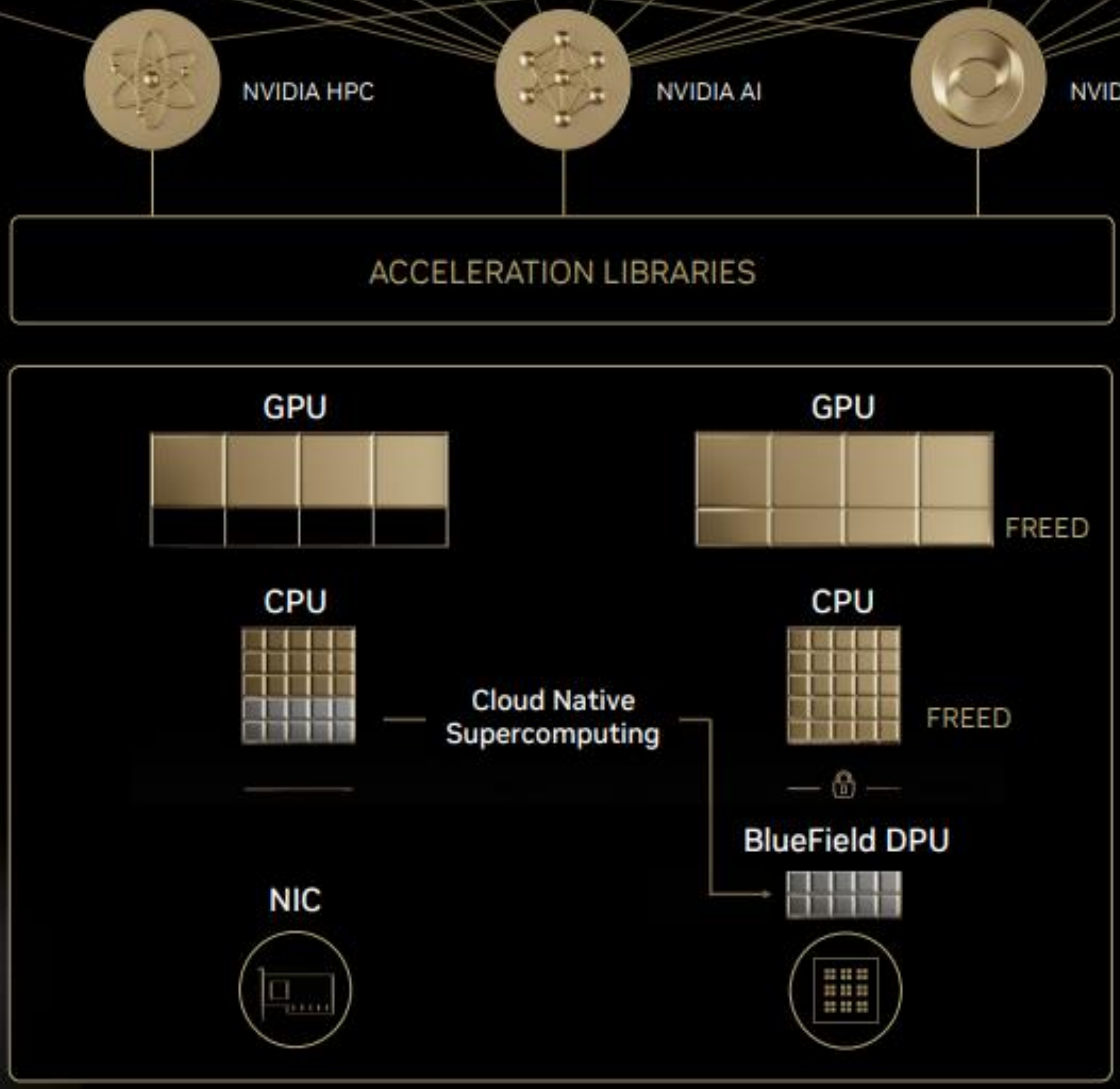
- Randomly selected nodes from cluster
- Run MPI AllReduce with 16, 32, 64, 128 nodes with PPN=1
- Significant performance improvement with both LLT and SAT protocols

Microsoft Azure

MODULUS MONAI RIVA MAXINE NEMO MERLIN CUOPT MORPHEUS TOKKIO AVATAR DRIVE ISAAC METROPOLIS HOLOSCAN

AI APPLICATION FRAMEWORK

PLATFORMS



NVIDIA HPC

NVIDIA AI

NVIDIA Omniverse

ACCELERATION LIBRARIES

GPU

CPU

NIC

Cloud Native Supercomputing

GPU

FREED

CPU

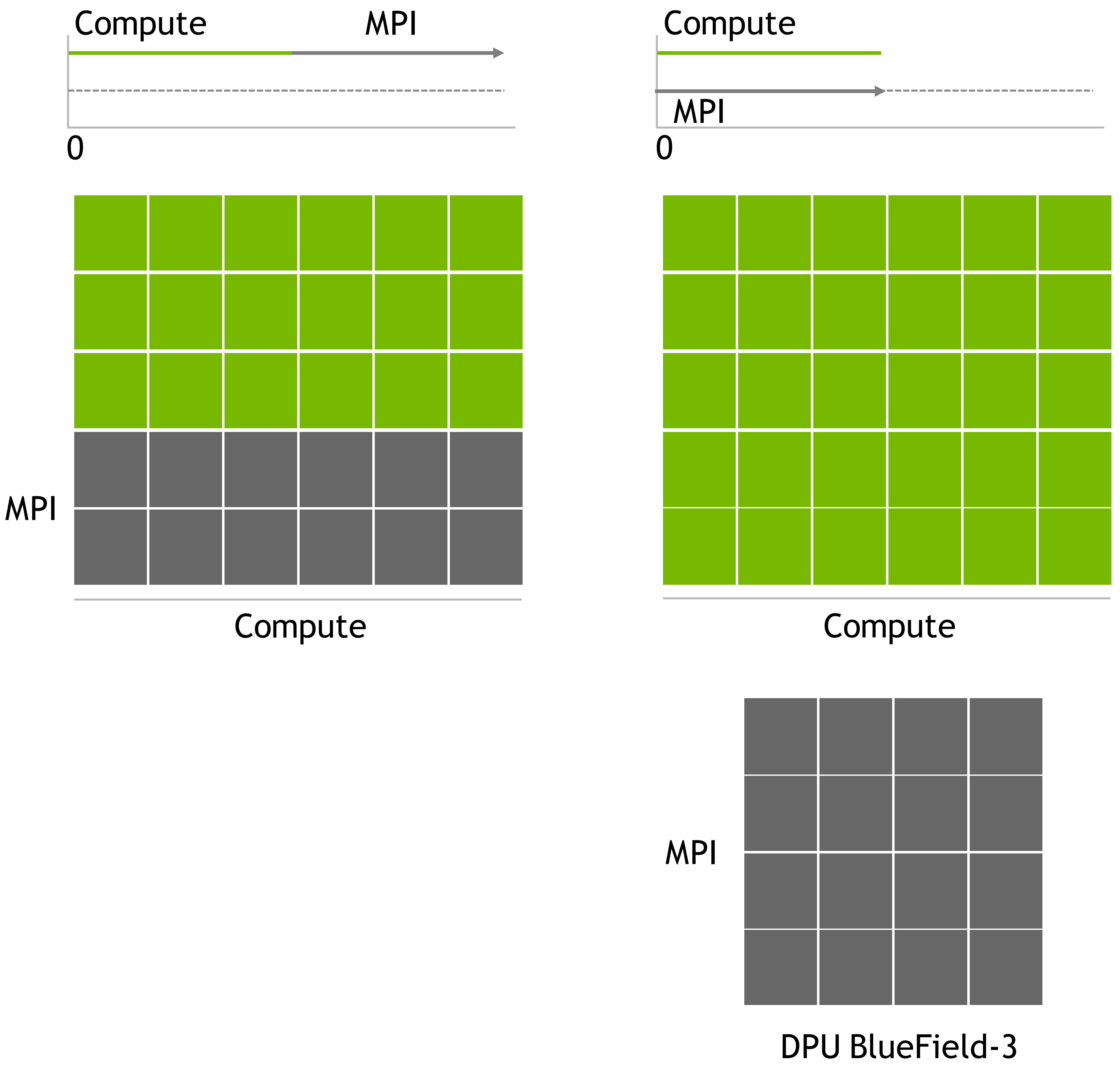
FREED

BlueField DPU

DPU for Scientific Computing

1.2x Higher Application Performance with MPI Acceleration

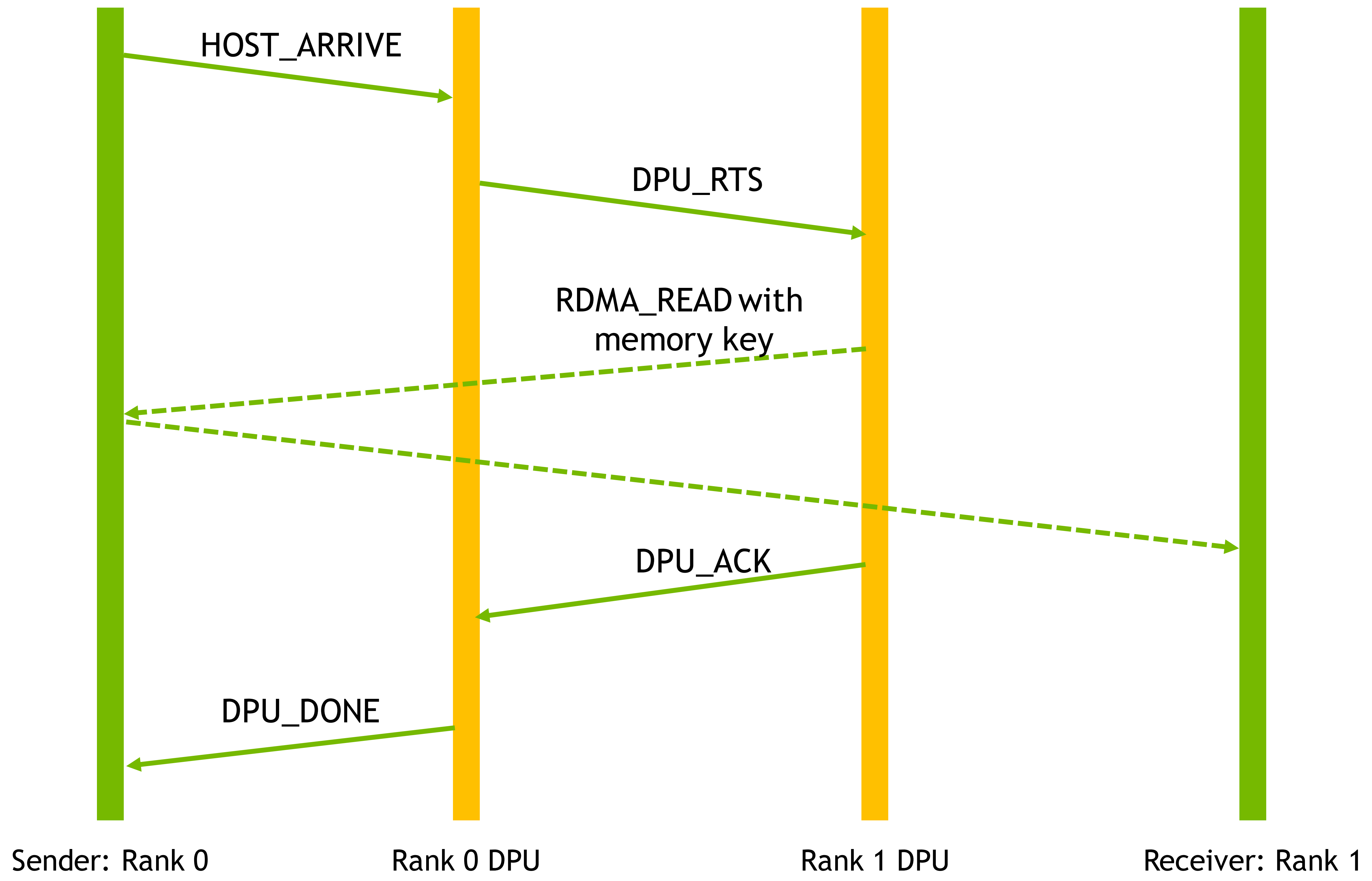
- Parallel communications impose large overhead on CPU centric platform
- Offloading and accelerating these operations maximize performance
- Enabling computations and communications to be performed in parallel



Los Alamos NATIONAL LABORATORY · TACC · Durham University · UNIVERSITY OF CAMBRIDGE · OAK RIDGE National Laboratory · Japan Meteorological Agency

Offloading and Accelerating Data Exchange Example

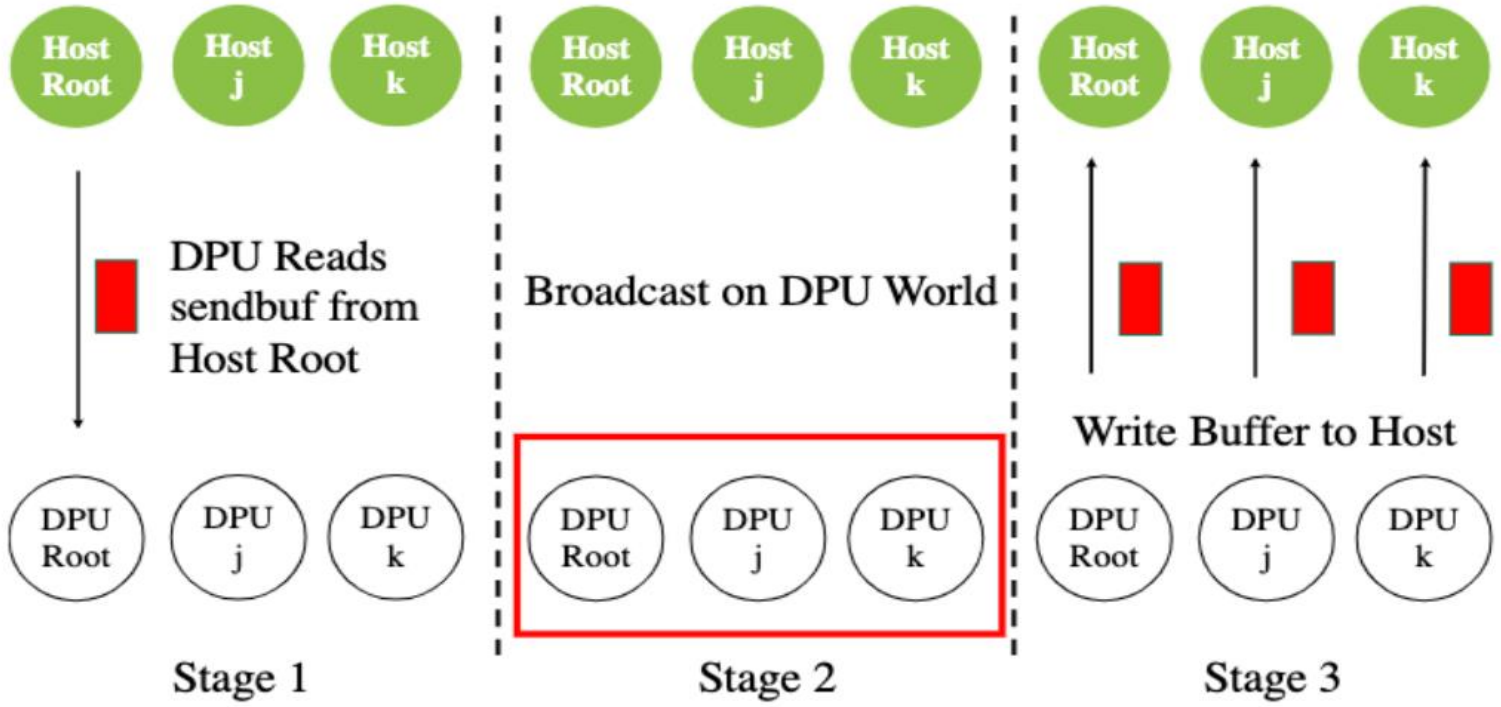
An Element of Collective Algorithm



Offload Framework

MVAPICH2

- Non-blocking collective operations are offloaded to a set of Worker processes
- BlueField is set to separated host mode
- Worker processes are spawned to the ARM cores of BlueField
- Once the application calls a collective, host processes prepare a set of metadata and provide it to the Worker processes
- Using these metadata, worker processes can access host memory through RDMA
- Worker processes progress the collective on behalf of the host processes
- Once message exchanges are completed, worker processes notify the host processes about the completion of the non-blocking operation

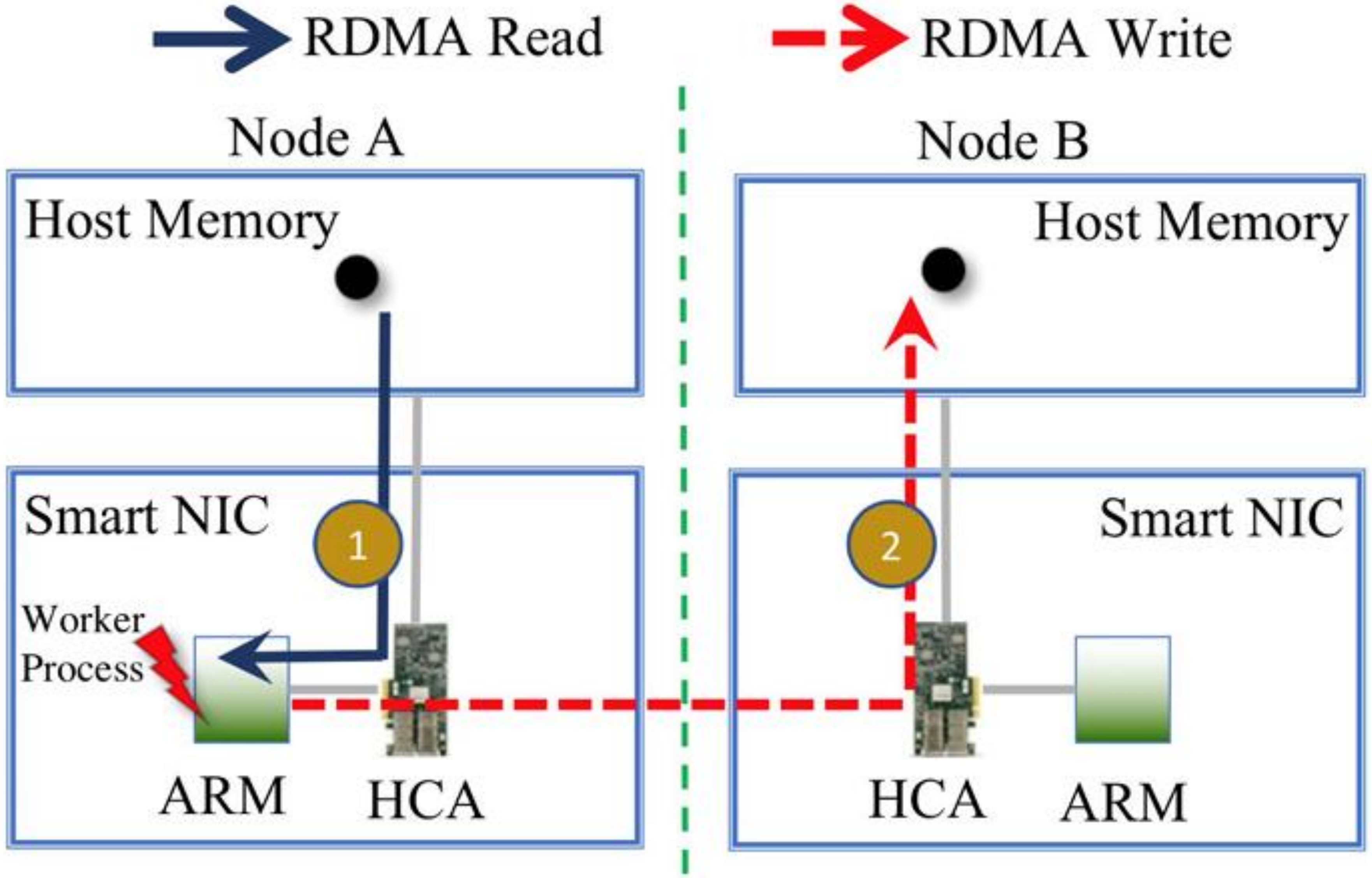


https://www.openfabrics.org/wp-content/uploads/2021-workshop-presentations/403_Panda_mpi-offload.pdf

Proposed Non-Blocking Alltoall Desgin

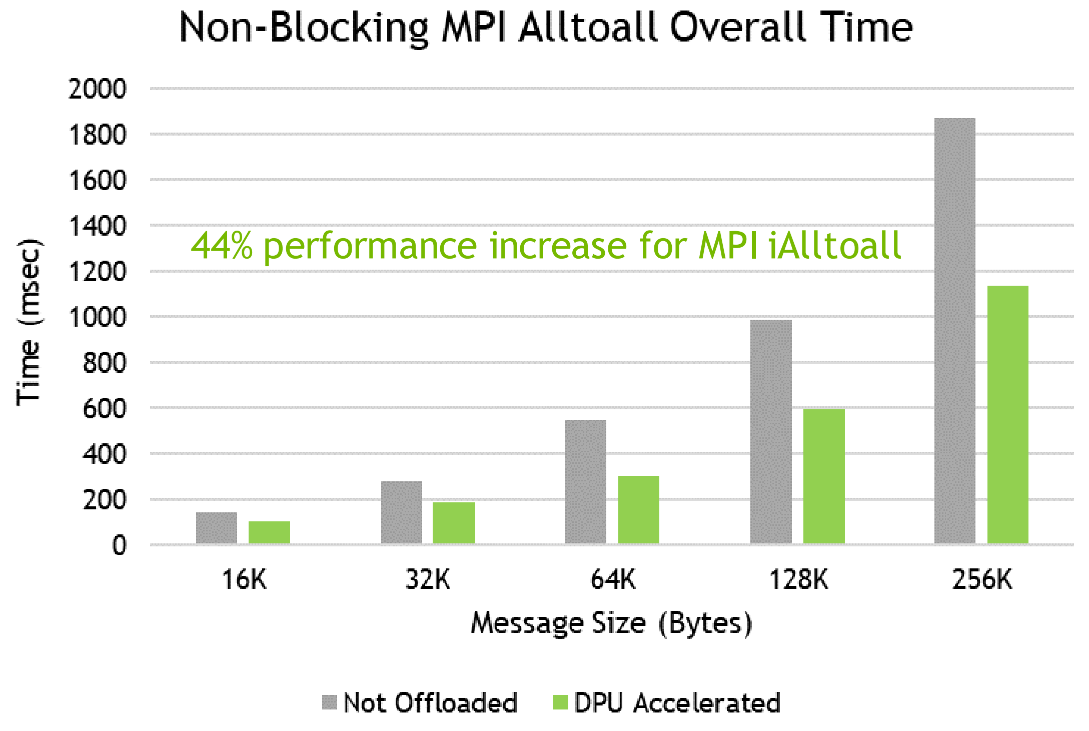
MVAPICH2

- Worker process performs RDMA Read to receive the data chunk from host main memory
- Once data is available in the ARM memory, worker process performs RDMA Write to the remote host memory

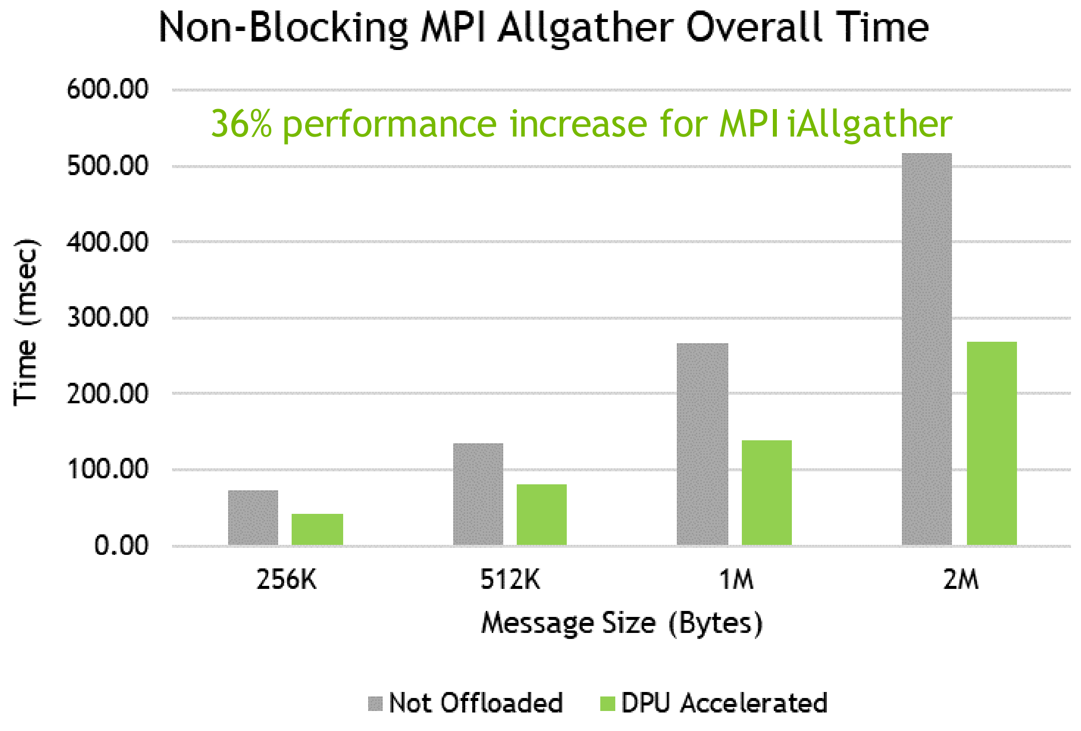


https://www.openfabrics.org/wp-content/uploads/2021-workshop-presentations/403_Panda_mpi-offload.pdf

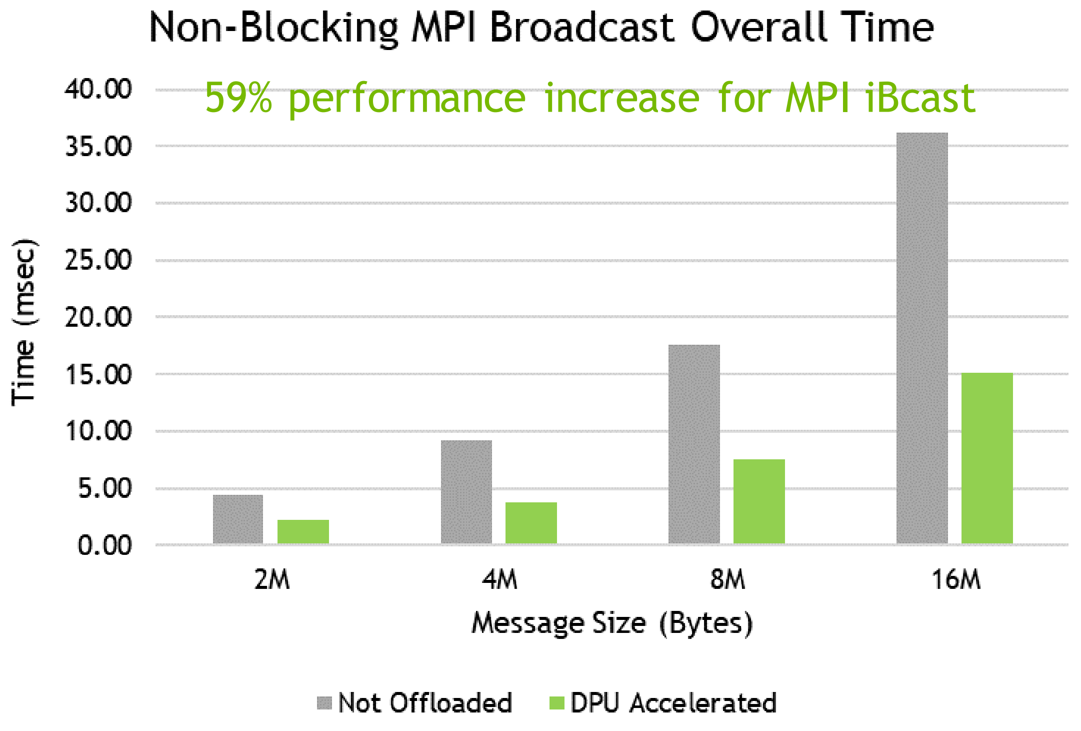
Non-Blocking MPI Performance



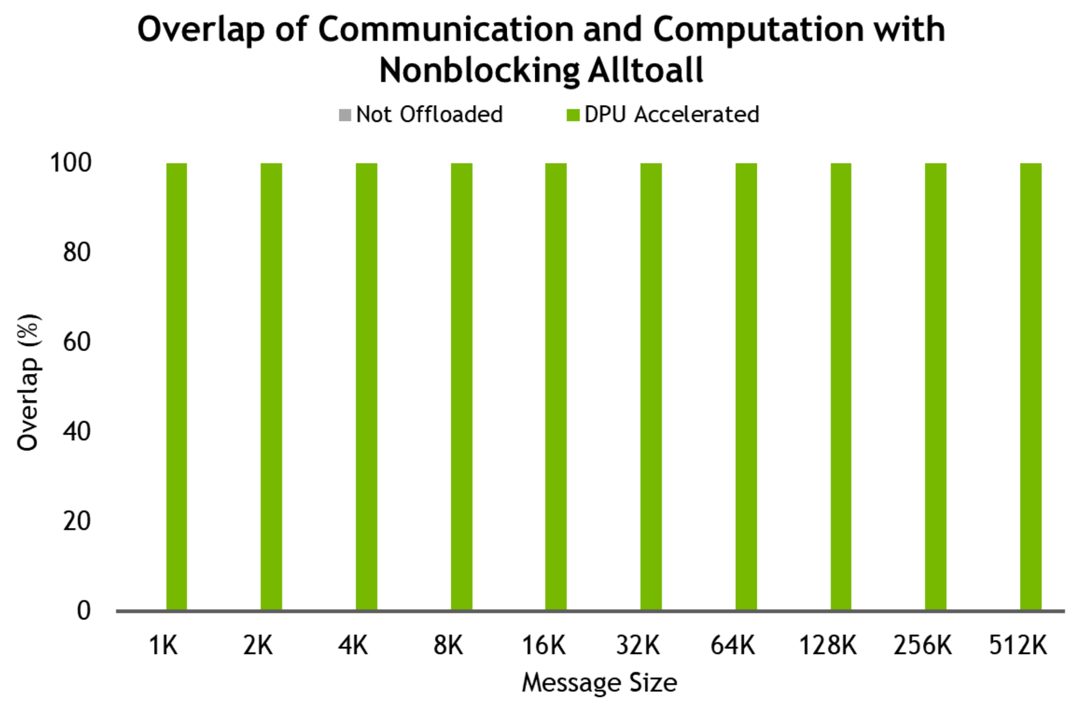
44% performance increase for MPI iAlltoall



36% performance increase for MPI iAllgather



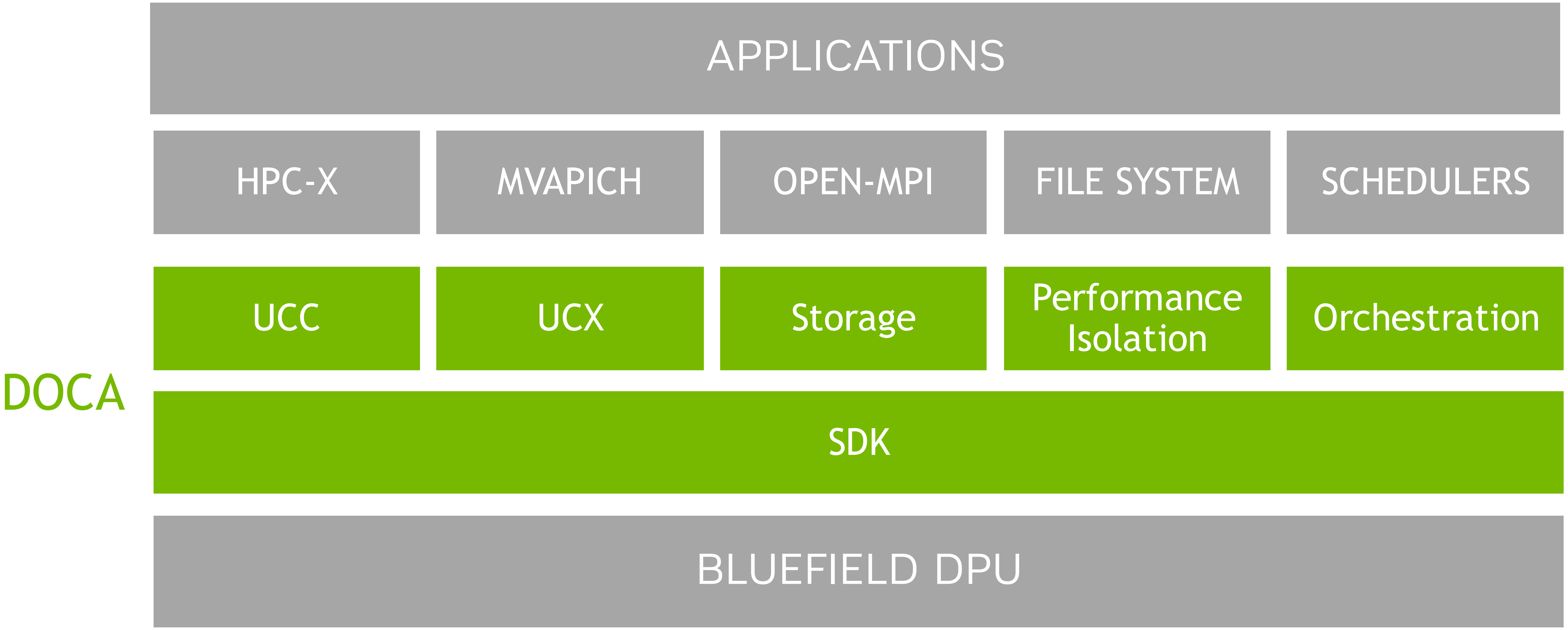
59% performance increase for MPI iBcast



100% Communication -Computation Overlap

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

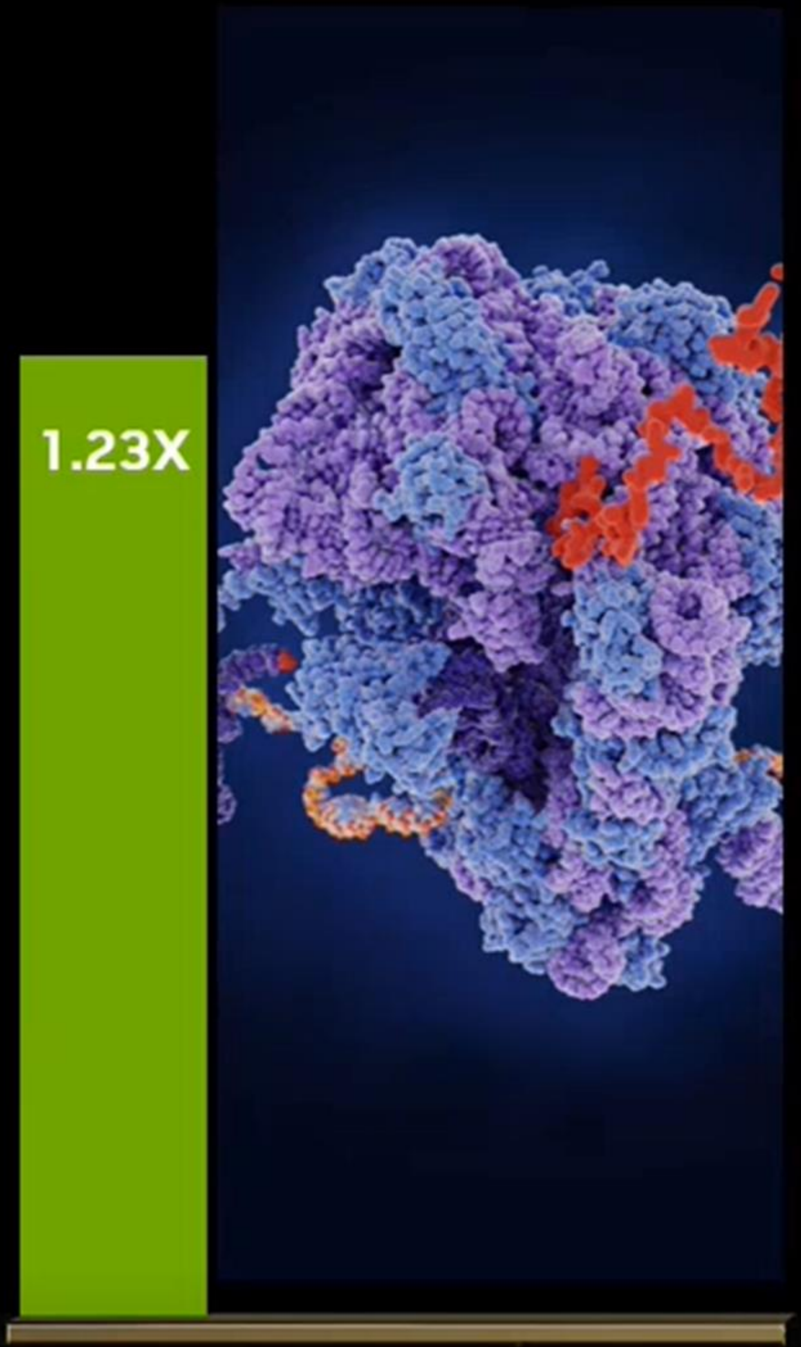
Accelerating HPC Applications with DOCA Services



Higher Performance and Cost Saving

With BlueField DPU and Quantum InfiniBand In-Network Computing

Octopus (Physics / Chemistry)



1.23	Higher Performance
1.17	Higher Performance / TCO$
1.19	Higher Performance/Watt

JOIN NVIDIA DEVELOPER PROGRAM TODAY

Supporting the Community That's Changing the World

LEARN MORE ABOUT DEVELOPER PROGRAM AND JOIN NOW.
https://developer.nvidia.com/developer-program

GET NVIDIA OPTIMIZED CONTAINERS, MODELS AND MORE.
https://developer.nvidia.com/downloads

EXPLORE NVIDIA INCEPTION AND APPLY TODAY.
https://www.nvidia.com/en-us/startups