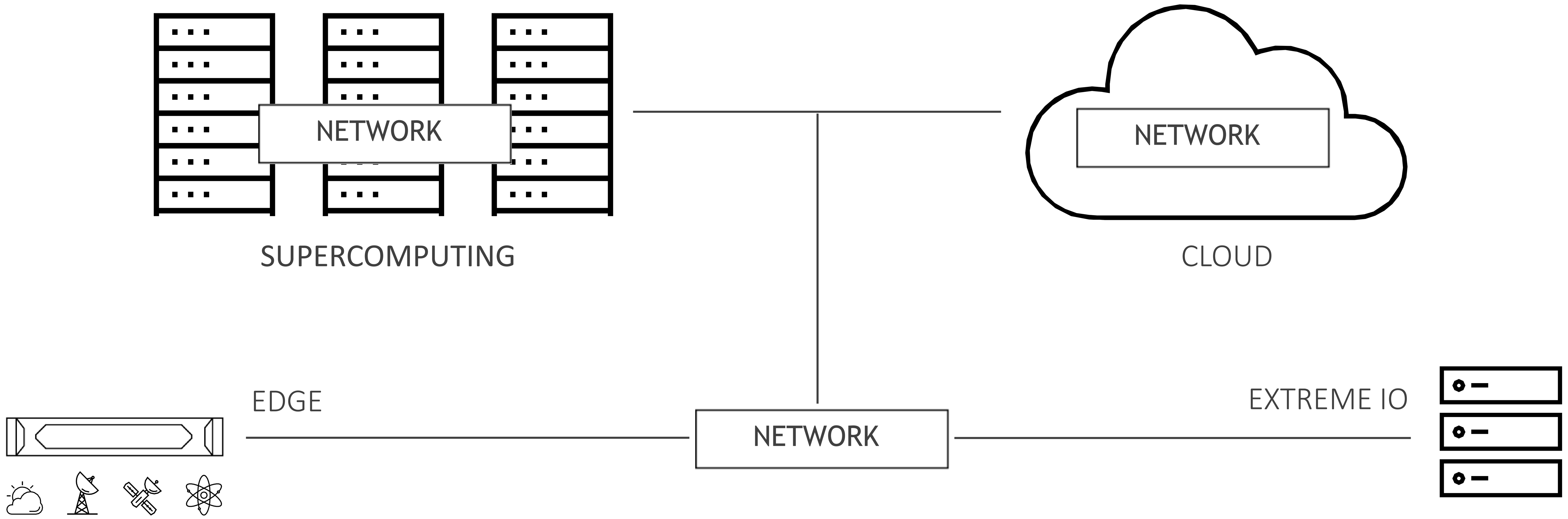
INTRODUCING CLOUD-NATIVE SUPERCOMPUTING BARE-METAL, SECURED SUPERCOMPUTING ARCHITECTURE

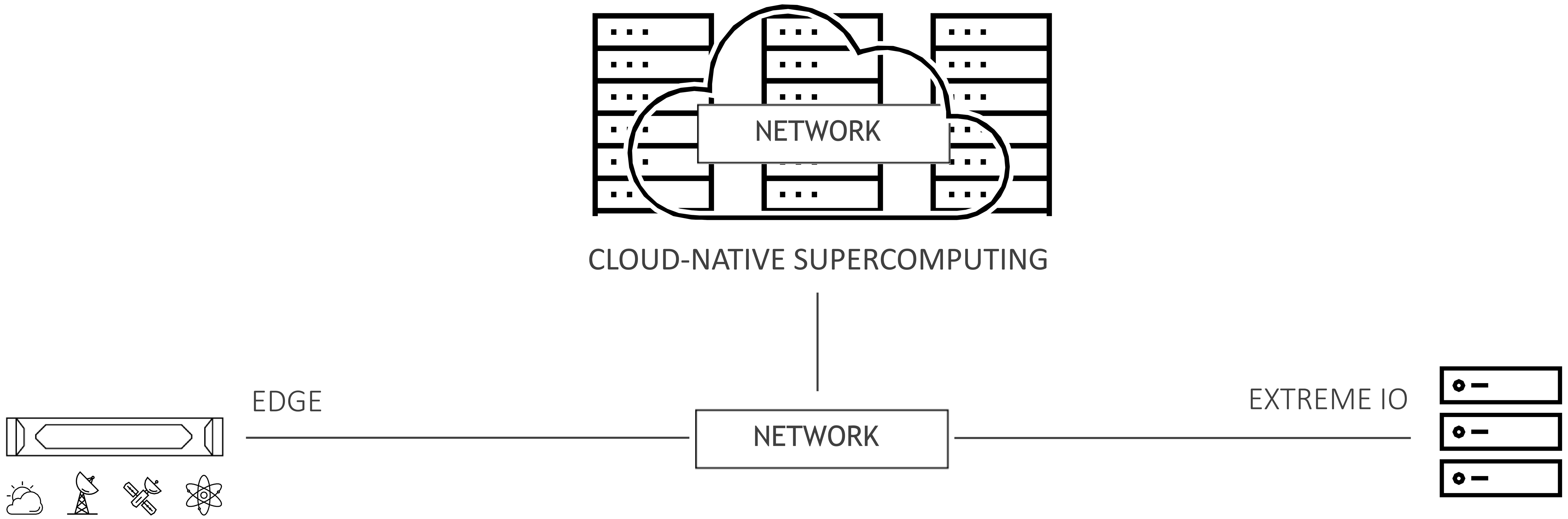
RICH GRAHAM, GILAD SHAINER, JITHIN JOSE

DECEMBER 2021

DIVERSITY OF APPLICATIONS REQUIRES ARCHITECTURAL FLEXIBILITY



DIVERSITY OF APPLICATIONS REQUIRES ARCHITECTURAL FLEXIBILITY



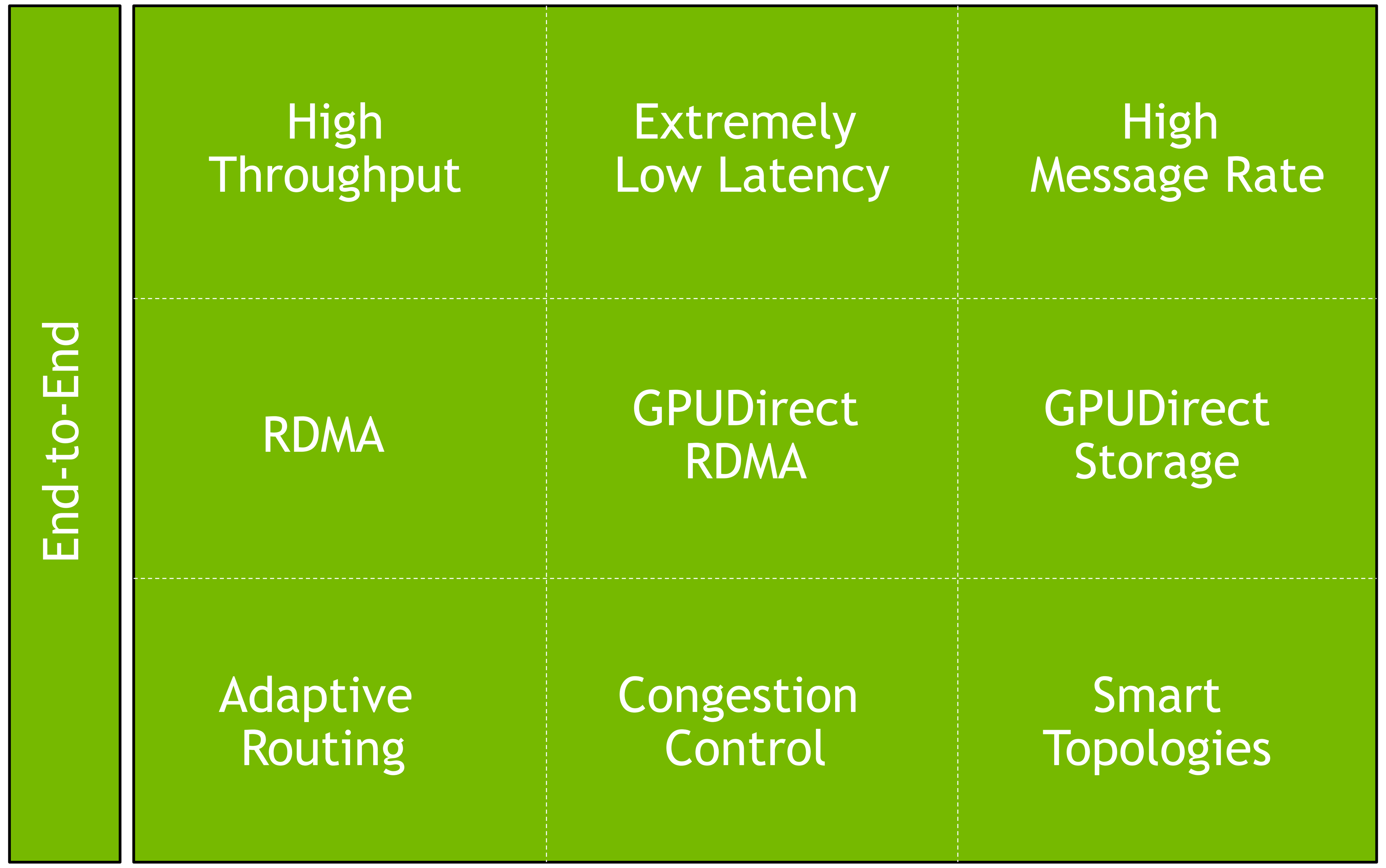
NVIDIA


NVIDIA INFINIBAND CLOUD-NATIVE SUPERCOMPUTING
QUANTUM-2 SWITCH
CONNECTX-7 INFINIBAND
BLUEFIELD-3 INFINIBAND

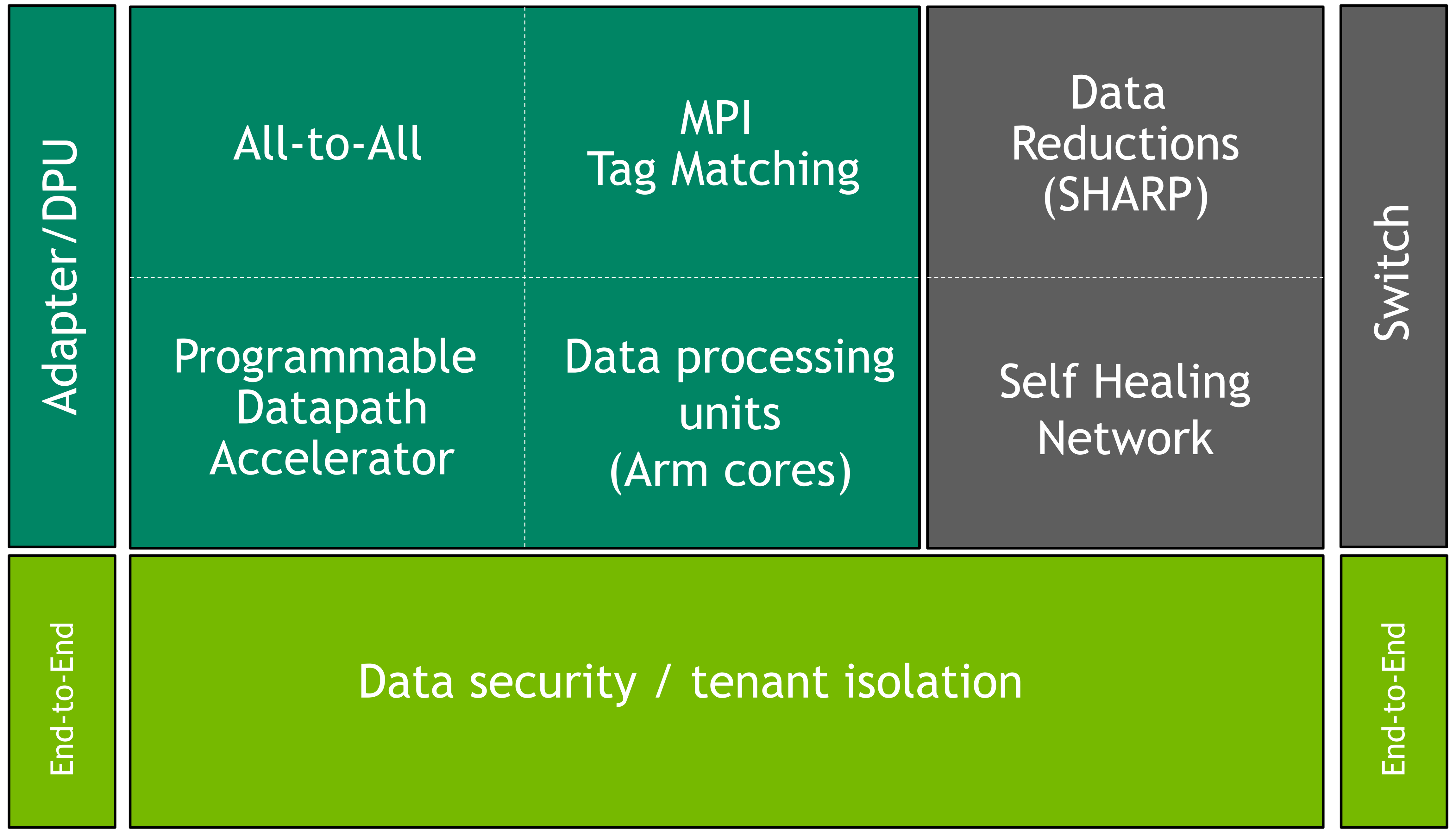
IN-NETWORK COMPUTING ACCELERATED SUPERCOMPUTING

Software-Defined, Hardware-Accelerated, InfiniBand Network

Most Advanced Networking



In-Network Computing



CLOUD-NATIVE SUPERCOMPUTING

Bare-metal Secured Infrastructure

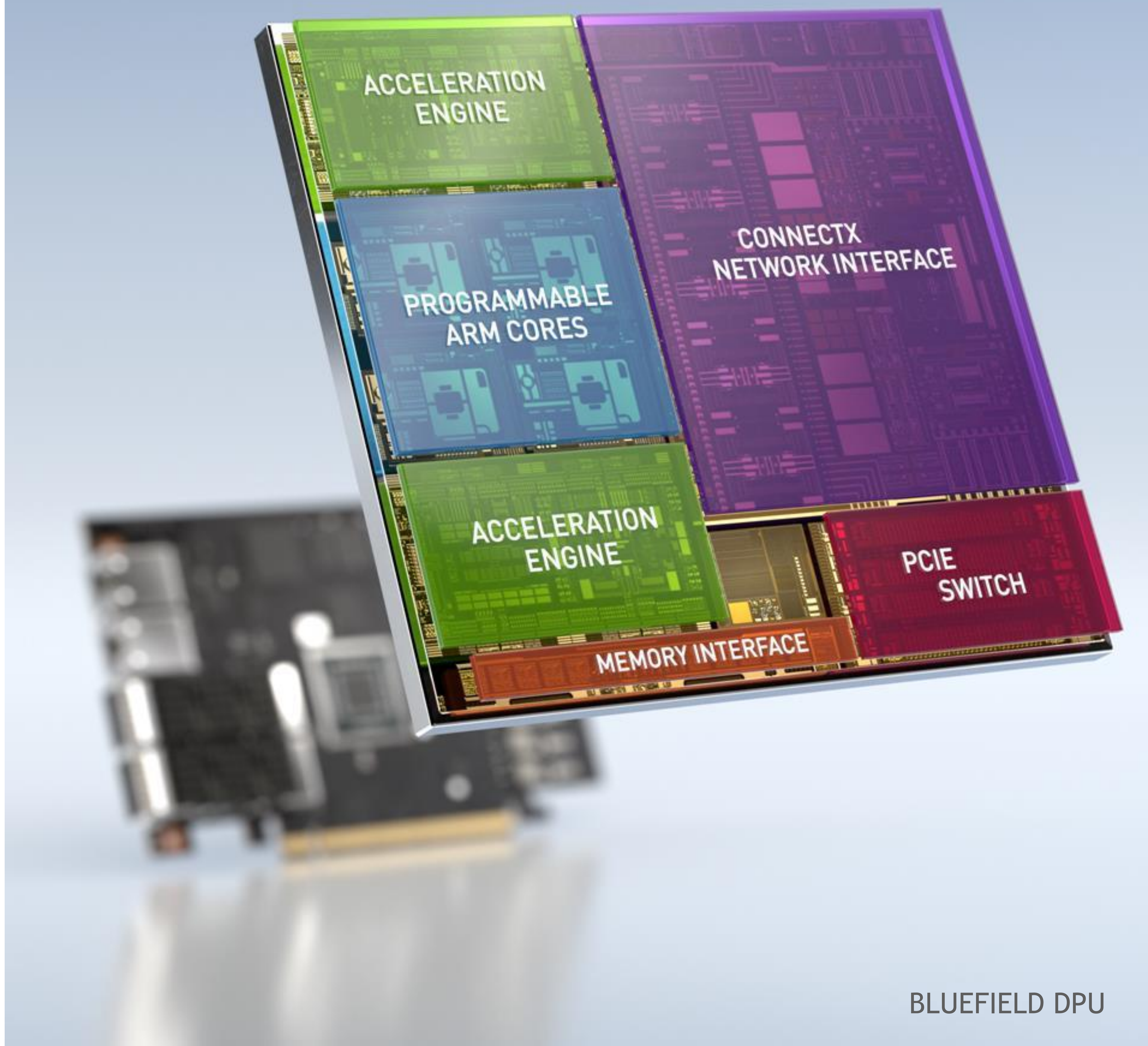
Higher Application Performance

From the Edge to the Main Data Center

BARE-METAL PERFORMANCE

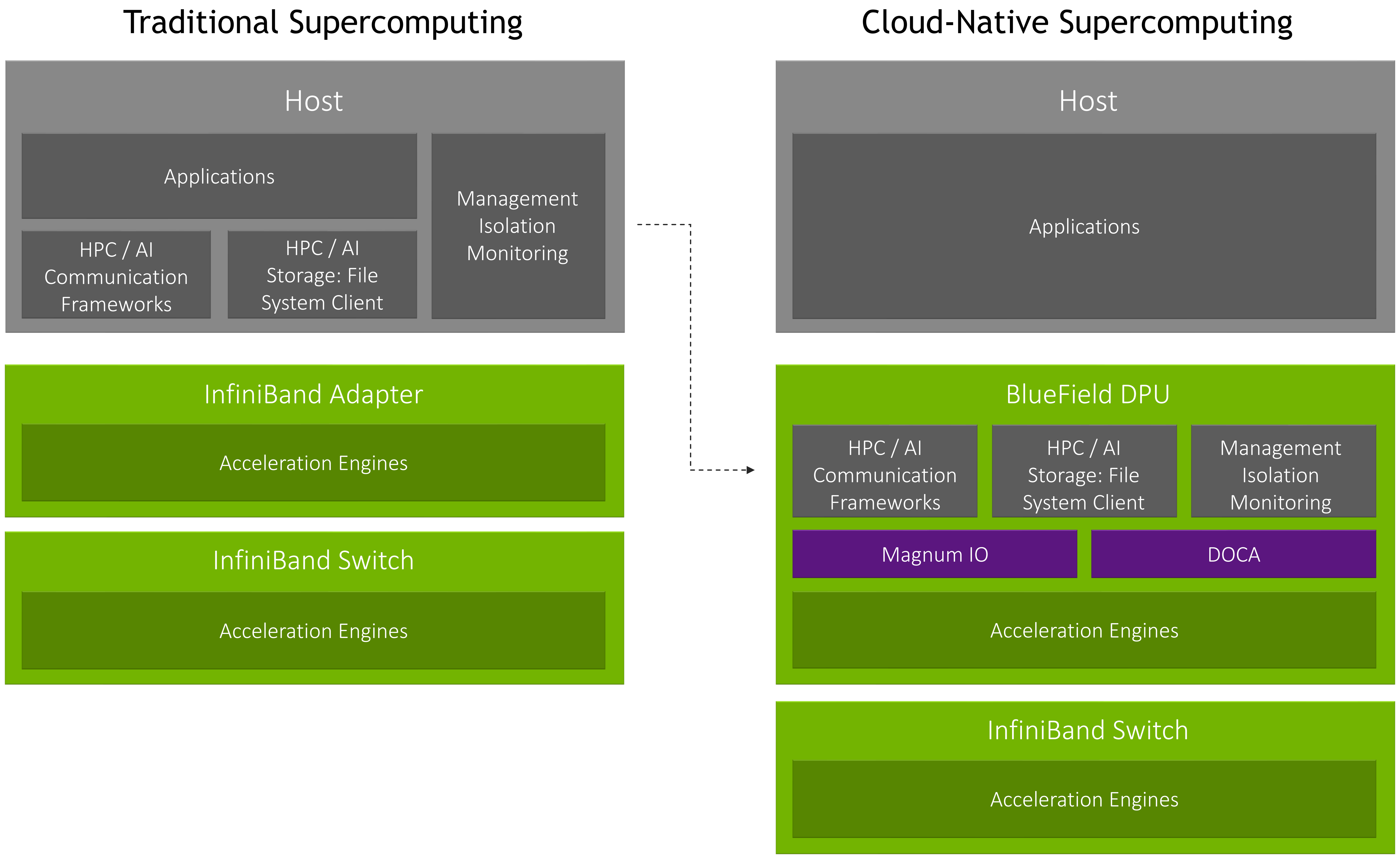
MULTI TENANCY

EDGE COMPUTING



BLUEFIELD DPU

CLOUD-NATIVE SUPERCOMPUTING INFRASTRUCTURE



MULTI-TENANT ISOLATION

Zero-Trust Architecture

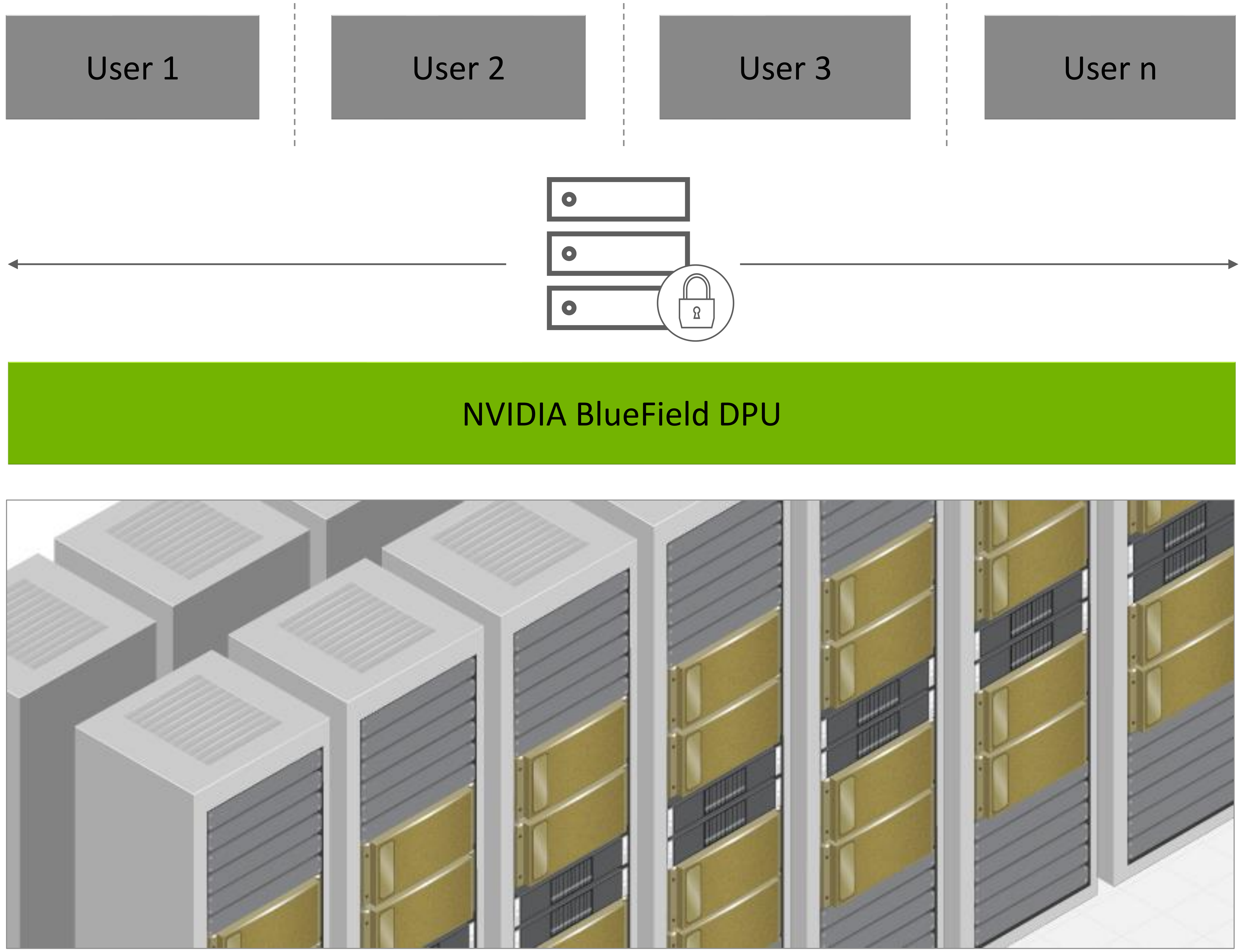
Secured Network Infrastructure and Configuration

Storage Virtualization

Tenant Service Level Agreement (SLA)

32K Concurrent Isolated Users on Single Subnet

Secure Partitioning with Bare-Metal Performance



HIGHER APPLICATION PERFORMANCE

DPU-Accelerated HPC Communications

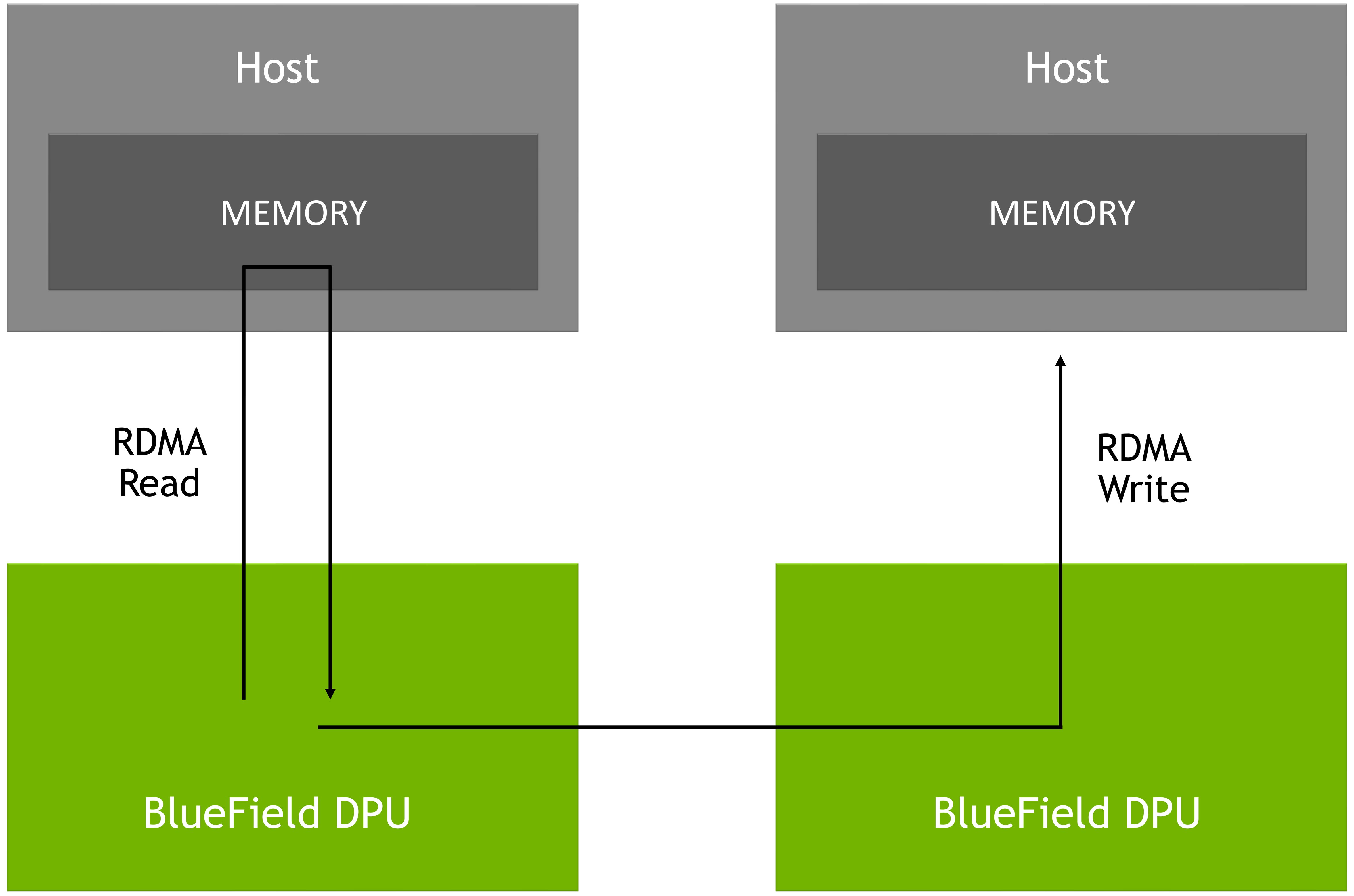
Collective Offloads

Active Messages

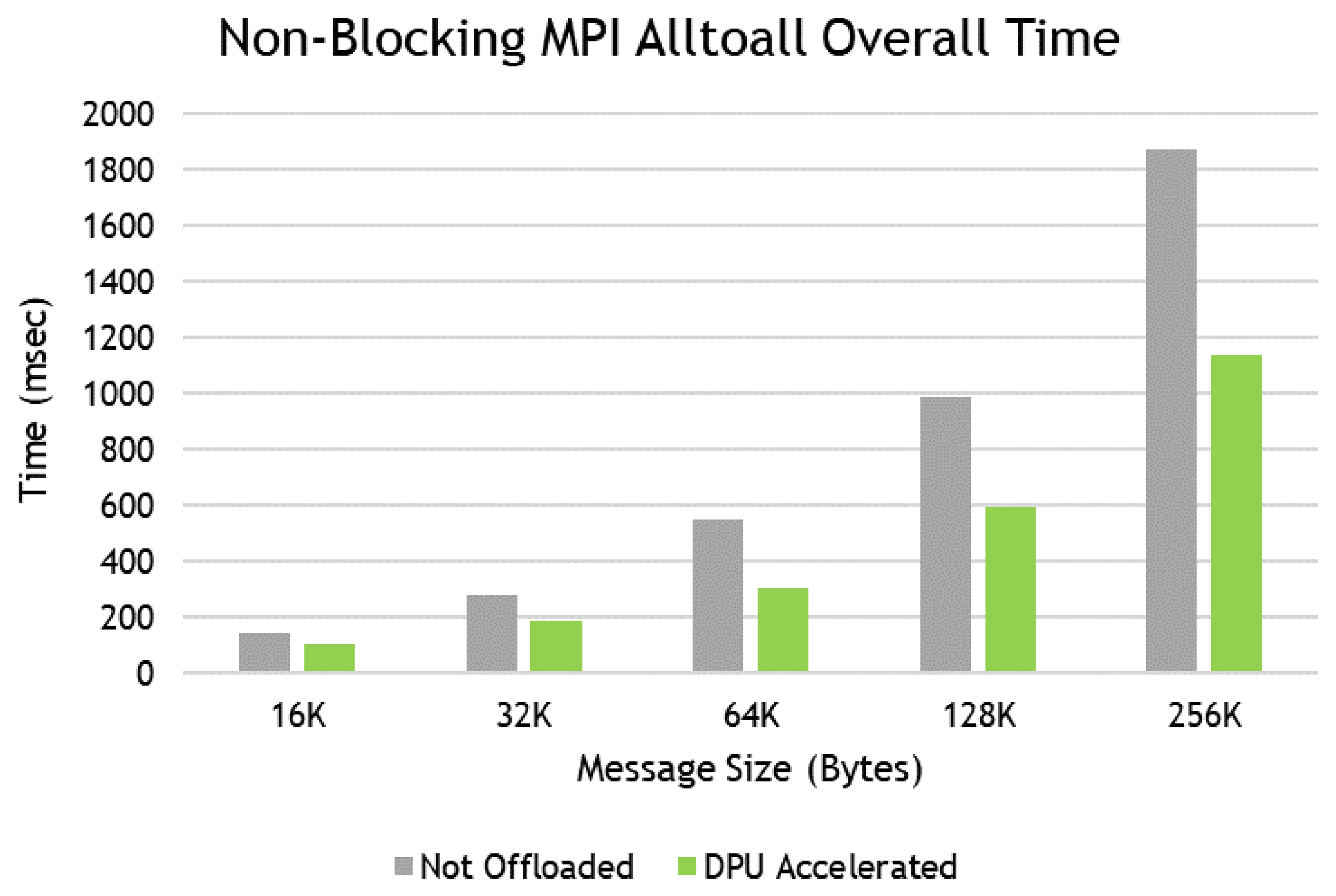
Smart MPI Progression

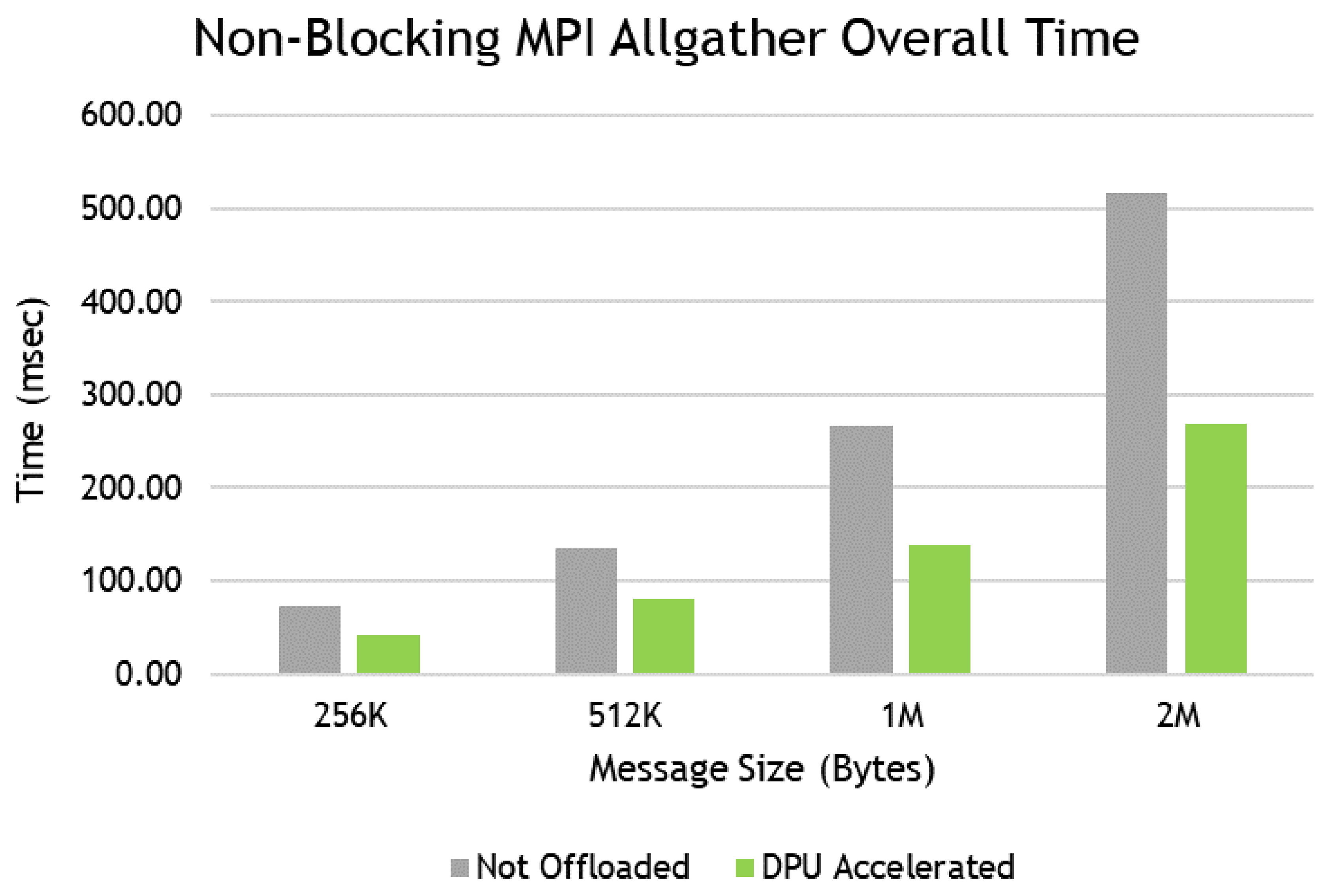
Data Compression

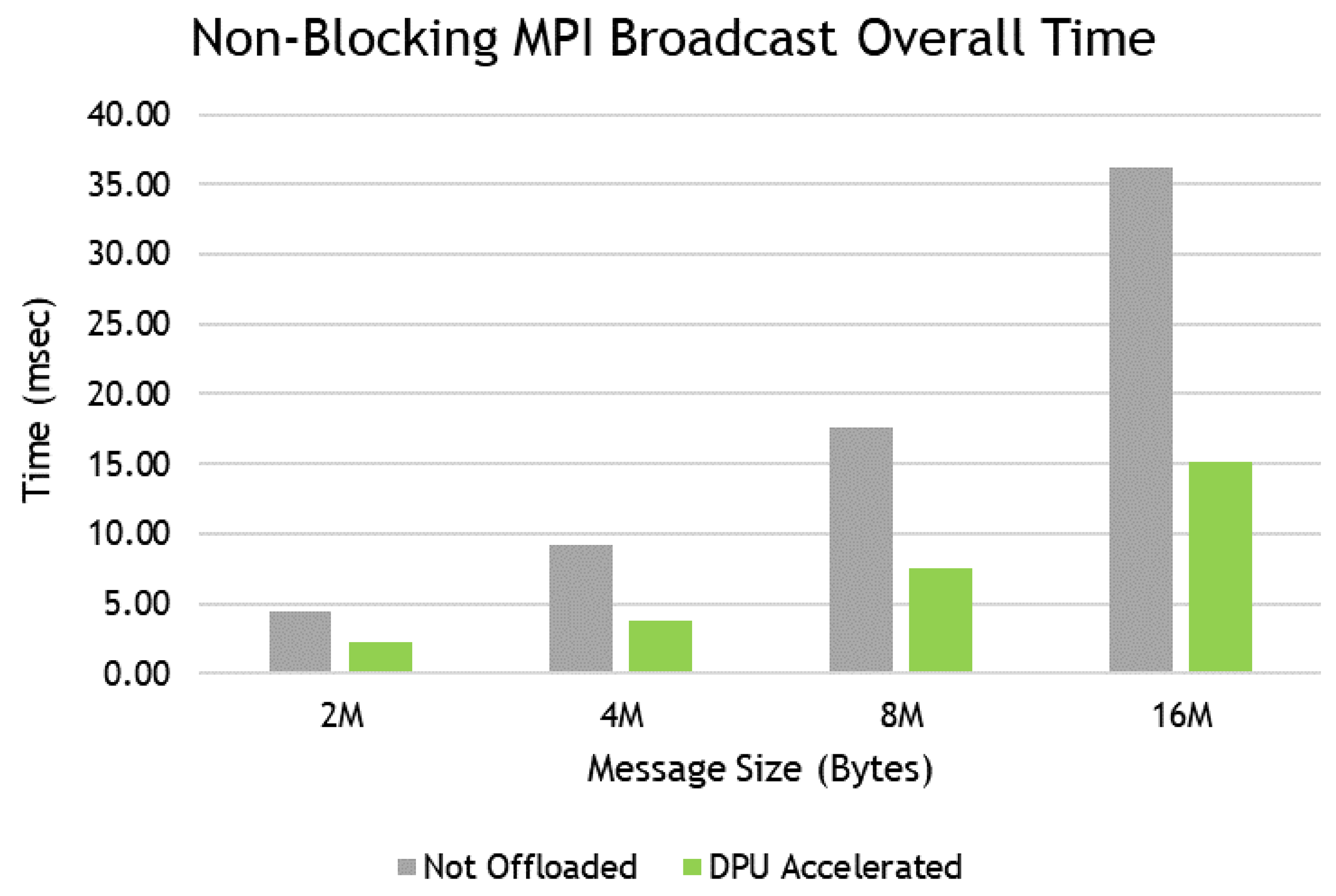
User-defined Algorithms



NON-BLOCKING MPI PERFORMANCE





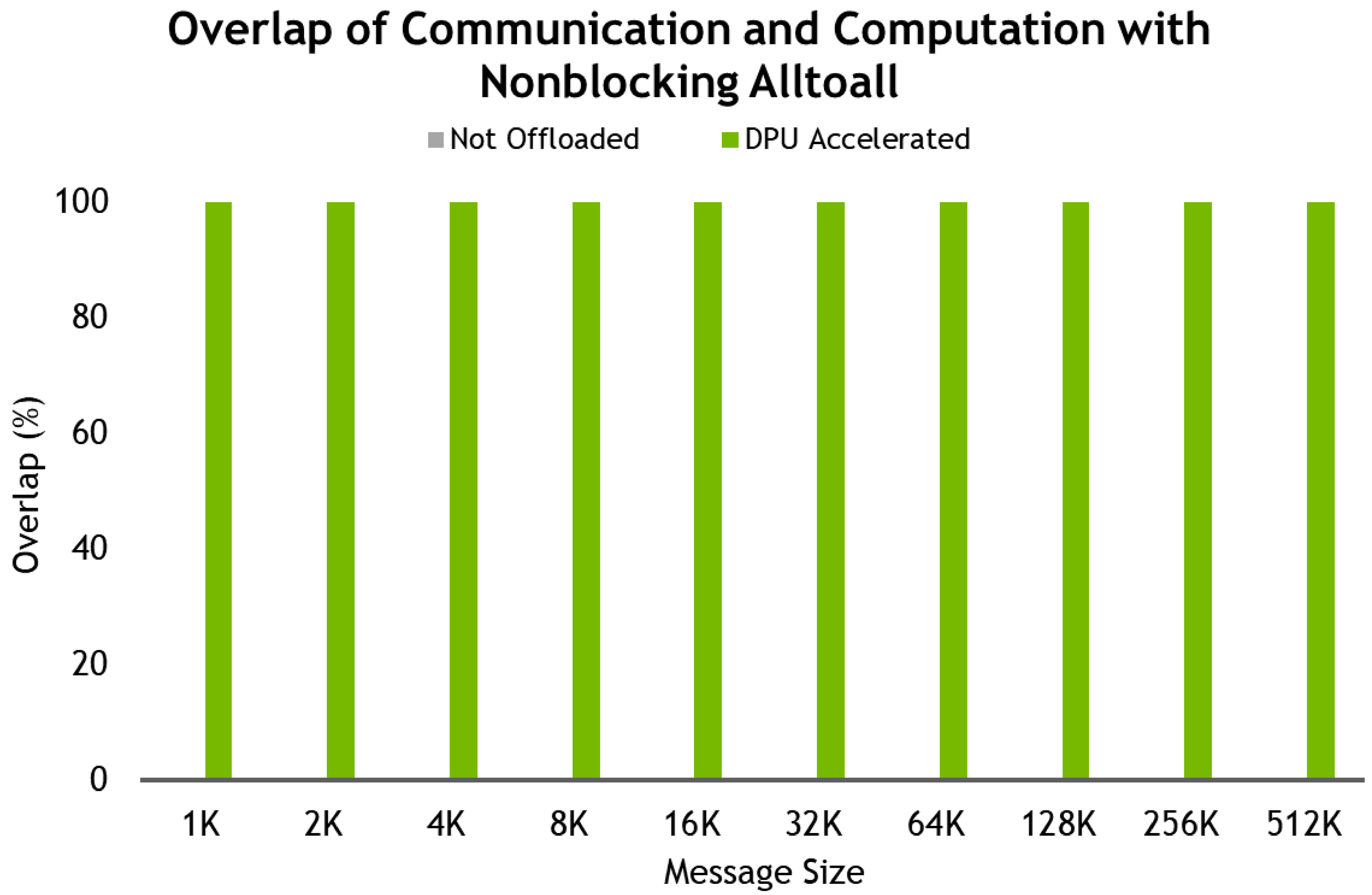


Courtesy of: Ohio State University MVAPICH Team and X-ScaleSolutions

MVAPICH X-ScaleSolutions

HIGHER APPLICATION PERFORMANCE

100% Communication – Computation Overlap



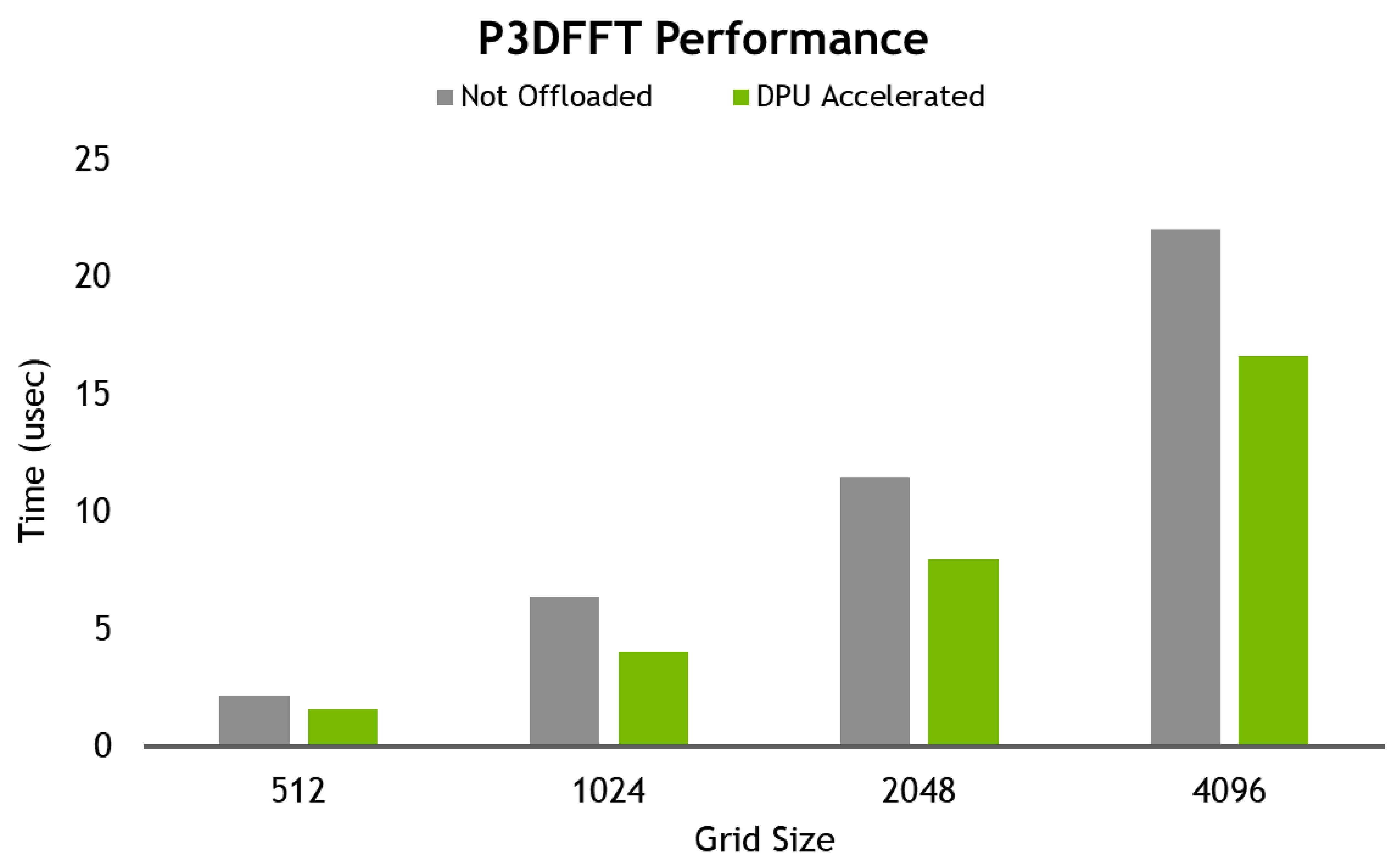
Courtesy of: Ohio State University MVAPICH Team and X-ScaleSolutions

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.

NVIDIA DOCA

Enabling Broad DPU Partner Ecosystem

Software Application Framework for BlueField DPUs

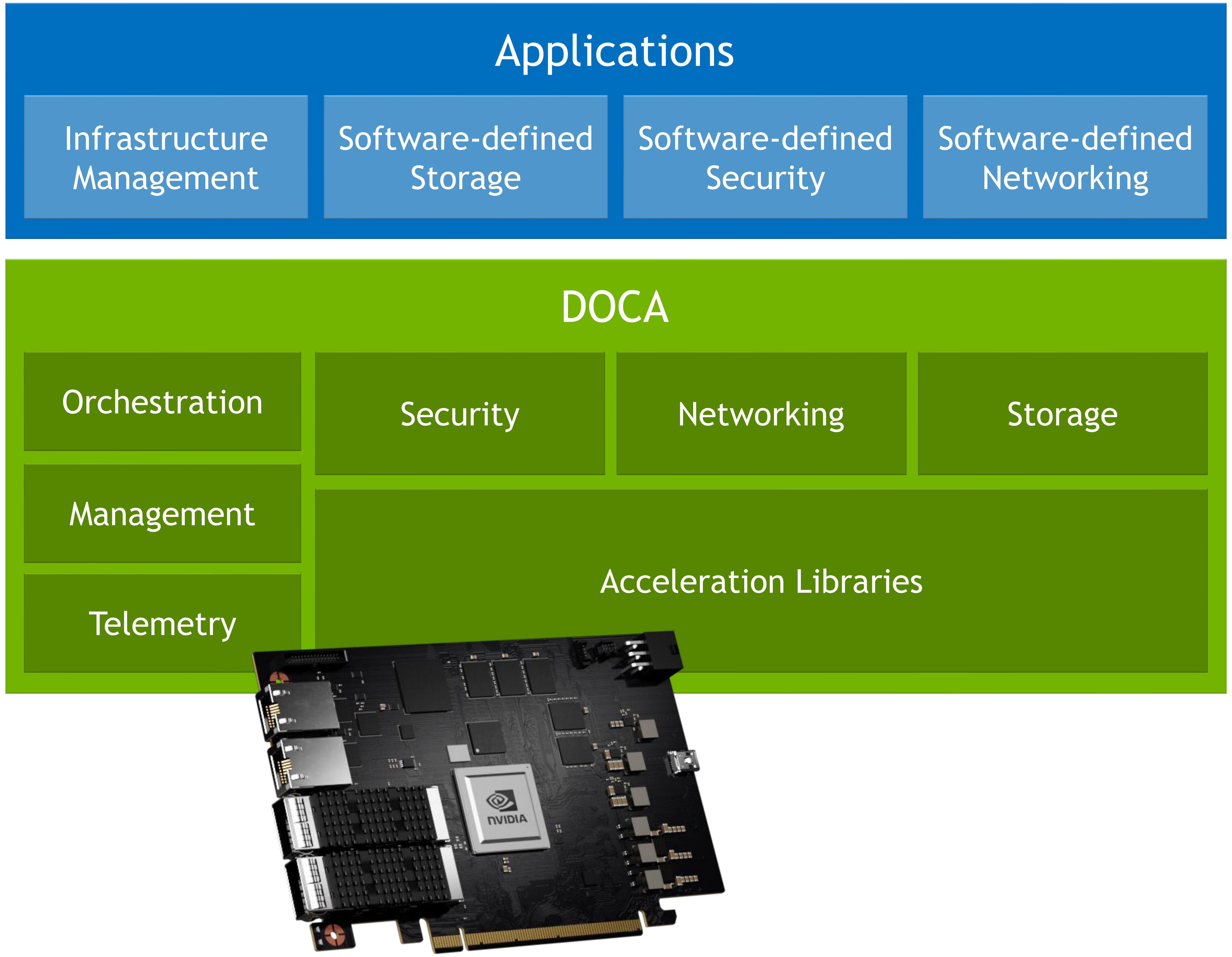
DOCA is for DPUs What CUDA is for GPUs

Protects Developer Investment for Future DPUs

Certified Reference Applications, APIs & Partner Solutions

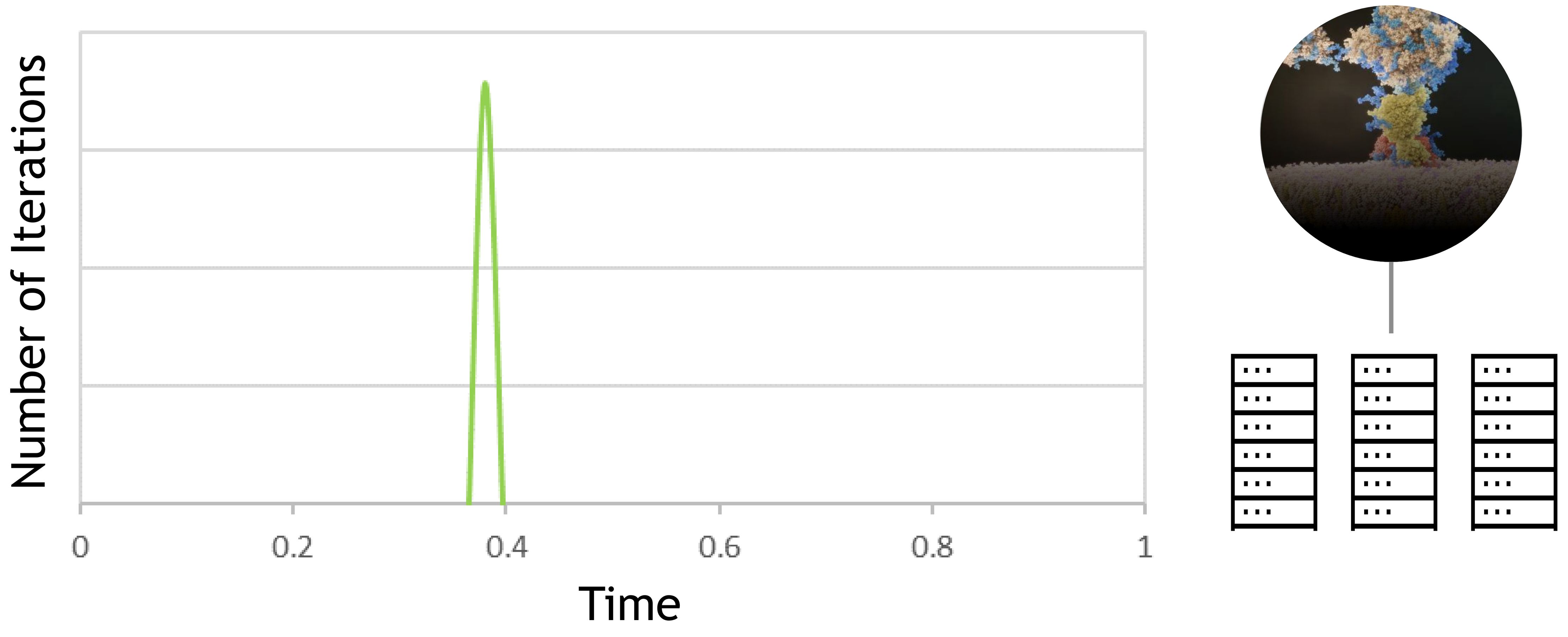
Rich Partner Ecosystem Across Industries and Workloads

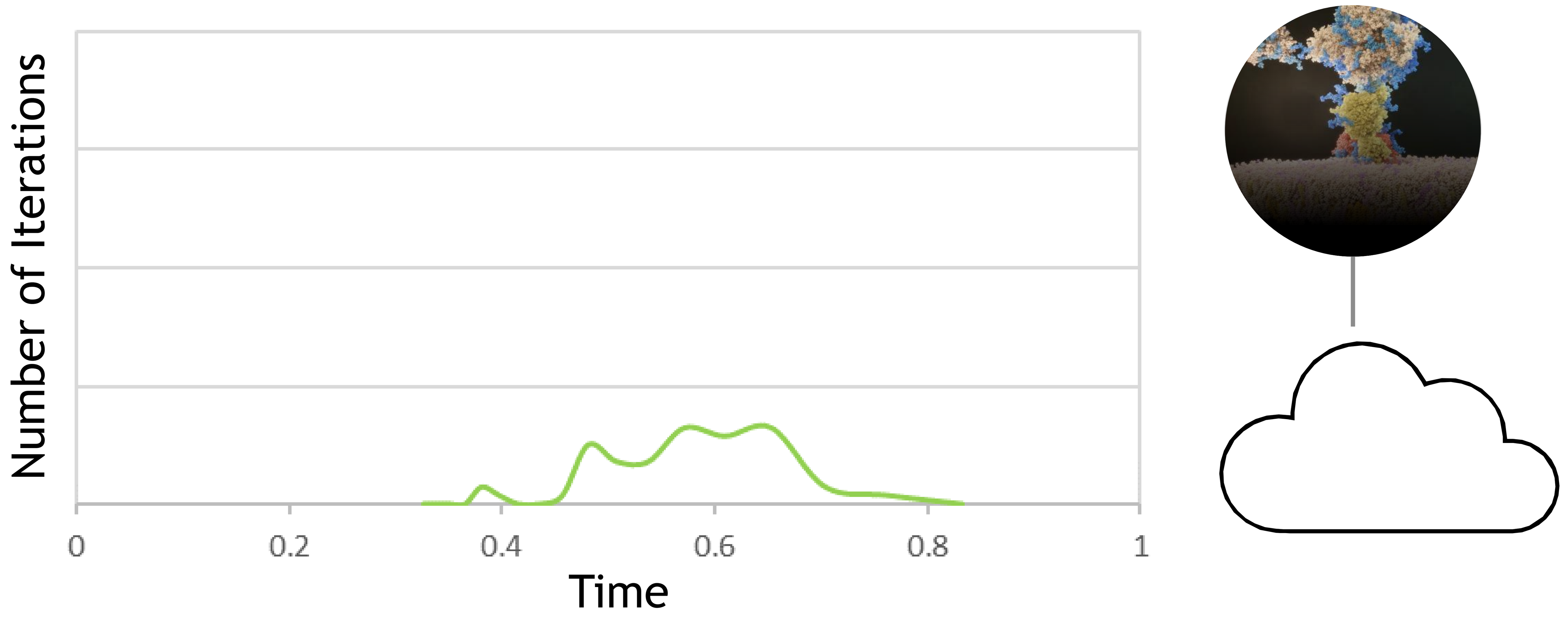
https://developer.nvidia.com/networking/doca



MULTI-TENANT SUPERCOMPUTING CLOUD — PERFORMANCE ISOLATION

Molecular Dynamics (LAMMPS) Example





HPC ON SUPERCOMPUTING

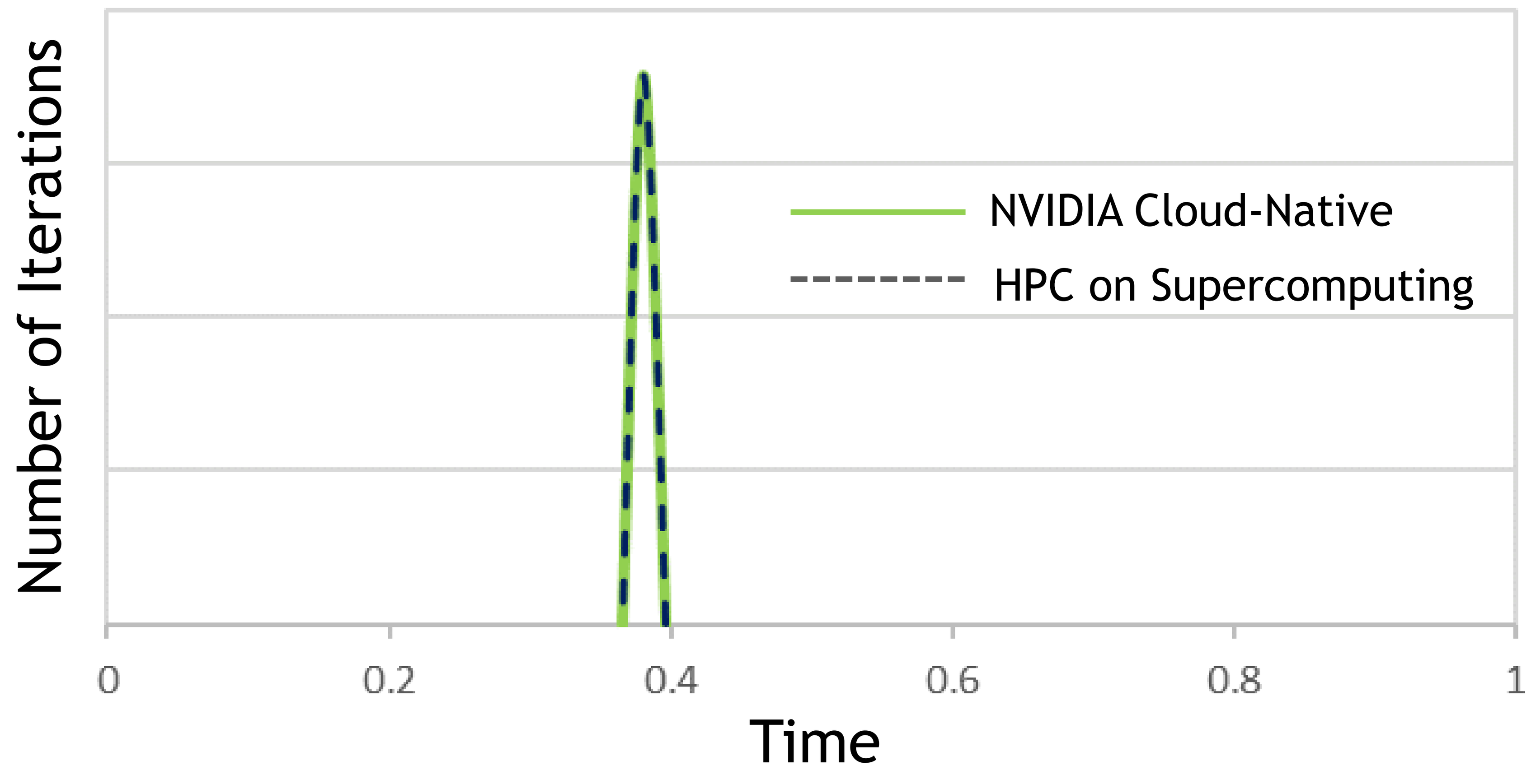
Molecular Dynamics (LAMMPS)

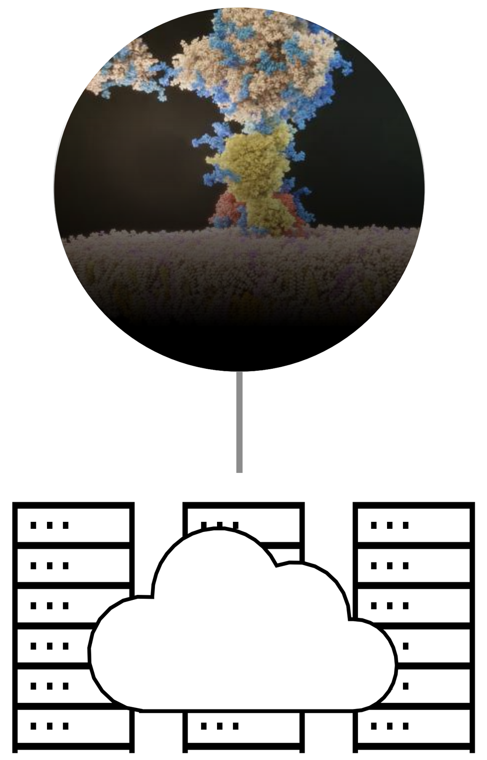
HPC ON THE CLOUD

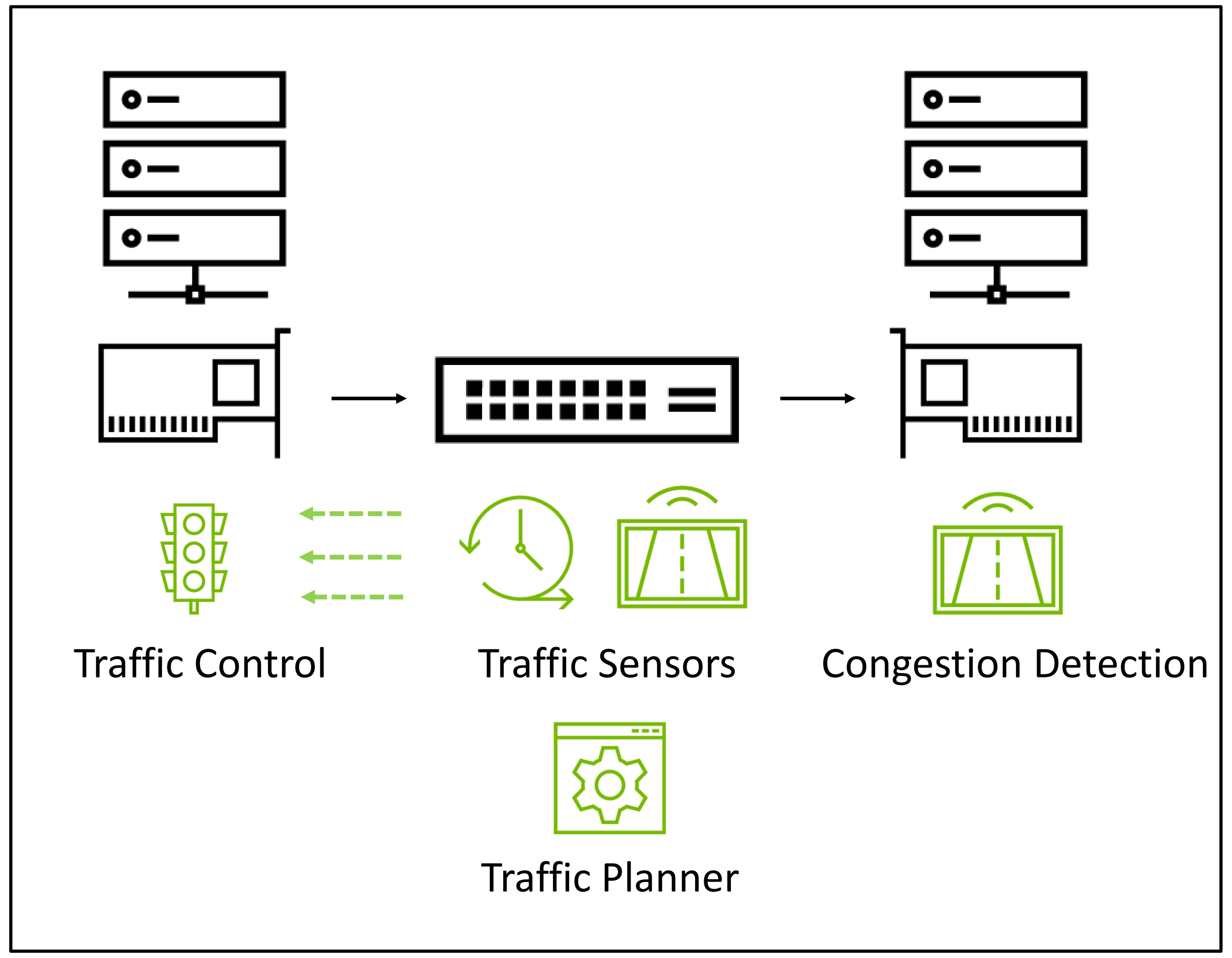
Molecular Dynamics (LAMMPS)

CLOUD NATIVE SUPERCOMPUTING PLATFORM

Performance Isolations via Telemetry Based Congestion Control







HPC ON CLOUD-NATIVE SUPERCOMPUTING

Molecular Dynamics (LAMMPS)

Proactive / Reactive

Telemetry Data
Time Sensors
Traffic Planners

PERFORMANCE ISOLATION - MICROSOFT AZURE

Quantum InfiniBand Congestion Control

HPC ON SUPERCOMPUTING

HPC ON THE CLOUD

HPC ON CLOUD-NATIVE SUPERCOMPUTING

Azure HPC/AI VM Series

Standard HPC VMs

Standard HPC Applications

High Compute/Memory + InfiniBand

*__HPC SKUs: H, HB, HC, HBv2, HBv3__

GPU VMs

Deep Learning, AI workloads

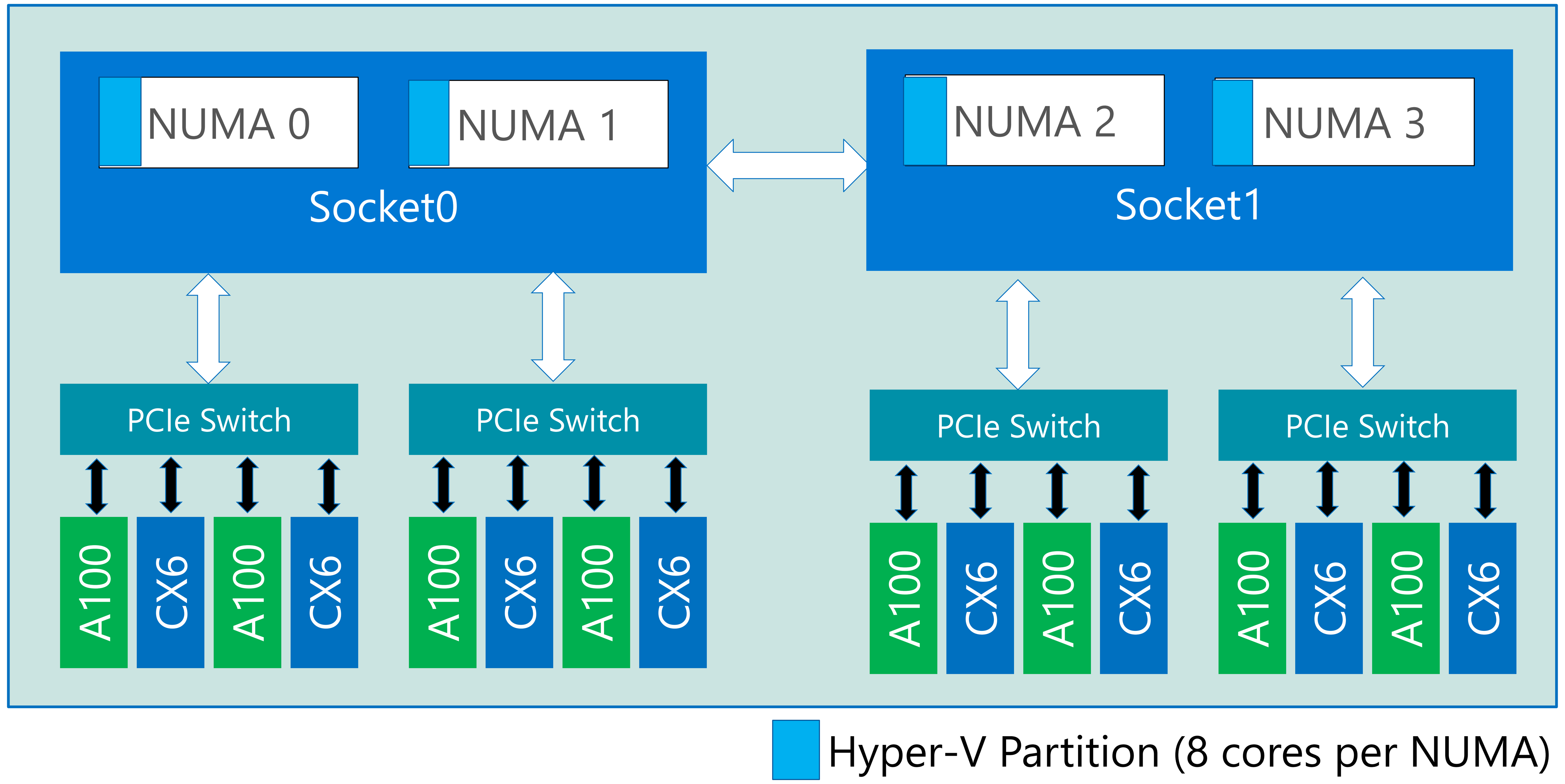
Visualization SKUs: NV series

*__Deep Learning/AI SKUs (InfiniBand): NC, ND series__

- "r" in VM type indicates RDMA support
- InfiniBand exposed to VMs using SR-IOV, offers full host bypass with full feature support
- *InfiniBand/RDMA enabled VMs: One VM per Host

Azure NDv4

- VM Specs:
  - AMD Rome (NPS=2)
  - VM Cores: 96 (48 per socket)
  - Memory: 900 GB
  - 8 x NVIDIA A100 GPUs
  - **8 x HDR 200Gbps InfiniBand**
  - Local Disk: 6.4 TB local NVMe SSD

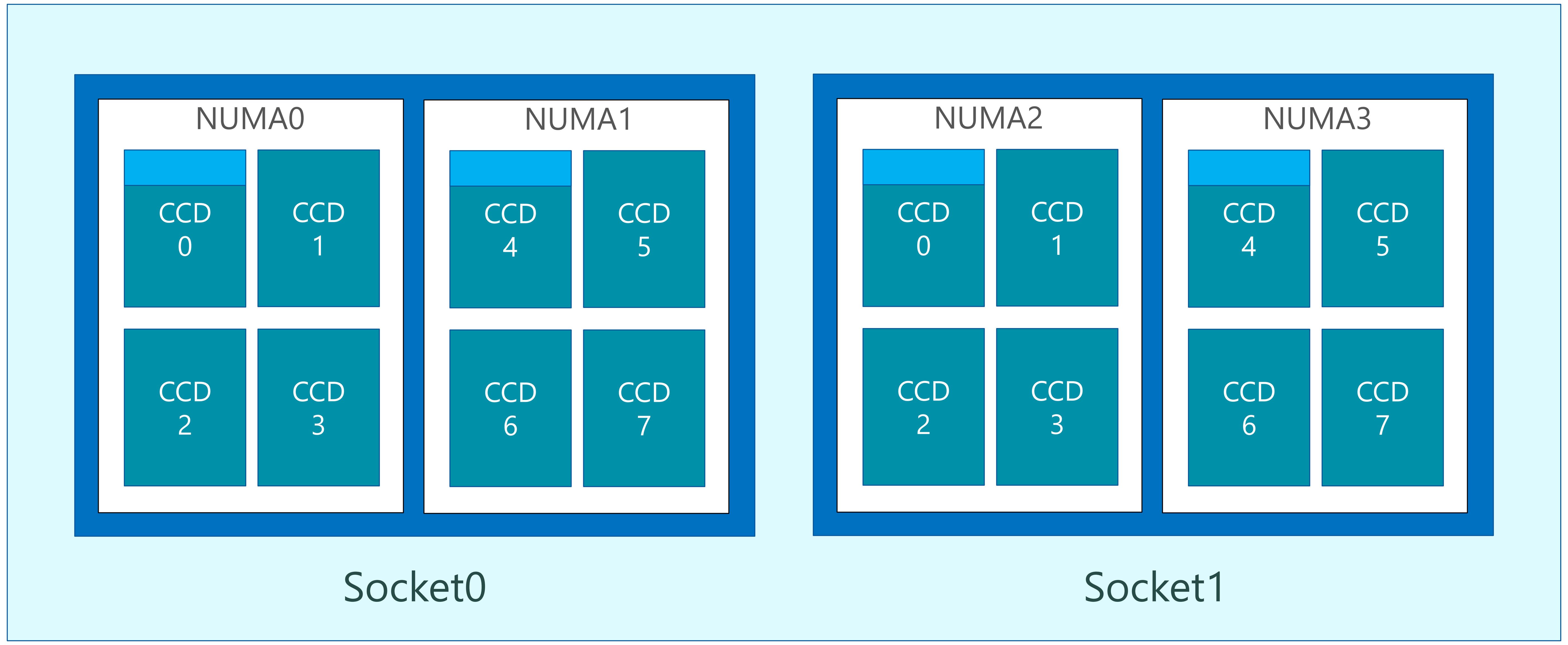


Standard_ND96asr_v4 (NDv4)

Azure HBv3 with Milan-X

AMD EPYC Milan

InfiniBand HDR 200Gbps



Hyper-V Partition (2 cores per NUMA)

- VM Specs:
  - AMD Milan-X (NPS = 2)
  - VM Cores: 120
  - Memory: 448 GB
  - Local Disk: 2 x 900 GB NVMe SSD
  - **Network: 200 Gbps HDR (SR-IOV)**

HBv3 VM Sizes (one VM per Host):

- Standard_HB120rs_v3 (all 120 cores)
- Standard_HB120-96rs_v3 (6 cores per CCD)
- Standard_HB120-64rs_v3 (4 cores per CCD)
- Standard_HB120-32rs_v3 (2 cores per CCD)
- Standard_HB120-16rs_v3 (1 cores per CCD)

Microsoft

InfiniBand Features in Azure

- **HB, HC, NDv2:** ConnectX-5
  - EDR 100 Gb/s InfiniBand
  - Up to 200 M messages/second
- **HBv2, HBv3, NDv4:** ConnectX-6
  - HDR 200 Gb/s InfiniBand
  - Up to 215 M messages/second

- **Dynamically Connected Transport (DCT)**
  - Reliable and scalable transport
  - Lesser Memory footprint
- **Hardware offload**
  - Collectives offload framework
  - Hardware tag matching
- **UD multicast (MCAST)**
  - Unreliable datagram (UD) based multicast
  - Create a mcast group and broadcast
- **SHARP**
  - Switch based collectives
- **Dynamic Routing**
  - Advanced Congestion Control
  - Adaptive Routing
- **Better Reliability**
  - SHIELD detects link failures and reroutes

GPUDirect RDMA

- Available on Azure NDv4
- Direct data path b/w A100 GPU and HDR200
- Each NIC/GPU pair gets peak b/w simultaneously over GPUDirect RDMA
- Combined GPUDirect RDMA b/w of **1.6 Tbps**
- Supports *all* GDR capable MPI libraries/middleware

```
hpcadmin@compute000000:~$ ./test_ib_gpu.sh  compute000000 compute000001 cpu
Pair 0:
 8388608     2922              0.00            196.09             0.002922--
 8388608     2920              0.00            195.96             0.002920
Pair 1:
 8388608     2928              0.00            196.49             0.002928
 8388608     2930              0.00            196.63             0.002930
Pair 2:
 8388608     2894              0.00            194.21             0.002894
 8388608     2896              0.00            194.34             0.002896
Pair 3:
 8388608     2883              0.00            193.47             0.002883--
 8388608     2881              0.00            193.34             0.002881
Pair 4:
 8388608     2893              0.00            194.14             0.002893
 8388608     2895              0.00            194.28             0.002895
Pair 5:
 8388608     2883              0.00            193.47             0.002883
 8388608     2885              0.00            193.61             0.002885
Pair 6:
 8388608     2922              0.00            196.09             0.002922--
 8388608     2920              0.00            195.96             0.002920
Pair 7:
 8388608     2913              0.00            195.48             0.002913
 8388608     2915              0.00            195.62             0.002915
```

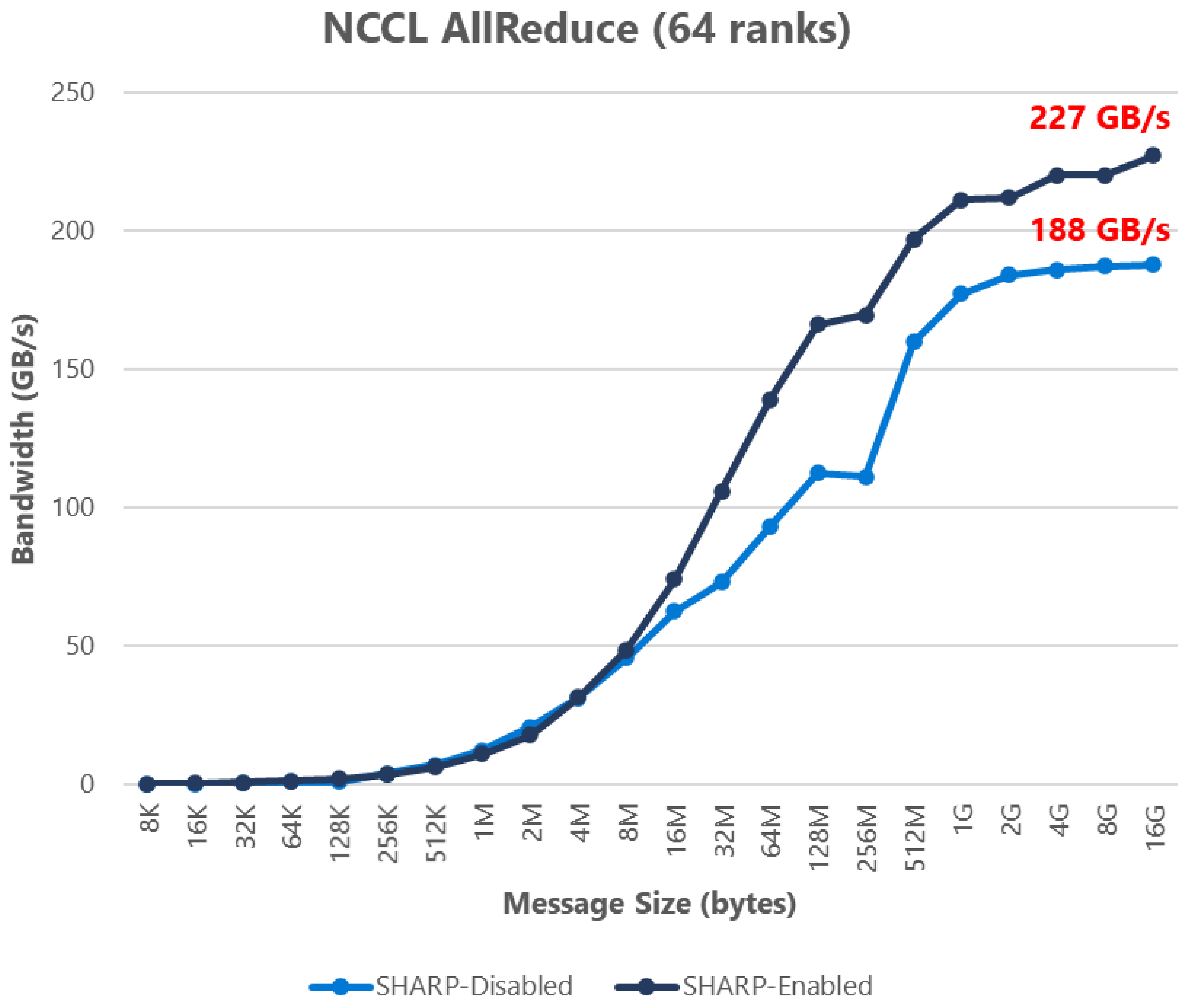
RDMA (Host Memory)

```
hpcadmin@compute000000:~$ ./test_ib_gpu.sh  compute000000 compute000001 gpu
Pair 0:
 8388608     2913              0.00            195.49             0.002913
 8388608     2913              0.00            195.49             0.002913
Pair 1:
 8388608     2914              0.00            195.55             0.002914
 8388608     2914              0.00            195.55             0.002914
Pair 2:
 8388608     2914              0.00            195.55             0.002914
 8388608     2914              0.00            195.55             0.002914
Pair 3:
 8388608     2915              0.00            195.62             0.002915
 8388608     2915              0.00            195.62             0.002915
Pair 4:
 8388608     2914              0.00            195.55             0.002914
 8388608     2914              0.00            195.55             0.002914
Pair 5:
 8388608     2915              0.00            195.62             0.002915
 8388608     2915              0.00            195.62             0.002915
Pair 6:
 8388608     2914              0.00            195.55             0.002914
 8388608     2914              0.00            195.55             0.002914
Pair 7:
 8388608     2915              0.00            195.62             0.002915
 8388608     2915              0.00            195.62             0.002915
hpcadmin@compute000000:~$
```

GPUDirectRDMA
(GPU Memory)

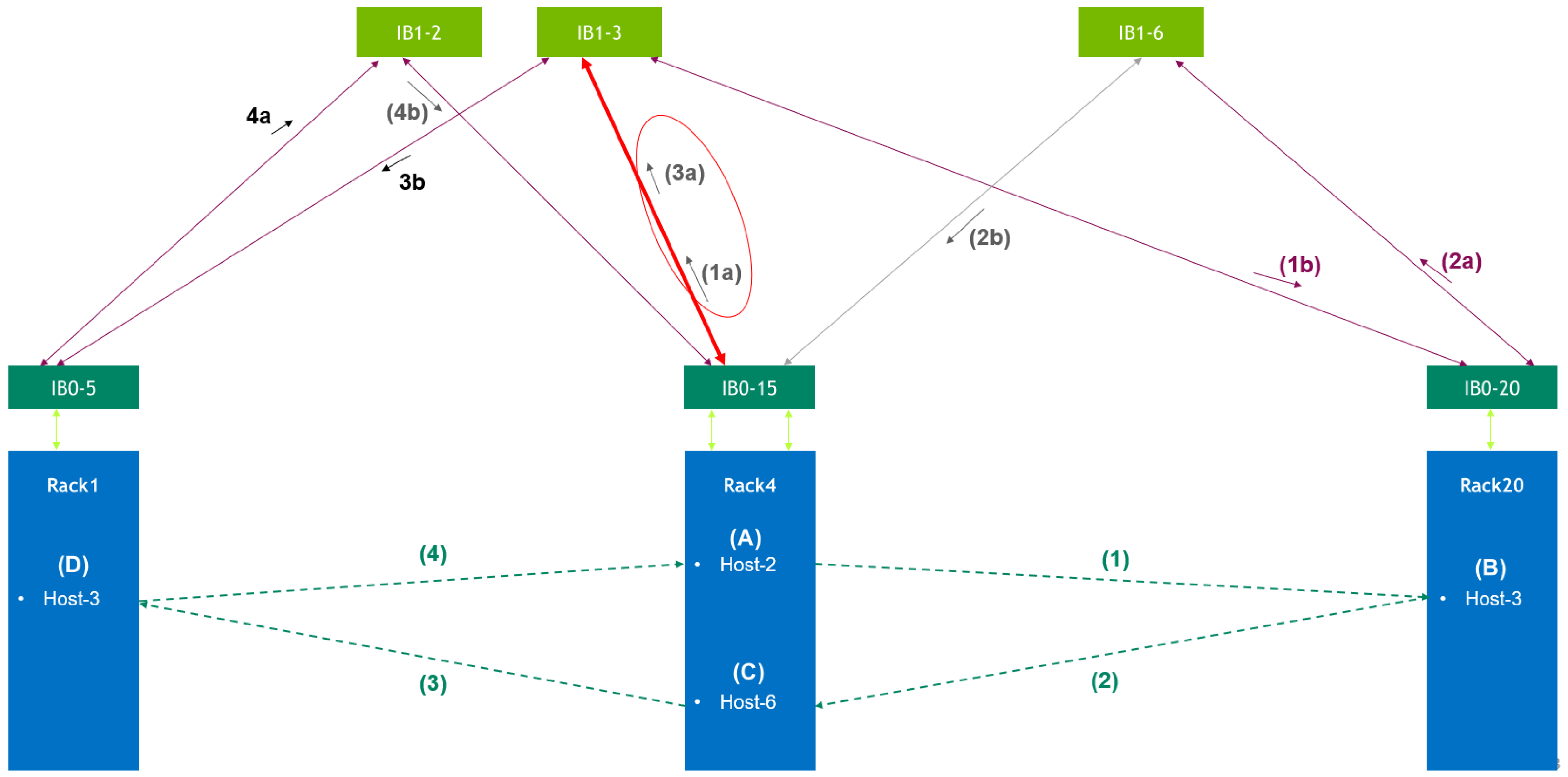
SHARP on NDv4

- Enabled on dedicated NDv4 clusters
- UCX-based Sharp-AM / SharpD communication
- Optimized SHARP tree initialization
- Connection keepalive
- GRH support

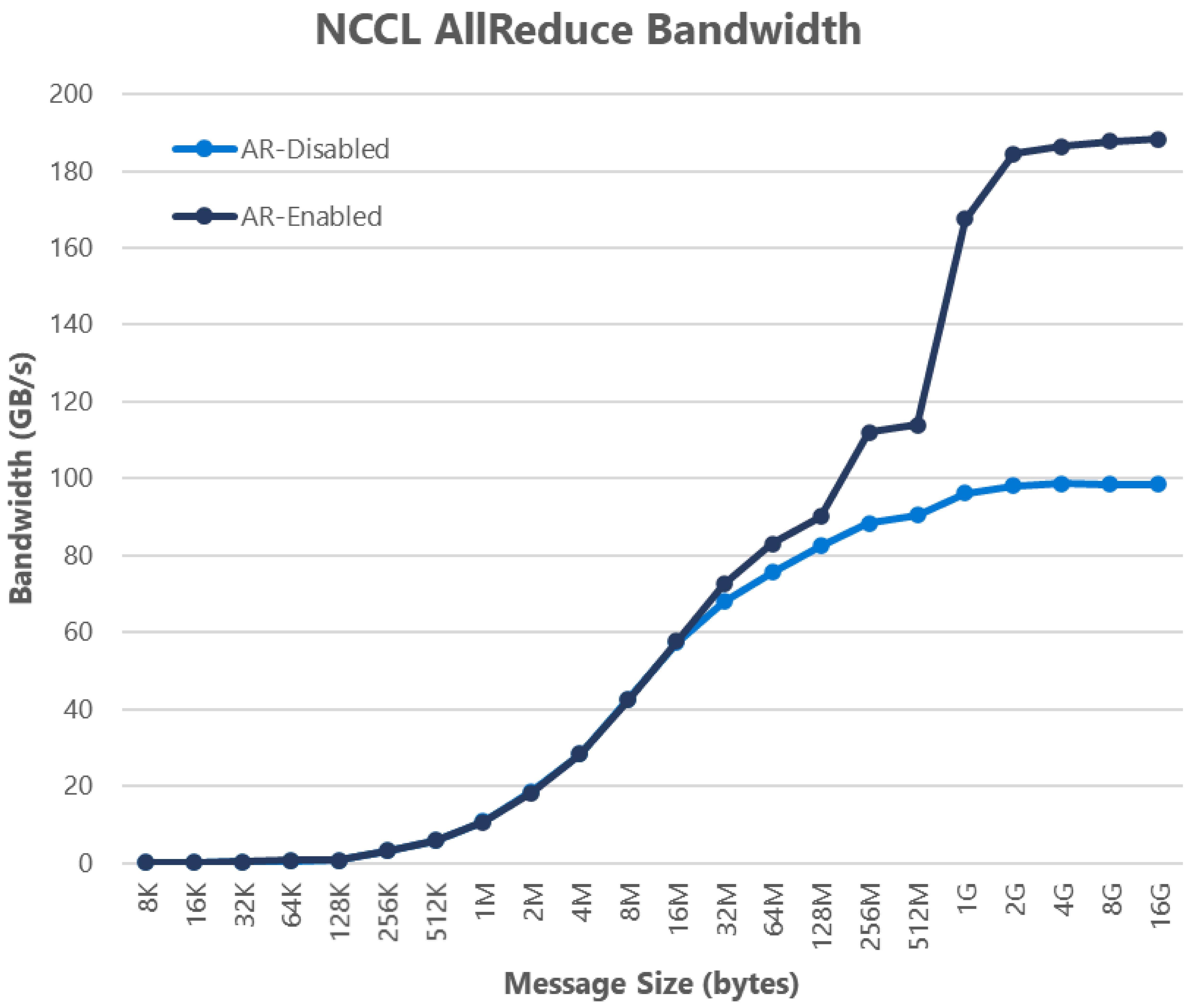


Microsoft

Adaptive Routing



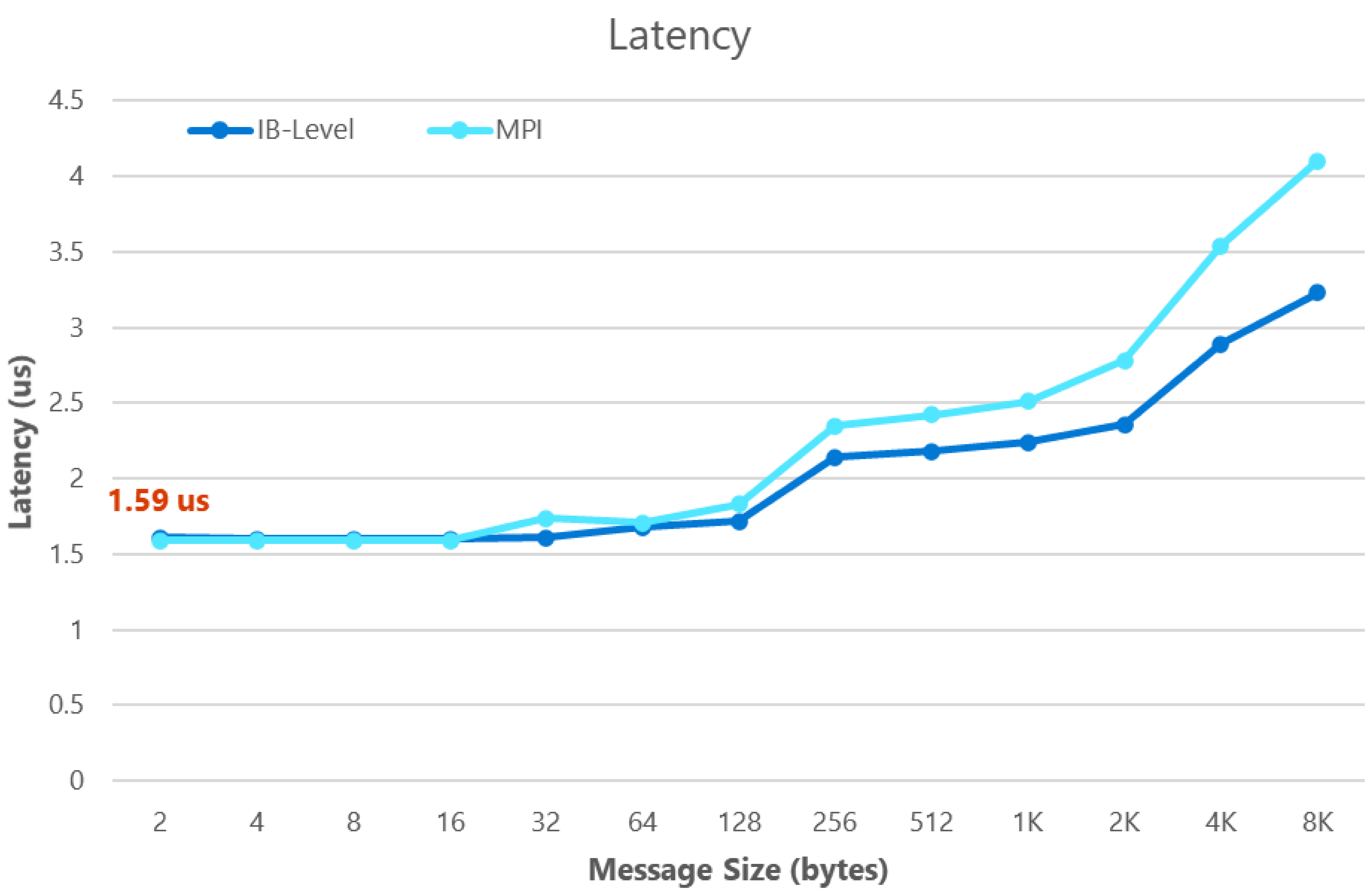
Communication paths during NCCL AllReduce

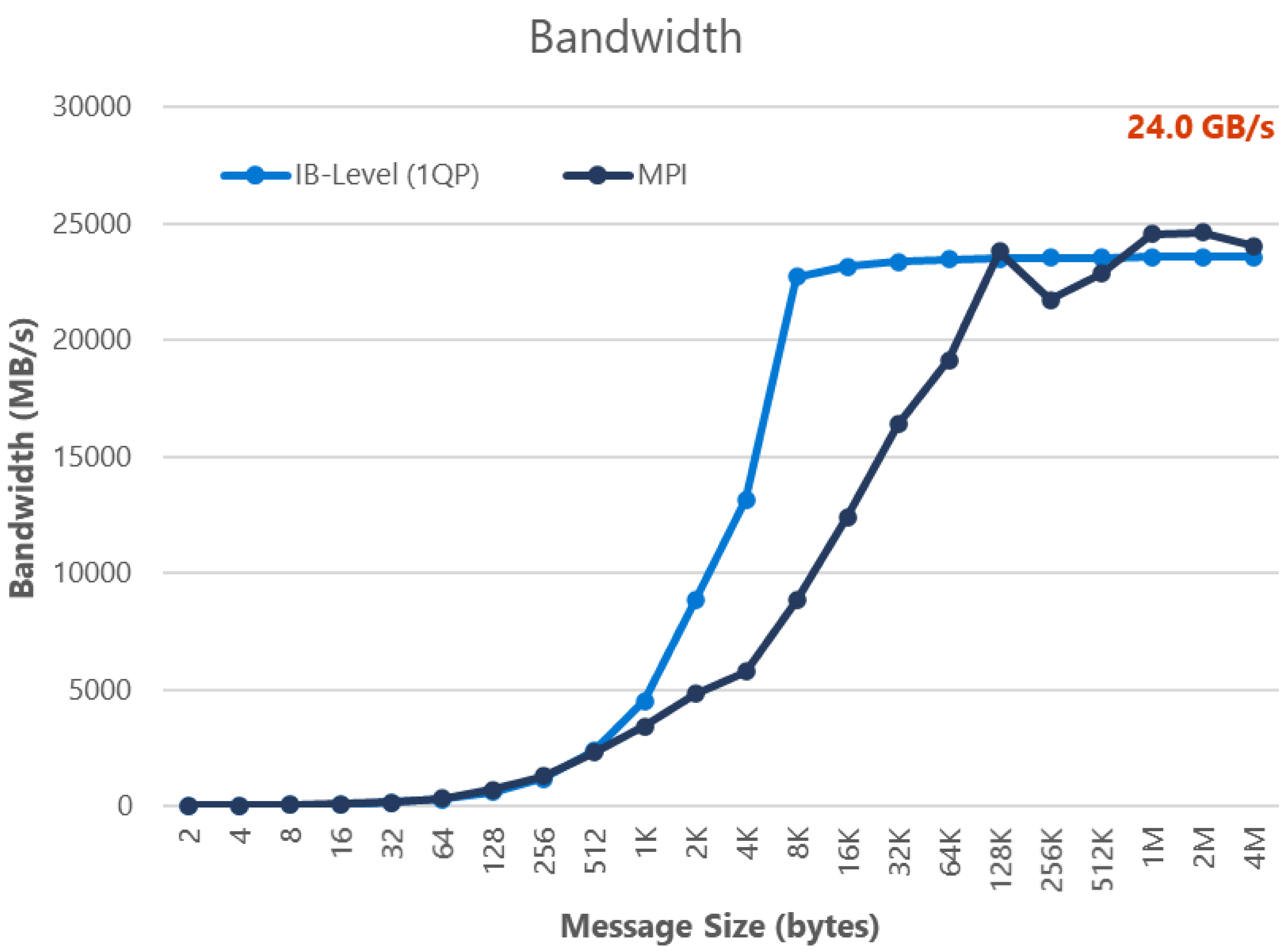


- Azure HPC InfiniBand Networks are non-blocking, and not oversubscribed
- Link contention can happen with static routing if a single link is being used by two or more communicating pairs
- Adaptive Routing allows packets to take different routes and offers stable performance

Microsoft

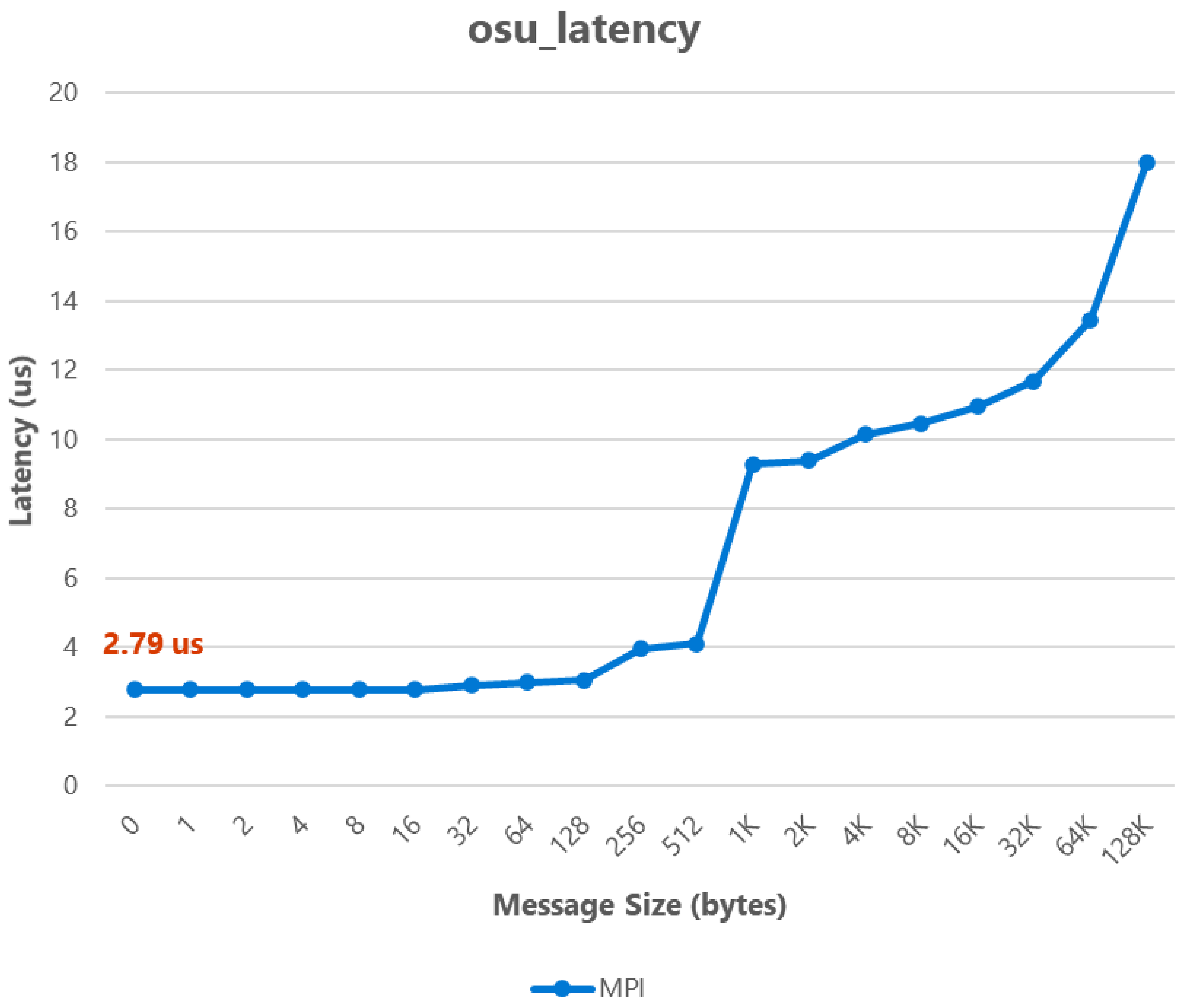
MPI Benchmarks on HBv3 (inter-node)

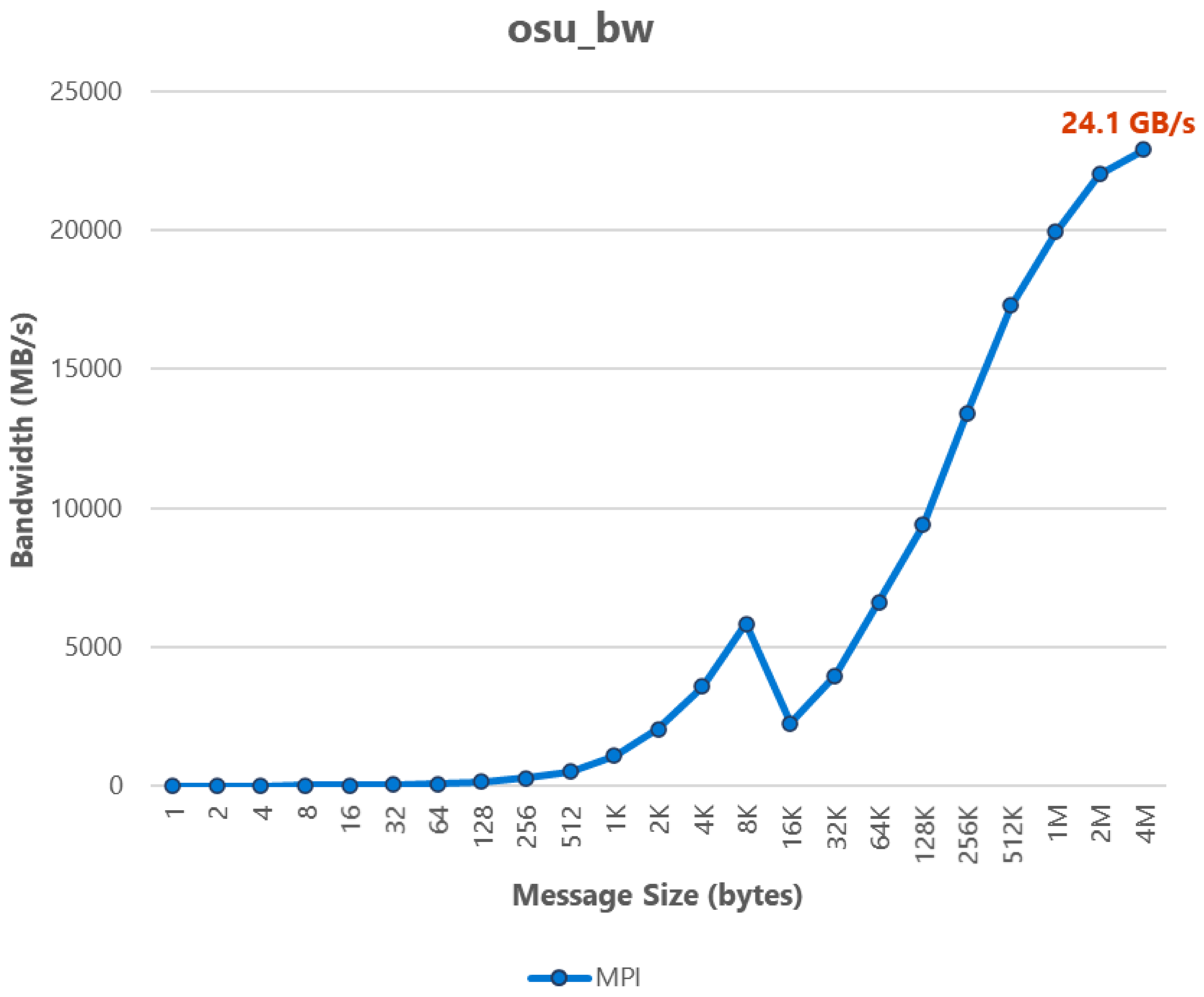




Microsoft

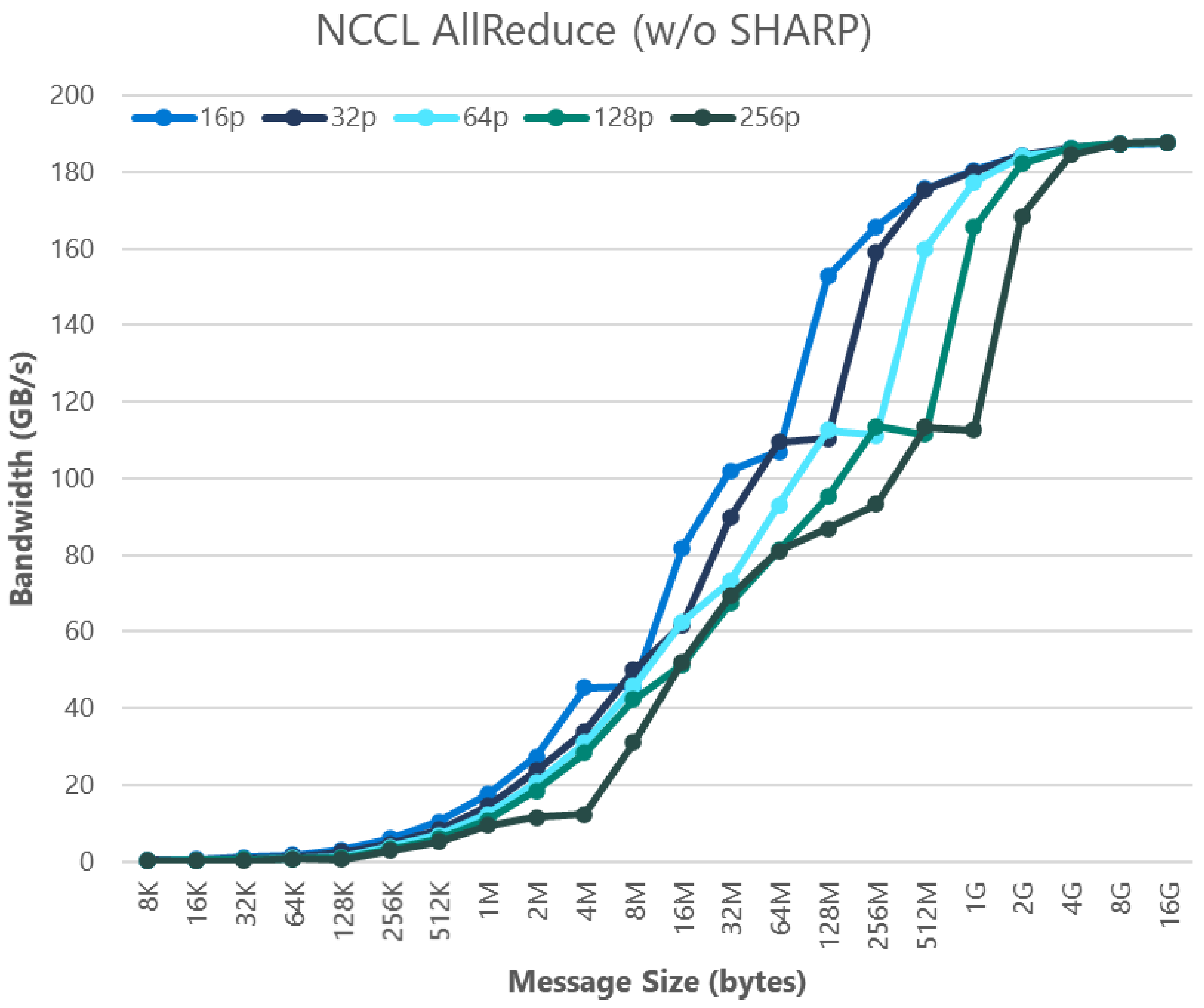
GPUDirect RDMA (MPI Level) on NDv4

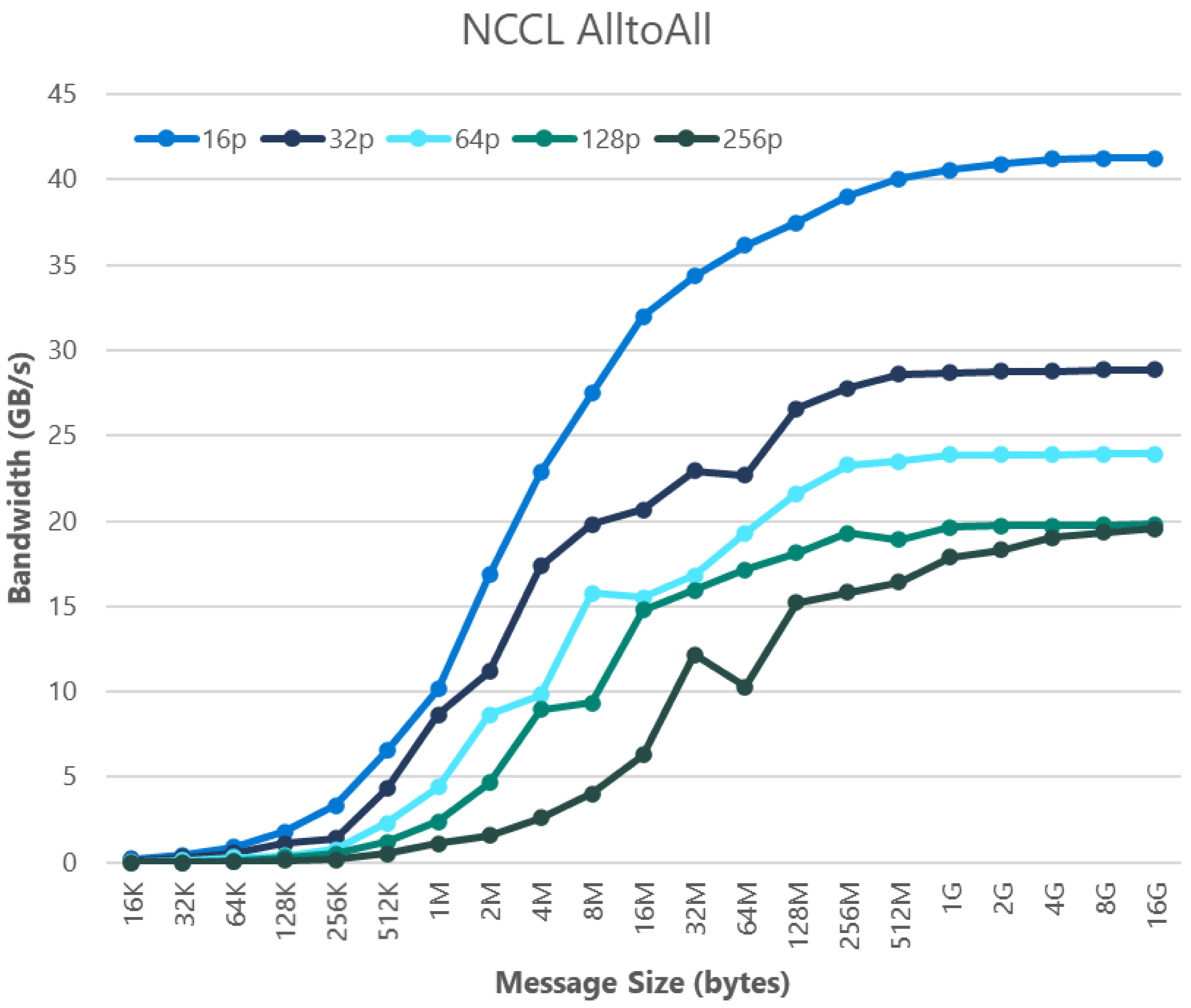




Microsoft

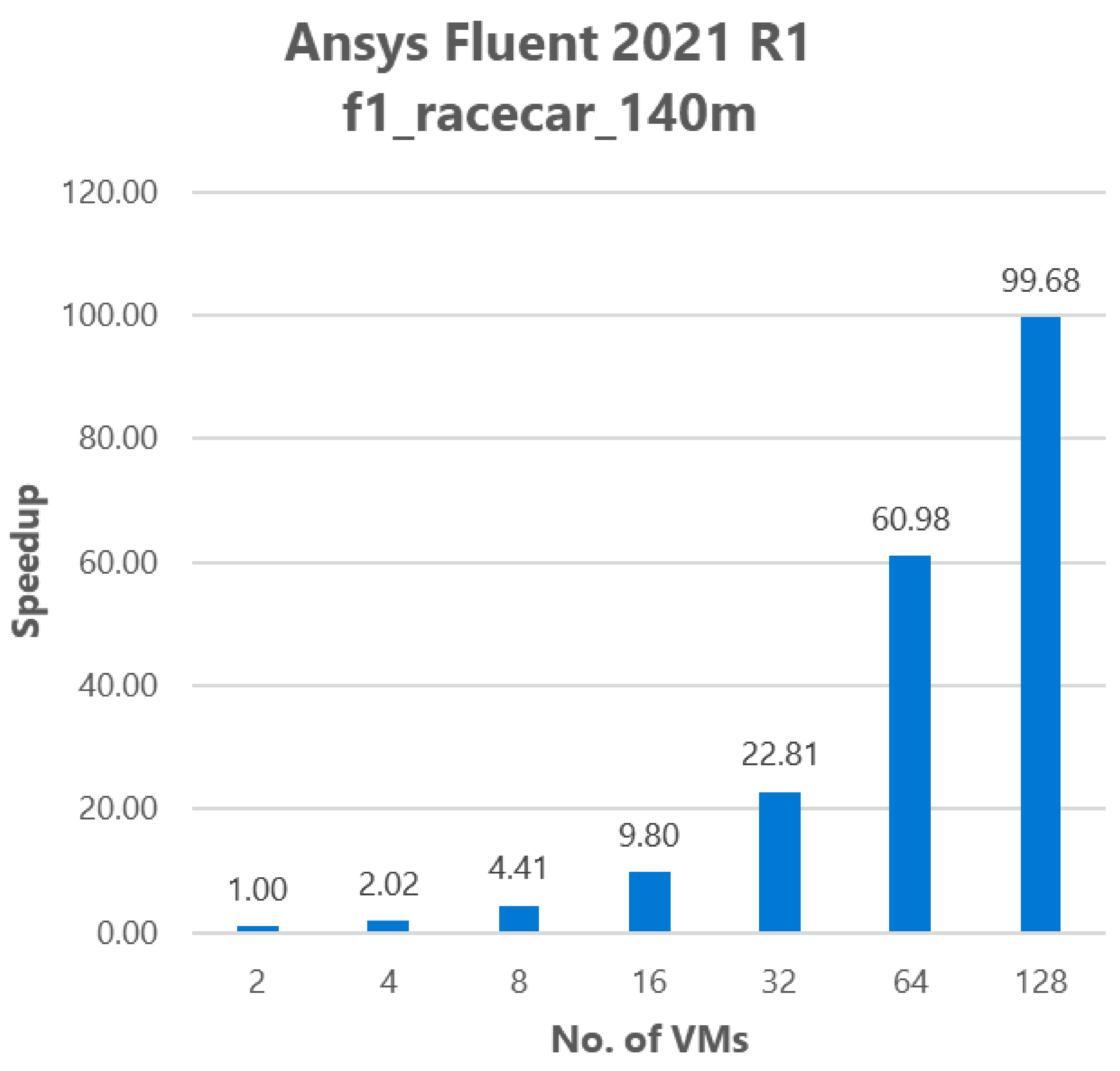
NCCL on NDv4



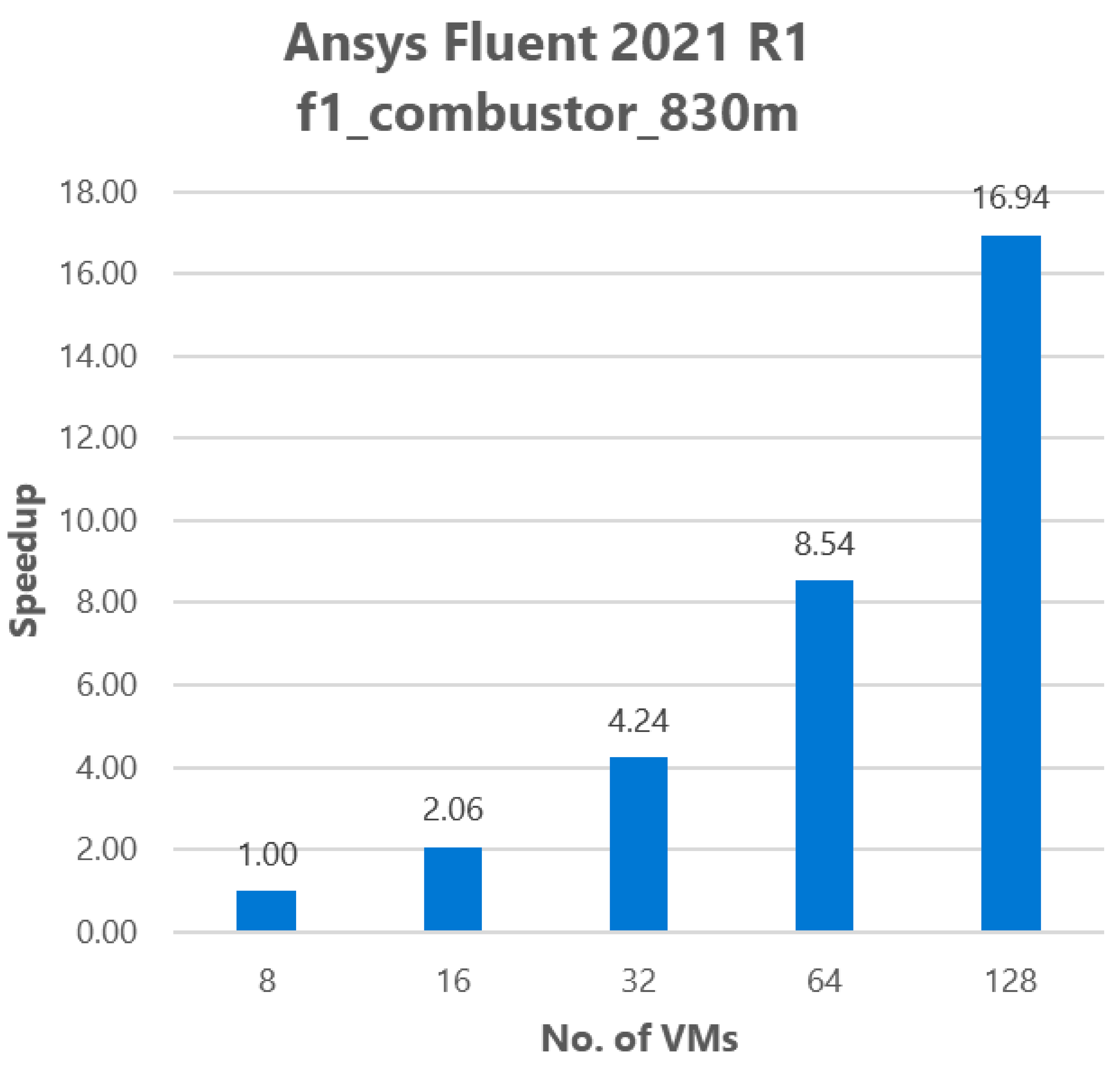


Microsoft

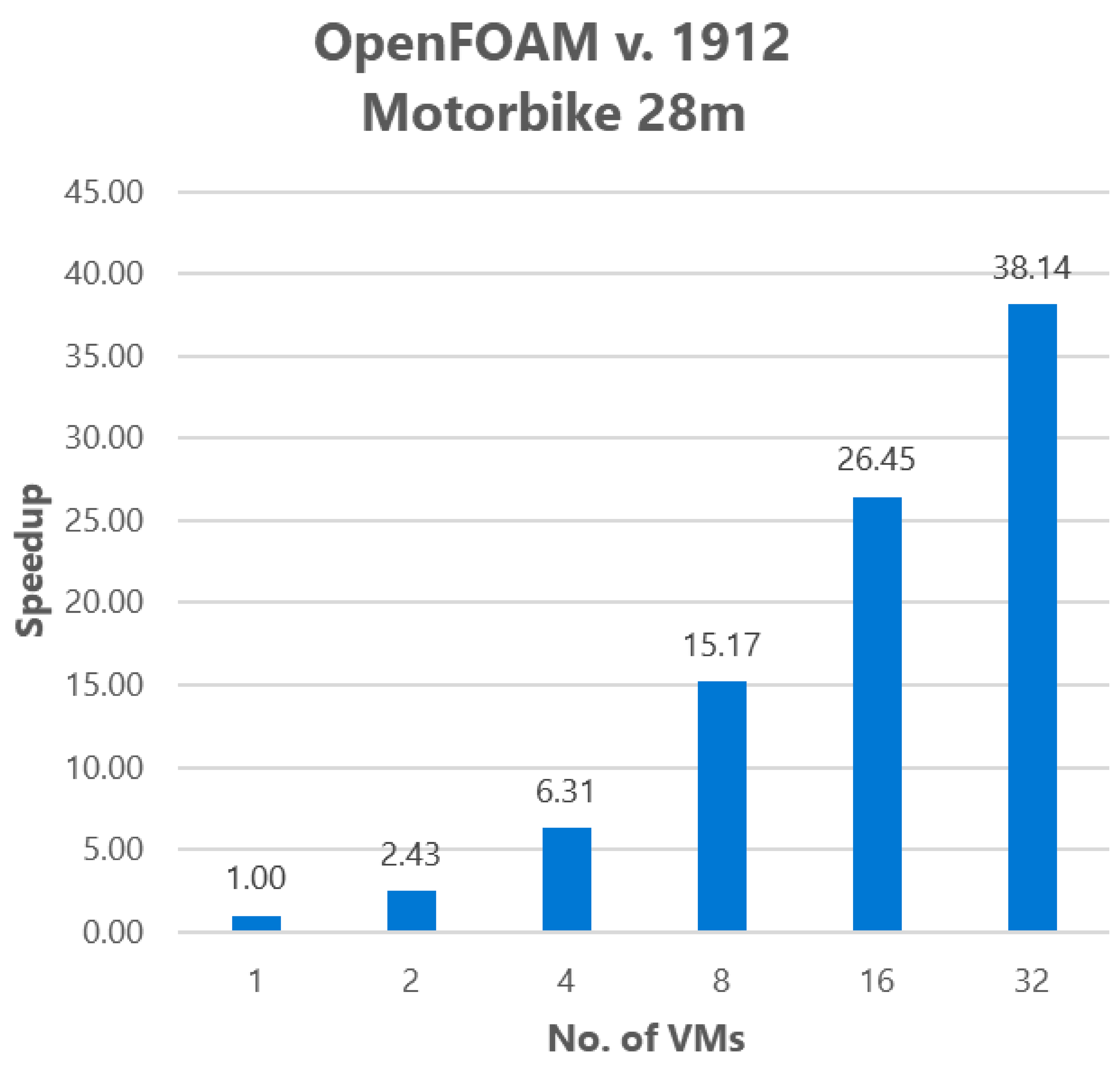
Scaling Efficiency on HBv3 (Milan-X)



156% scaling efficiency



106% scaling efficiency



119% scaling efficiency

https://aka.ms/MilanXPerf

Conclusion

- Supercomputer on Cloud is real!
- Azure HPC Cloud powered with InfiniBand in Top500, Graph500 top spots
    - Rank 10 in Top500 Nov. 2021
    - Rank 17 in Graph500 Nov. 2020
- Azure HPC democratizes Supercomputer!

Thank you!

jijos@microsoft.com

Microsoft


NVIDIA