Microsoft

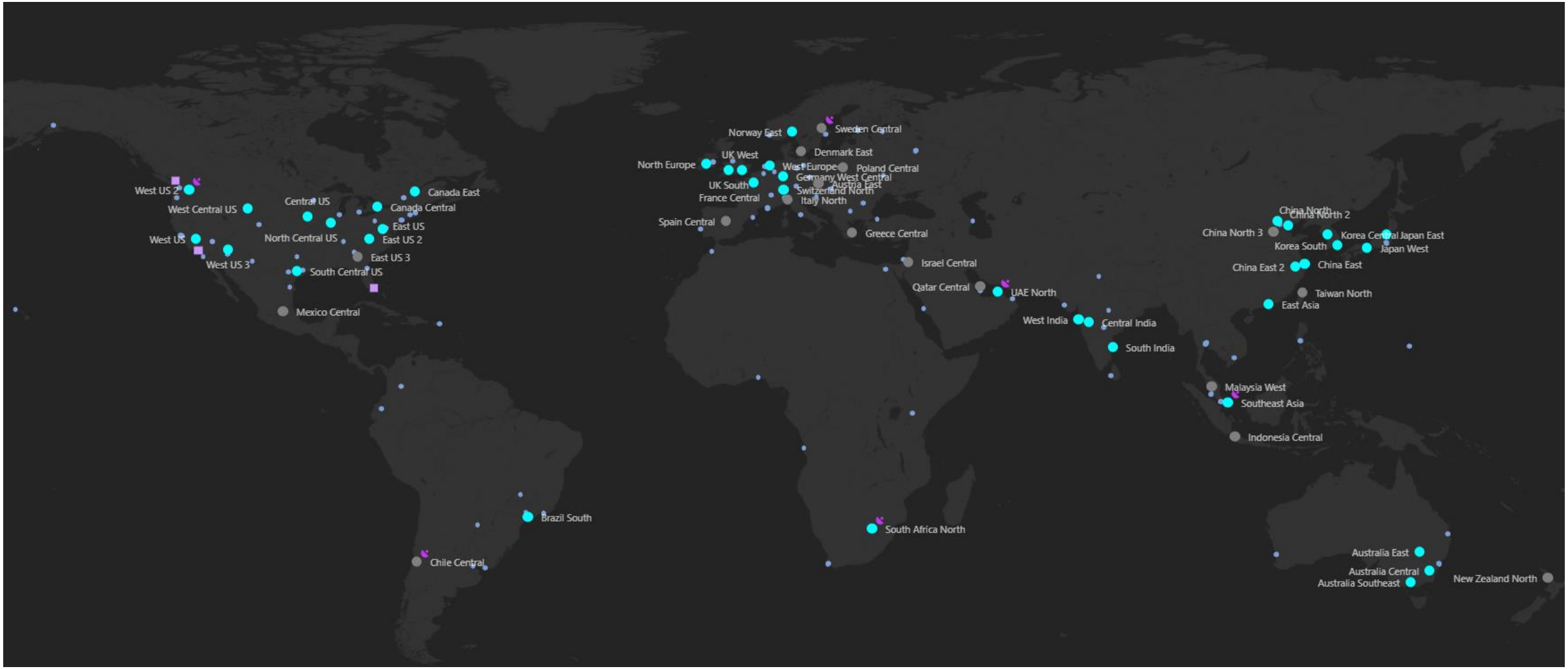
MVAPICH2 on Azure HPC and AI Clusters

Jithin Jose, Microsoft
jijos@microsoft.com

Agenda

- **Overview of Azure HPC**
- Azure HBv3, NDv4
- Latest advances
- MVAPICH2-Azure on HBv3
- MVAPICH2-GDR on NDv4
- Performance Highlights
- Conclusion

Global Infrastructure



Compliance and data residency

Service availability

Pricing

Azure VM Series

Entry Level VMs

Dev/Test
Workloads

General Purpose VMs

Common Applications,
Web Servers, etc.

Compute Optimized VMs

Gaming, Analytics

Large Memory VMs

Large Databases

>80,000 IOPs Premium Storage

Low Latency,
high throughput applications

Storage Optimized VMs

No-SQL Databases,
Data warehousing

InfiniBand

Standard HPC VMs

HPC Applications

GPU VMs

Deep Learning, AI
workloads

FPGA*

FPGA Microservices –
AI/Edge

Cray Services in Azure

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

*InfiniBand/RDMA enabled VMs: One VM per Host

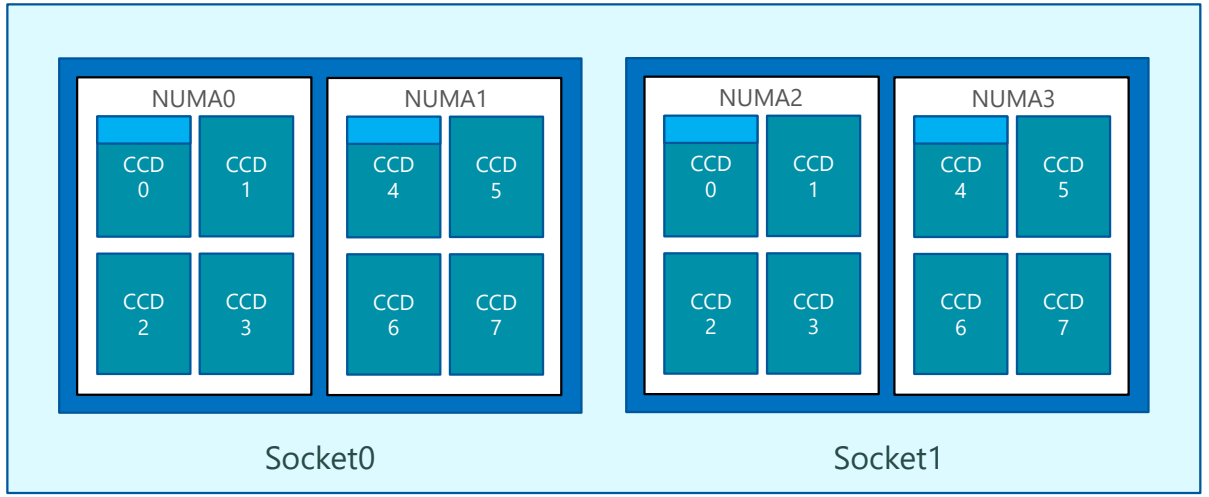
Agenda

- Overview of Azure HPC
- **Azure HBv3, NDv4**
- Latest advances
- MVAPICH2-Azure on HBv3
- MVAPICH2-GDR on NDv4
- Performance Highlights
- Conclusion

Azure HBv3

AMD EPYC Milan

InfiniBand HDR 200Gbps



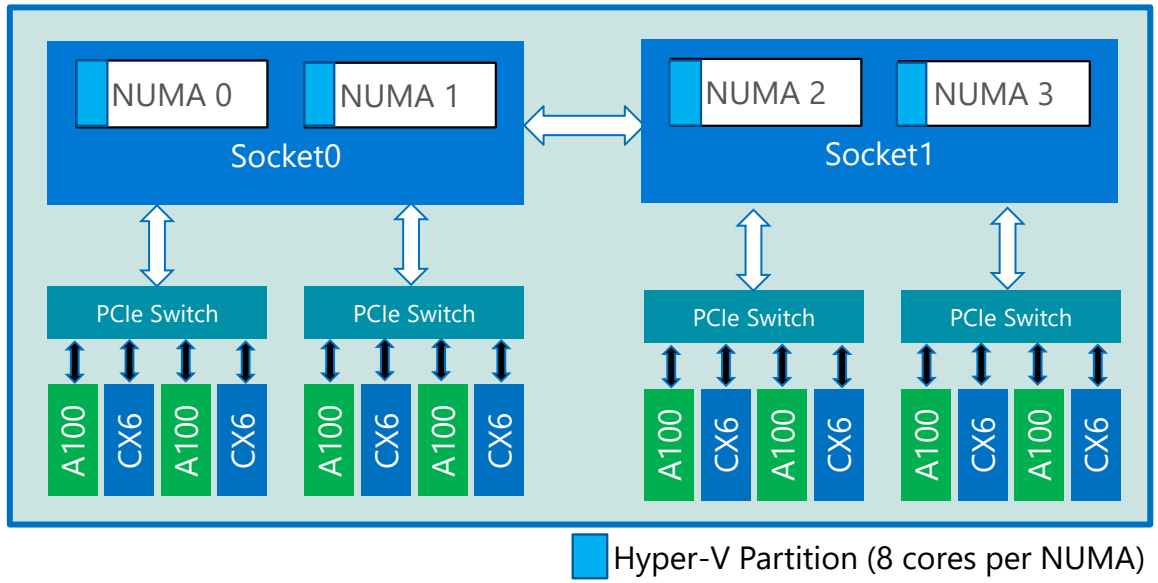
- VM Specs:
  - AMD Milan (NPS = 2)
  - VM Cores: 120
  - Memory: 448 GB
  - Local Disk: 2 x 900 GB NVMe SSD
  - Network: 200 Gbps HDR (SR-IOV)

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)

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



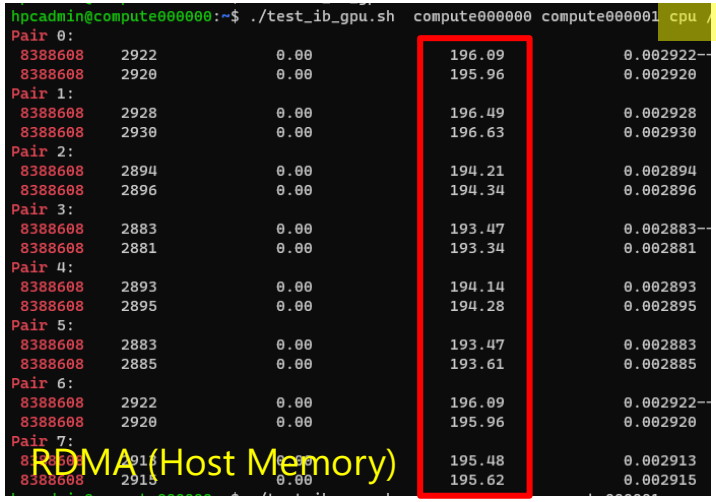
Standard_ND96asr_v4 (NDv4)

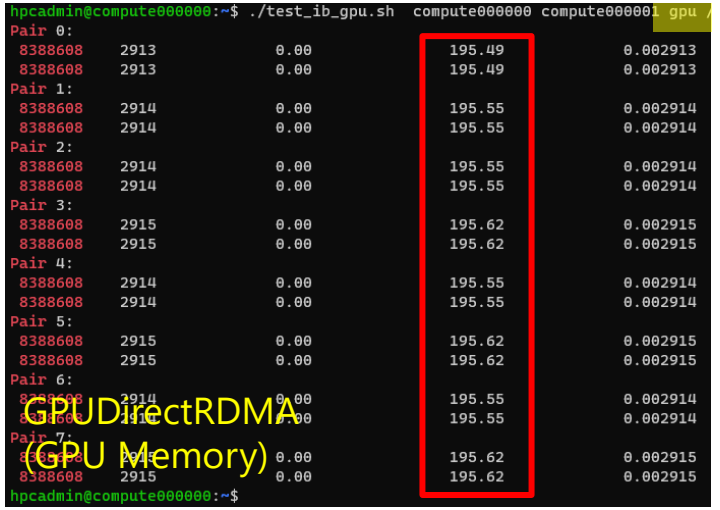
Agenda

- Overview of Azure HPC
- Azure HBv3, NDv4
- **Latest advances**
- MVAPICH2-Azure on HBv3
- MVAPICH2-GDR on NDv4
- Performance Highlights
- Conclusion

GPUDirect RDMA

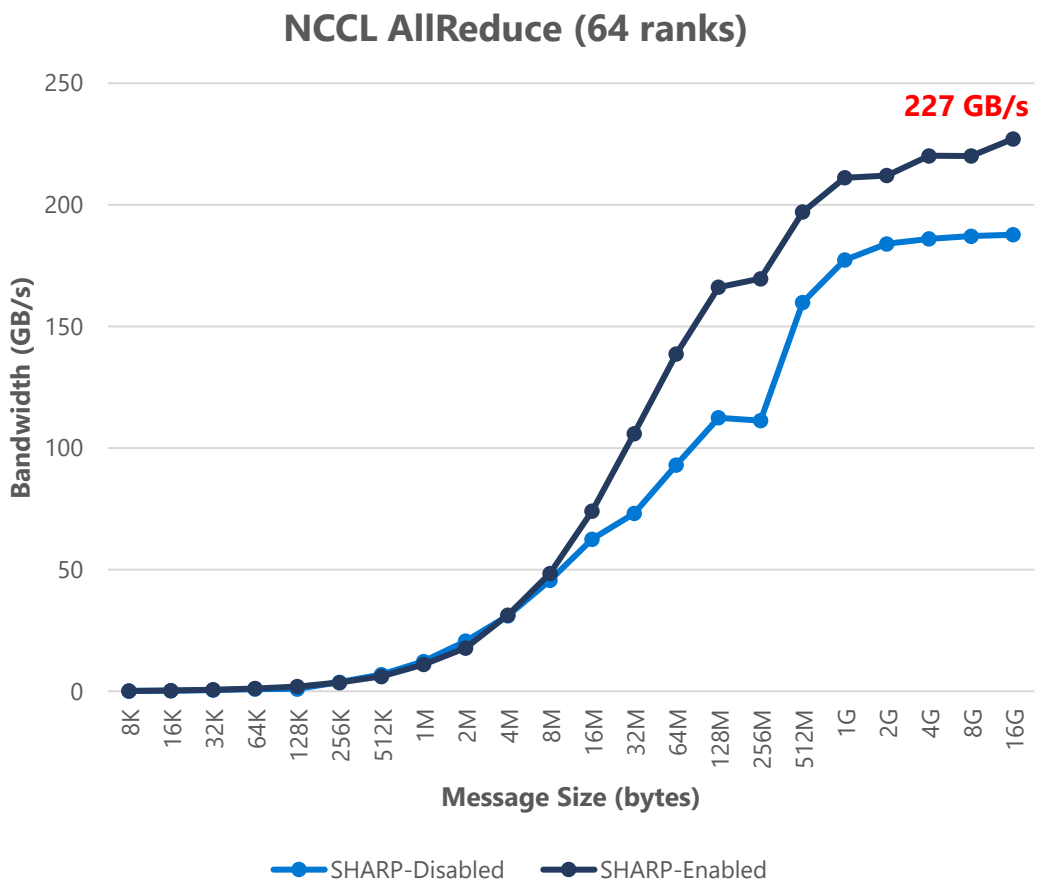
- Available on Azure NDv4
- Direct data path b/w A100 GPU and HDR200
- Each NIC/GPU pair gets peak b/w simultaneously
- Combined GPUDirect RDMA b/w of **1.6 Tbps**
- Supports *all* GDR capable MPI libraries/middleware (including MVAPICH2-GDR)





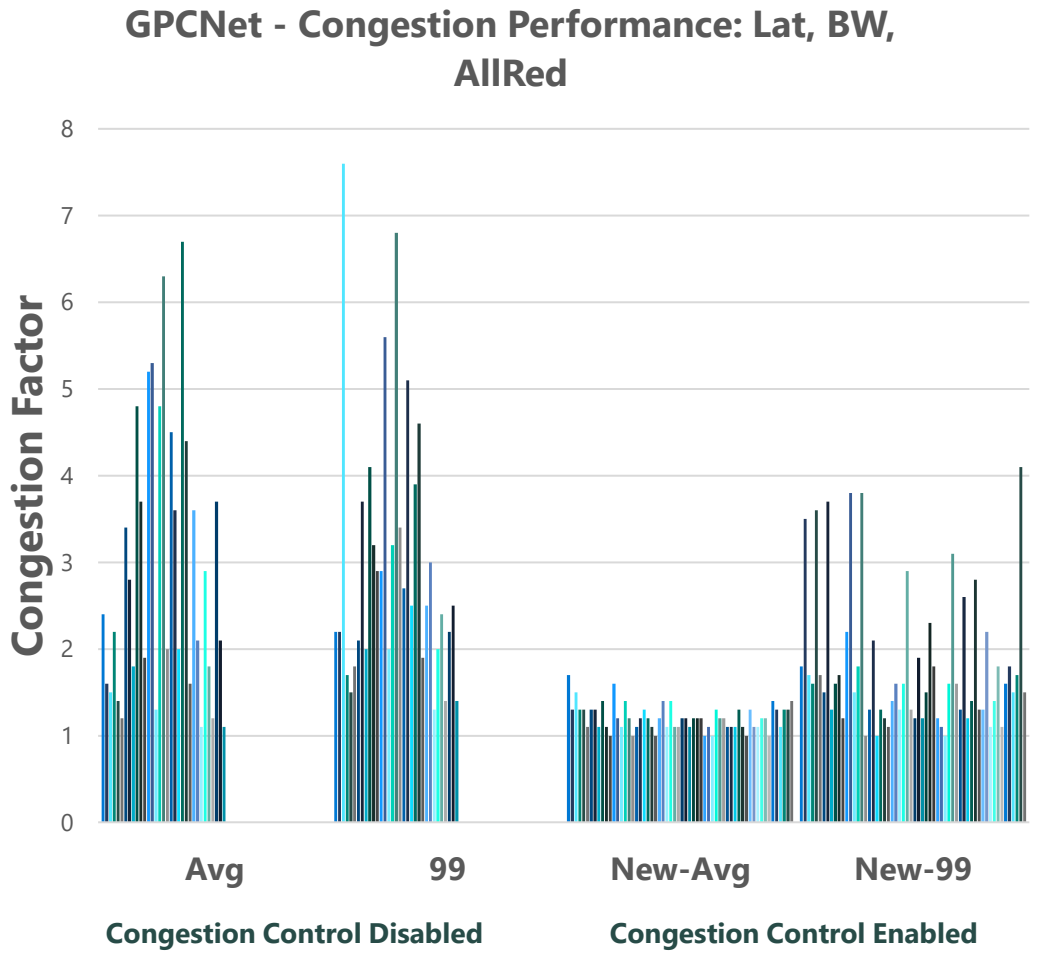
SHARP

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

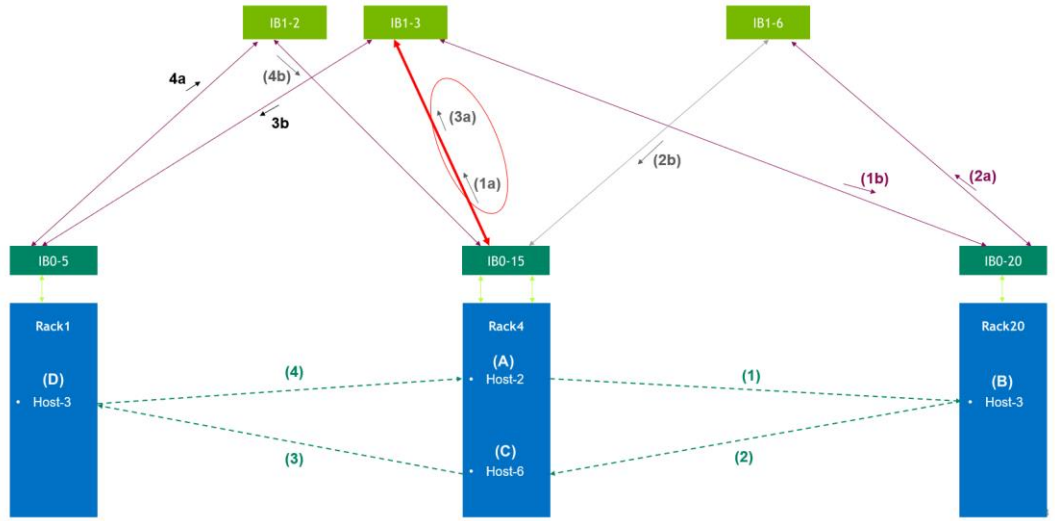


Congestion Control

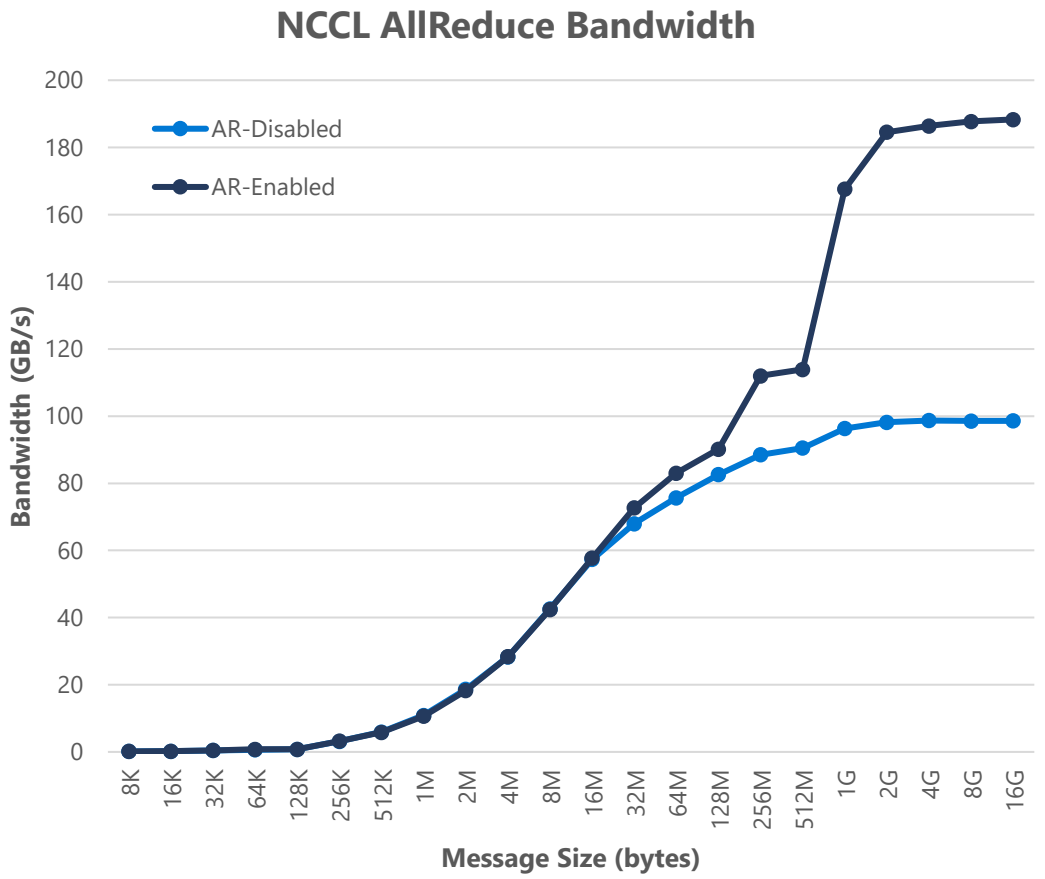
- Available on all VM Series with HDR
- Transparent to customer applications
- Avoids congestion, Improve tail latencies
- Critical in public multi-customer environments



Adaptive Routing



Communication paths during NCCL AllReduce



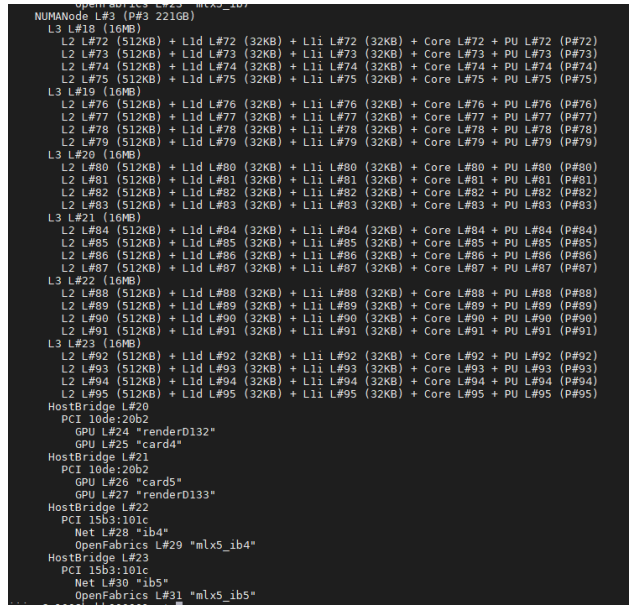
Impact of Adaptive Routing

- Congestion can happen with static routing if a single link is being used by two or more communicating pairs
- AR avoids congestion and offers stable performance

VM Counters, Topology

- NUMA topology
  - NUMA distance
  - L3, L2, PCIe topology info
- VM Performance Counters
  - Select Counters enabled on NDv4
- IB Topology to VMs
  - sharp_cmd topology
  - SLURM topology format

```
node distances:
node   0   1   2   3
  0:  10  12  32  32
  1:  12  10  32  32
  2:  32  32  10  12
  3:  32  32  12  10
```

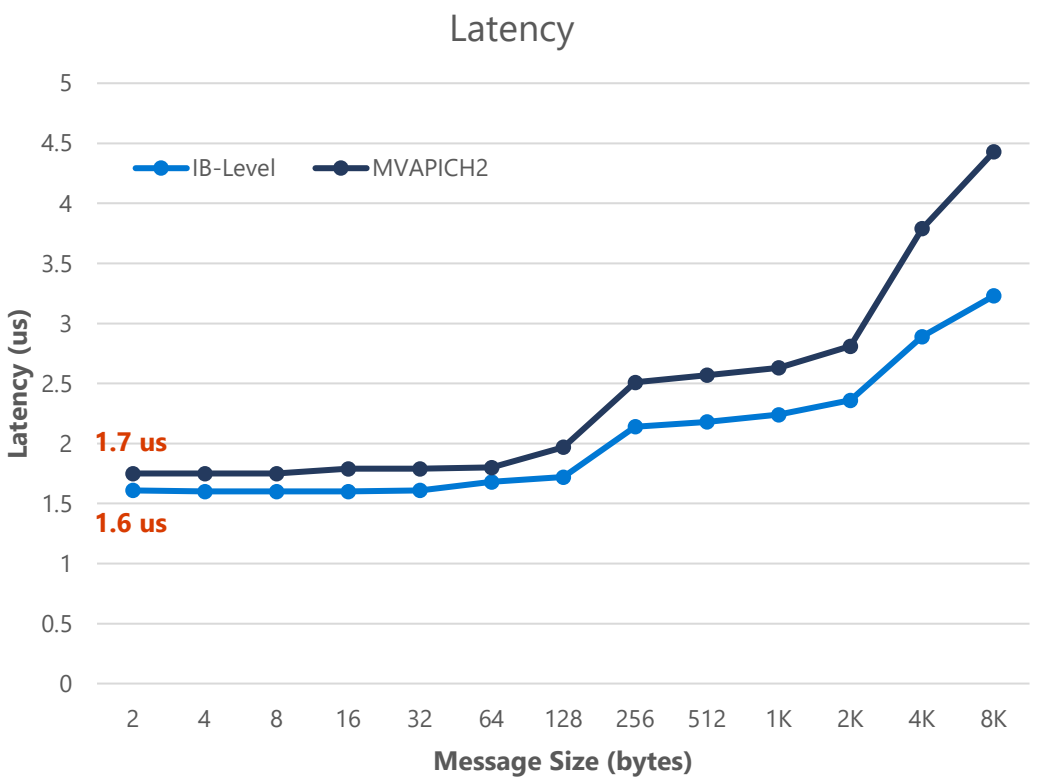


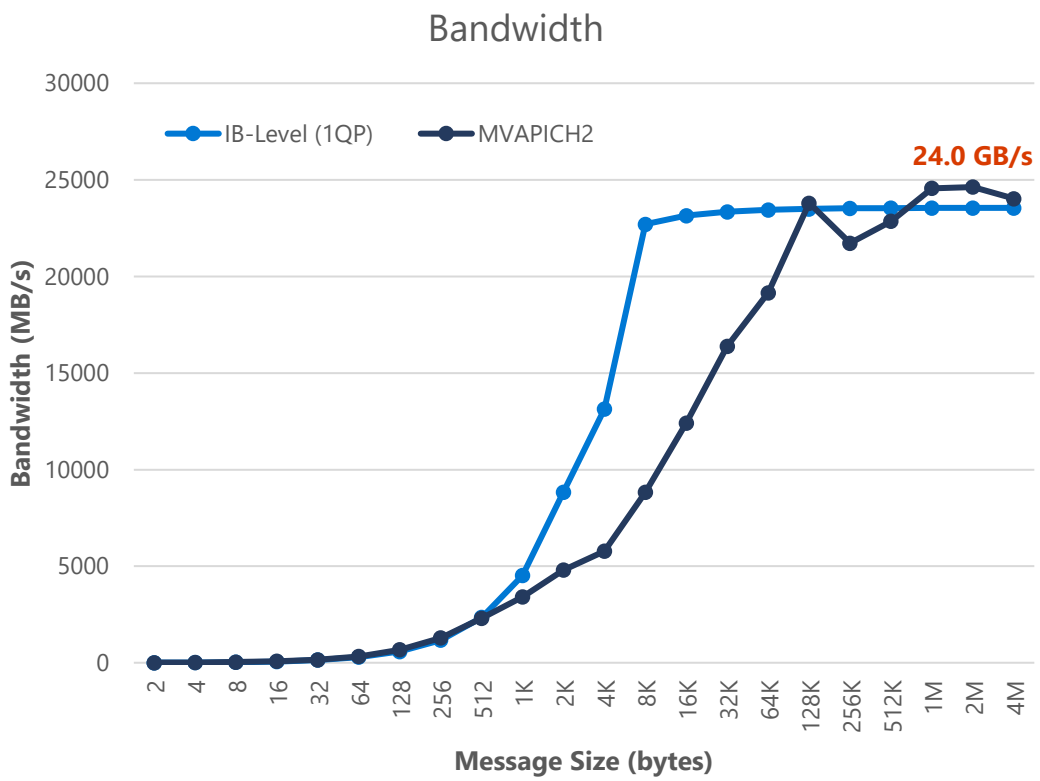

```
jijos@a1008bcbb000001:/opt/AMDuProf_3.4-498/bin$ ./AMDuProfCLI info  --list cacheline-events

        ALIAS        Description
        -----        -----------
    ldst-count       Number of load/stores
      ld-count       Number of loads
      st-count       Number of stores
    cache-hitm       Number of loads serviced by modified cache
lcl-cache-hitm       Number of loads serviced by modified cache on the local node
rmt-cache-hitm       Number of loads serviced by modified cache on remote node
  lcl-dram-hit       Total local DRAM hit
  rmt-dram-hit       Total remote DRAM hit
       l3-miss       Total l3 miss
    st-dc-miss       Number of store instruction which incurred a data cache miss
```

Agenda

- Overview of Azure HPC
- Azure HBv3, NDv4
- Latest advances
- **MVAPICH2-Azure on HBv3**
- MVAPICH2-GDR on NDv4
- Performance Highlights
- Conclusion

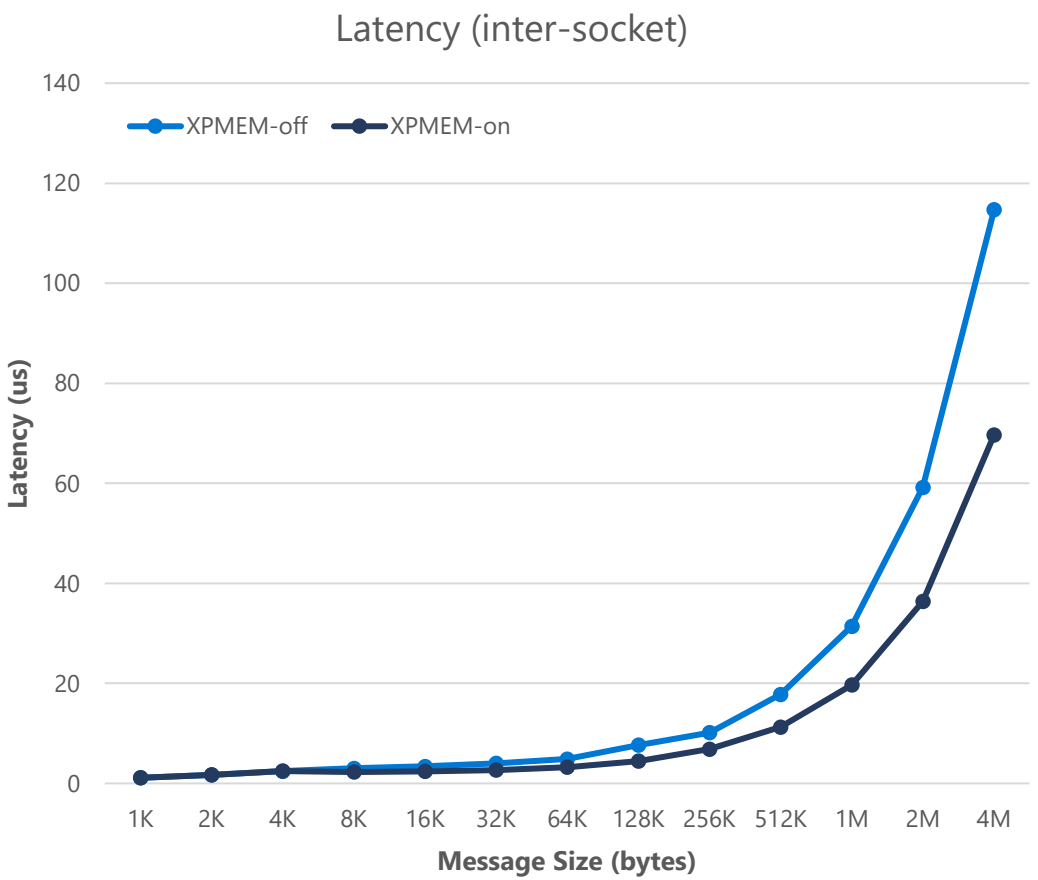
MVAPICH2-Azure on HBv3 (inter-node)

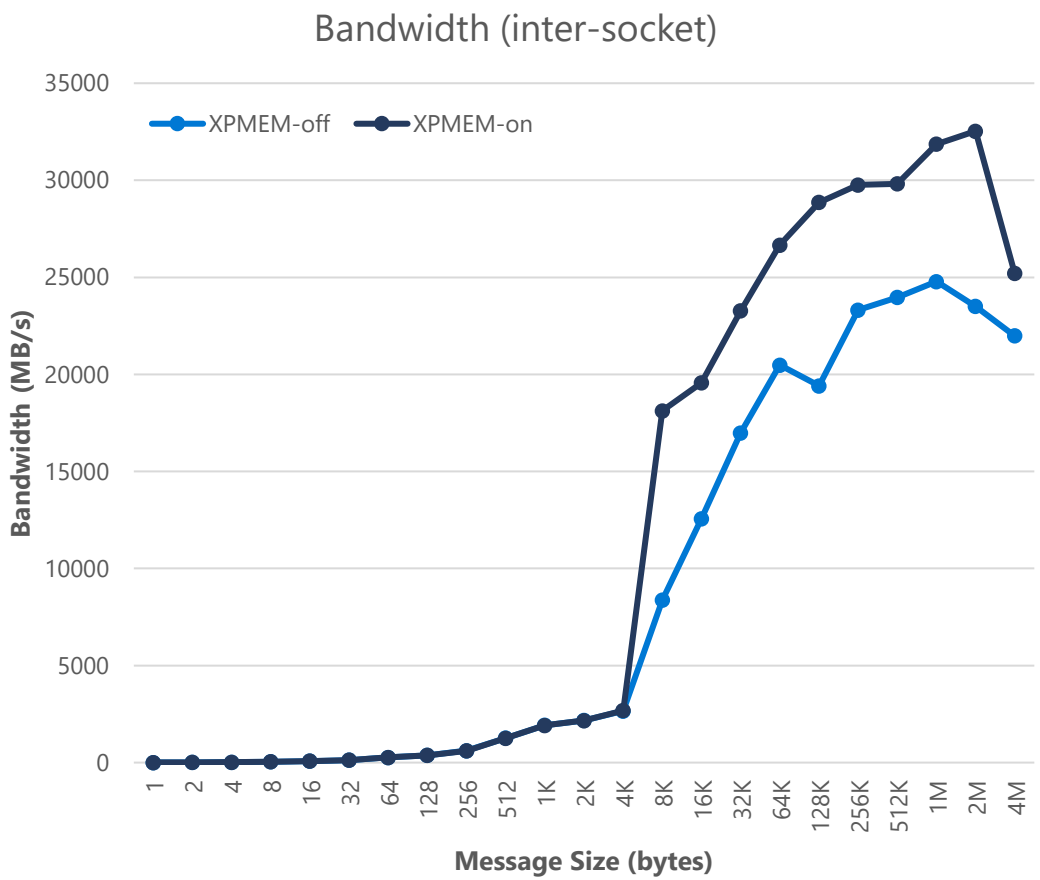




Environment parameters: MV2_NUM_QP_PER_PORT=2 MV2_IBA_EAGER_THRESHOLD=66560 MV2_VBUF_TOTAL_SIZE=66560 MV2_RNDV_PROTOCOL=RPUT

MVAPICH2-Azure on HBv3 (intra-node)

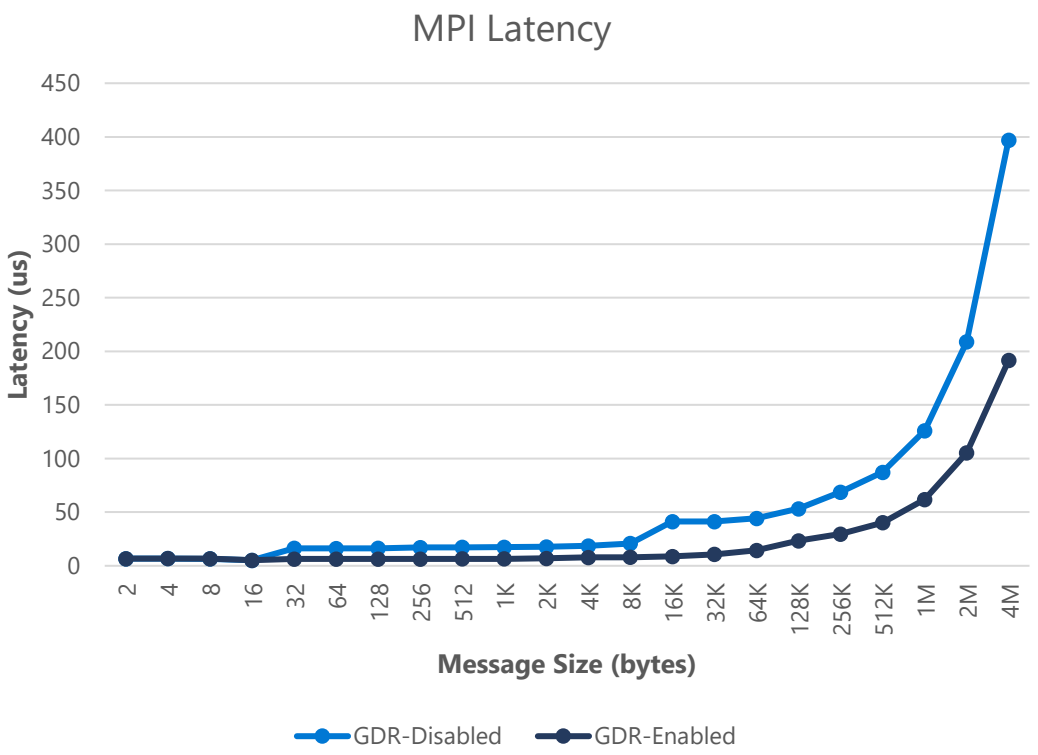


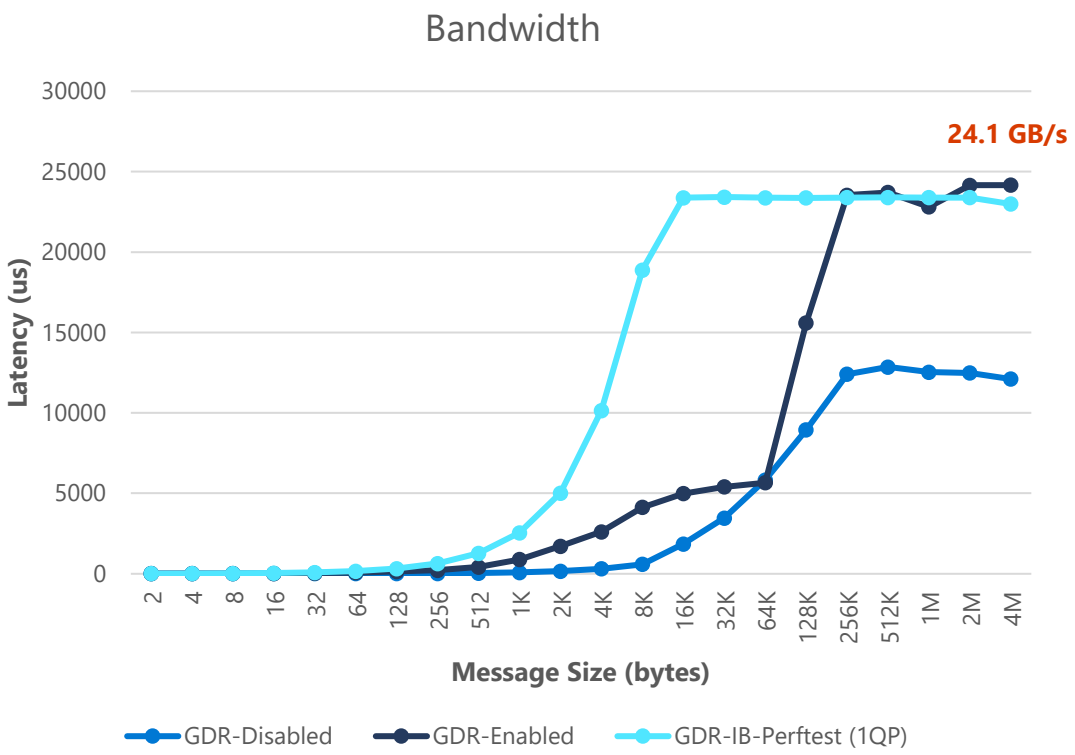


Agenda

- Overview of Azure HPC
- Azure HBv3, NDv4
- Latest advances
- MVAPICH2-Azure on HBv3
- **MVAPICH2-GDR on NDv4**
- Performance Highlights
- Conclusion

MVAPICH2-GDR on NDv4

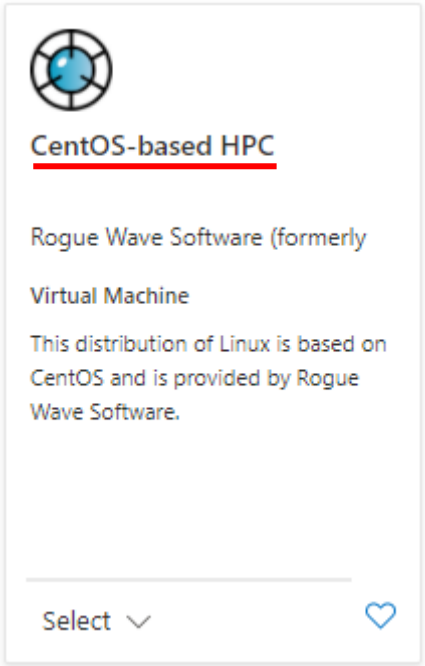


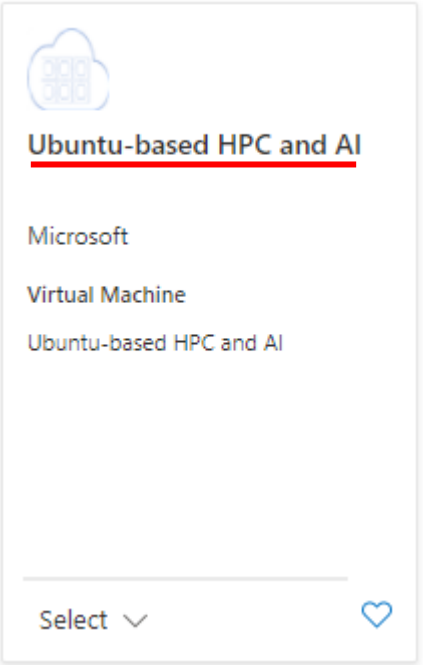


Environment parameters: MV2_NUM_QP_PER_PORT=4 MV2_IBA_EAGER_THRESHOLD=66560 MV2_VBUF_TOTAL_SIZE=66560 MV2_RNDV_PROTOCOL=RPUT MV2_CUDA_BLOCK_SIZE=131072 MV2_USE_GPUDIRECT_RDMA=1 MLX5_RELAXED_PACKET_ORDERING_ON=all MV2_GPUDIRECT_LIMIT=4194304 MV2_USE_CUDA=1 MV2_IBA_HCA=mlx5_ib0 CUDA_VISIBLE_DEVICES=0

Azure HPC VM Images

- Optimized VM Images for HPC/AI workloads
- Mellanox OFED
- Pre-configured IPoIB
- InfiniBand based MPI Libraries
  - HPC-X, IntelMPI, **MVAPICH2**, OpenMPI
- Communication Runtimes
  - Libfabric, OpenUCX
- Optimized HPC libraries
  - Blis, FFTW, Flame, MKL
- Recommended Compilers
- GPU Drivers
- NCCL, NCCL RDMA Sharp Plugin, SharpD
- Other optimizations

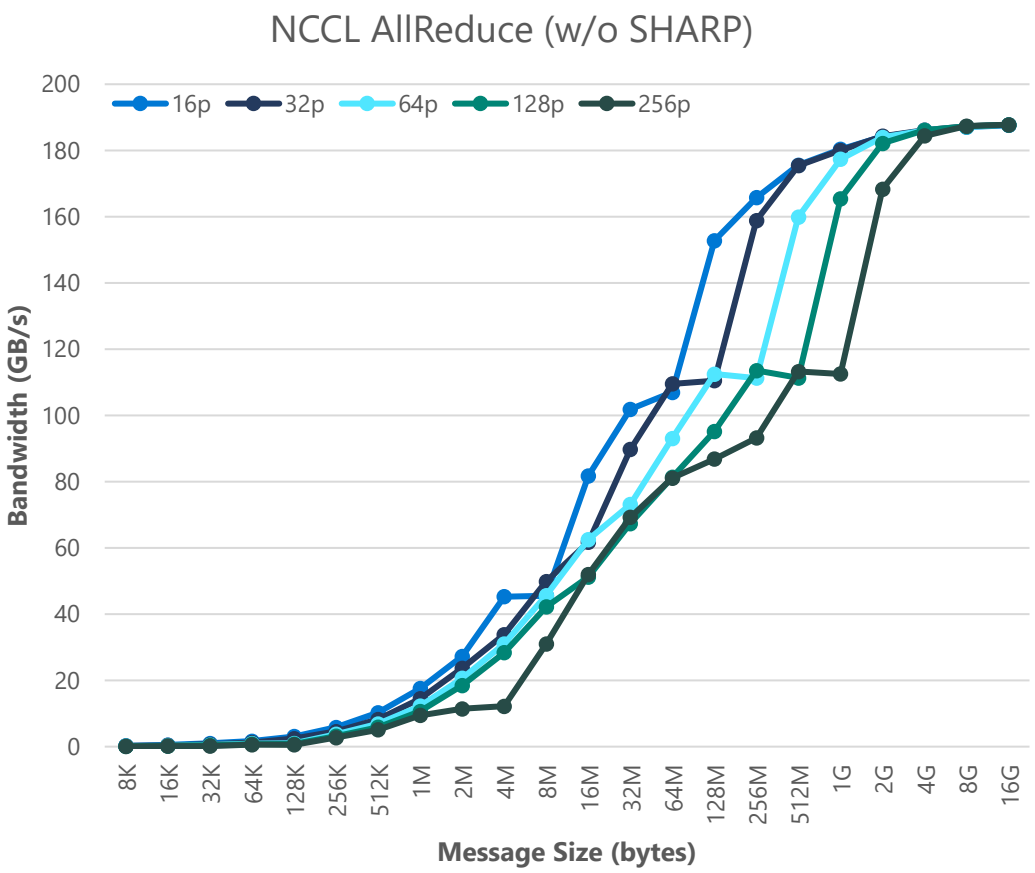


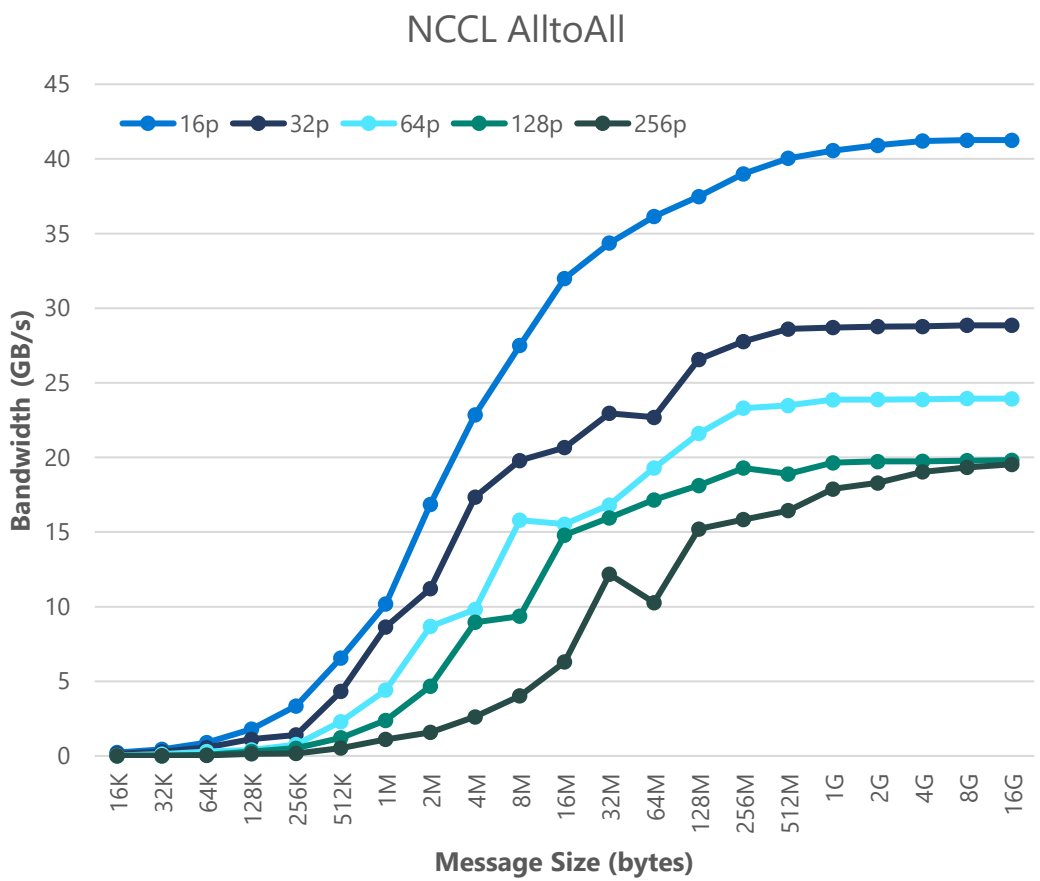


Agenda

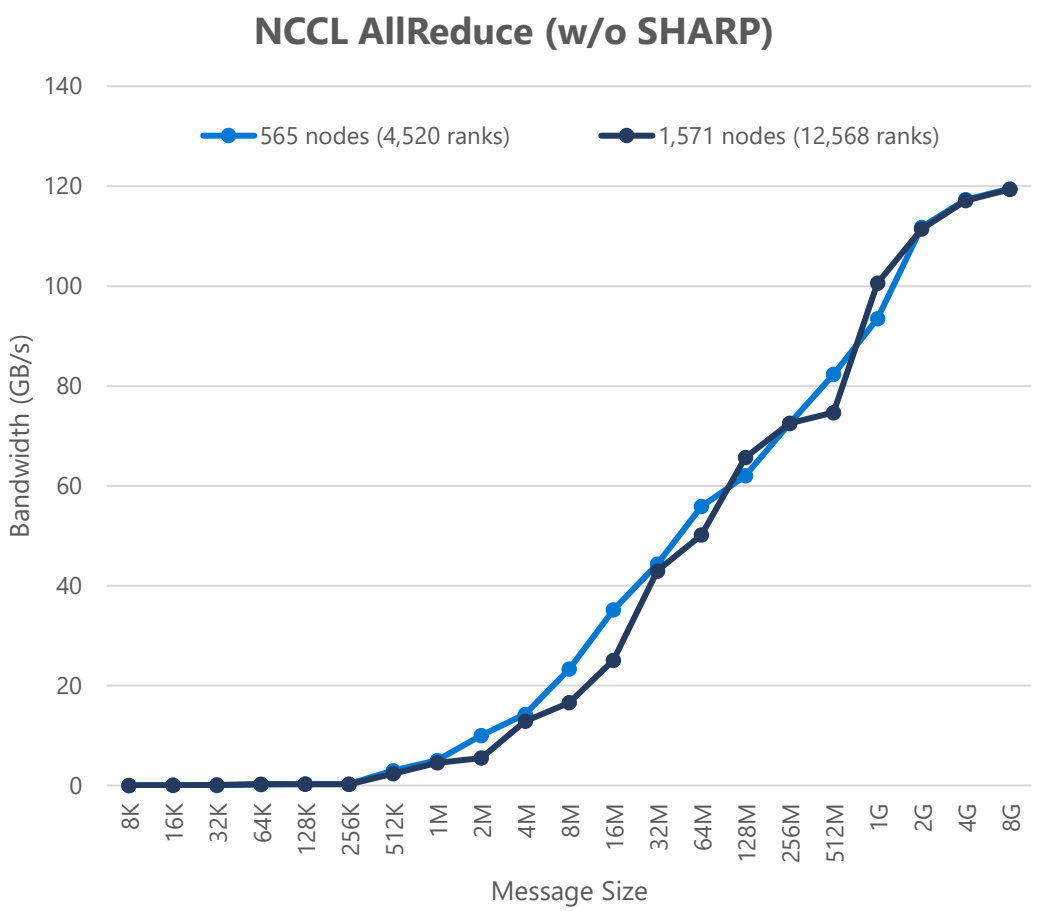
- Overview of Azure HPC
- Azure HBv3, NDv4
- Latest advances
- MVAPICH2-Azure on HBv3
- MVAPICH2-GDR on NDv4
- **Performance Highlights**
- Conclusion

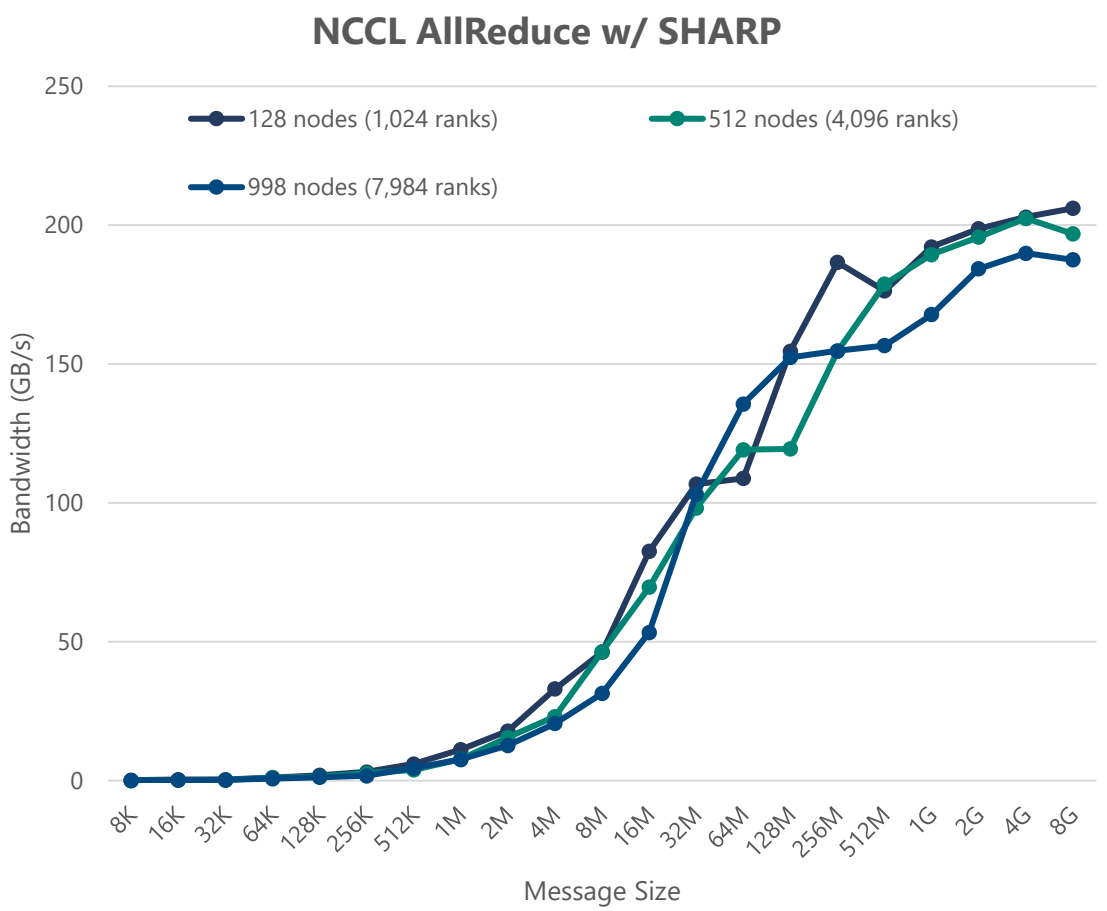
NCCL on NDv4


NCCL AllReduce (w/o SHARP)
16p
32p
64p
128p
256p
Bandwidth (GB/s)
0
20
40
60
80
100
120
140
160
180
200
8K
16K
32K
64K
128K
256K
512K
1M
2M
4M
8M
16M
32M
64M
128M
256M
512M
1G
2G
4G
8G
16G
Message Size (bytes)

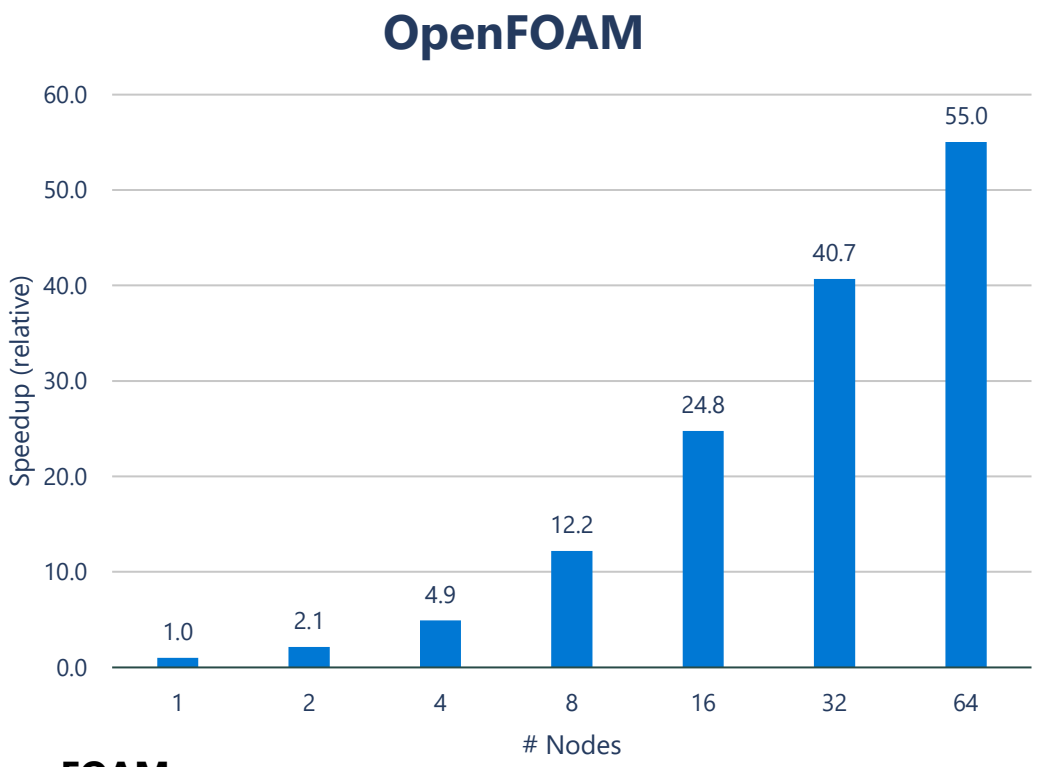

NCCL AlltoAll
16p
32p
64p
128p
256p
Bandwidth (GB/s)
0
5
10
15
20
25
30
35
40
45
16K
32K
64K
128K
256K
512K
1M
2M
4M
8M
16M
32M
64M
128M
256M
512M
1G
2G
4G
8G
16G
Message Size (bytes)

NCCL at Scale on NDv4



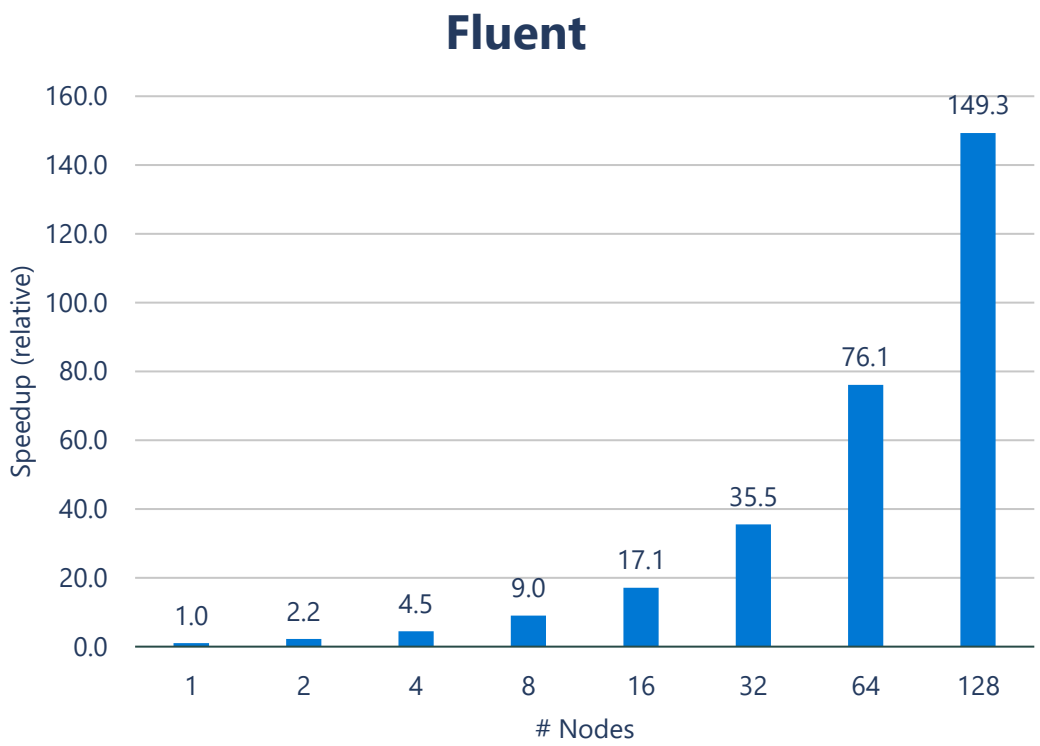


HBv3 Performance Advantage



OpenFOAM

- Motorbike 28M cells mesh
- Azure HBv3 VM instances (up to 7,680 MPI ranks)



Fluent v19.5

- F1 Race Car – 140 M
- Azure HBv3 VM instances (up to 15,104 MPI ranks)

Agenda

- Overview of Azure HPC
- Azure HBv3, NDv4
- Latest advances
- MVAPICH2-Azure on HBv3
- MVAPICH2-GDR on NDv4
- Performance Highlights
- **Conclusion**

Conclusion

- Supercomputer on Cloud is real!
- Azure HPC Cloud made into Top500, Graph500
- Cloud democratize Supercomputer
- High Performance middleware such as MVAPICH2 enables cutting edge technology
  - Deliver High Scalability and Performance

Resources

Getting Started

- High Performance Computing (HPC) on Azure

HPC VM Images

- Azure HPC VM Images
- GitHub Repository

HPC VM Series

- Azure VM sizes - HPC - Azure Virtual Machines

HPC VM Deployment

- Sample HPC VM deployment scripts
- Azure CycleCloud
- MUG '20 Tutorial

GPU VM Series

- Azure VM sizes - GPU - Azure Virtual Machines

Azure HPC Blogs

- Azure Compute - Microsoft Tech Community

We are Hiring!

jijos@microsoft.com


Microsoft
Thank you