



MVAPICH2 on Azure HPC and AI Clusters

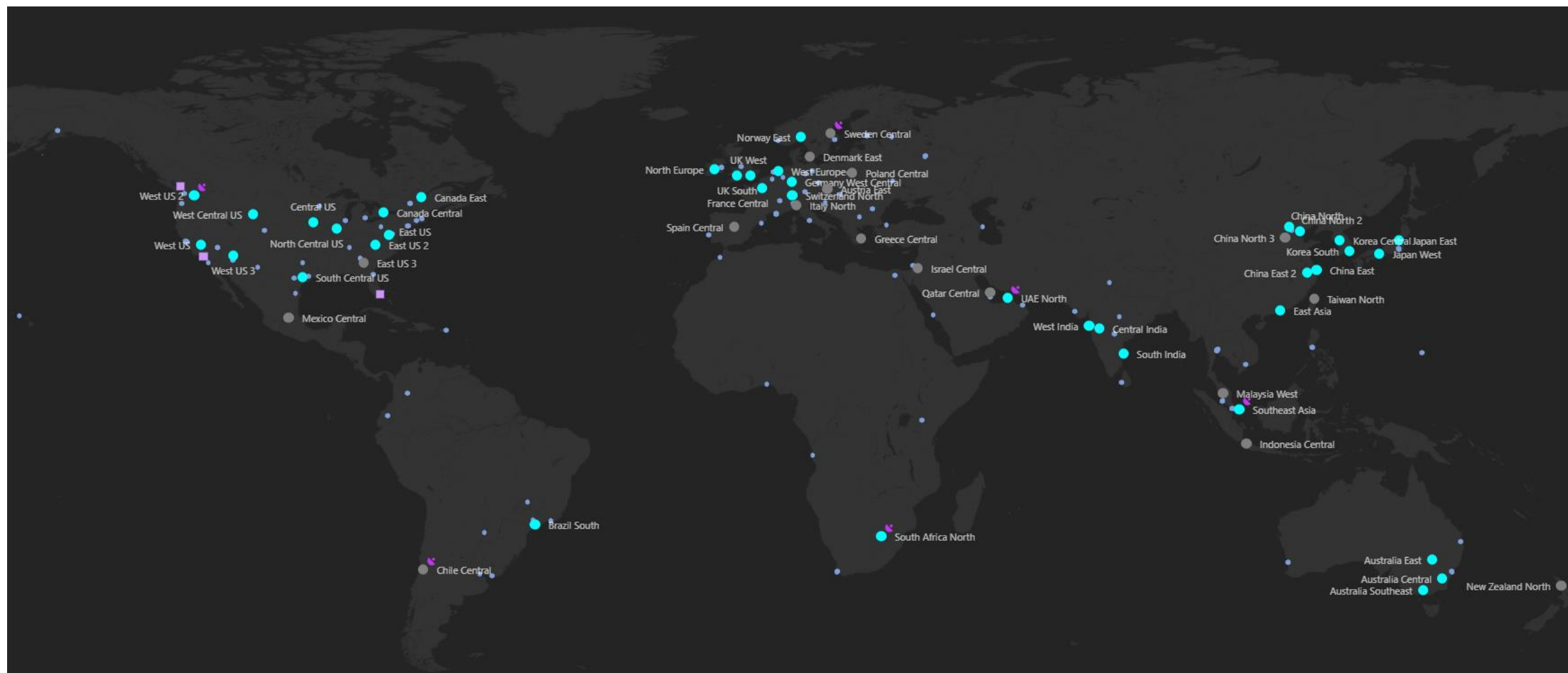
Jithin Jose, Microsoft
jjjos@microsoft.com

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



**Standard HPC
VMs**

HPC Applications



GPU VMs

Deep Learning, AI
workloads

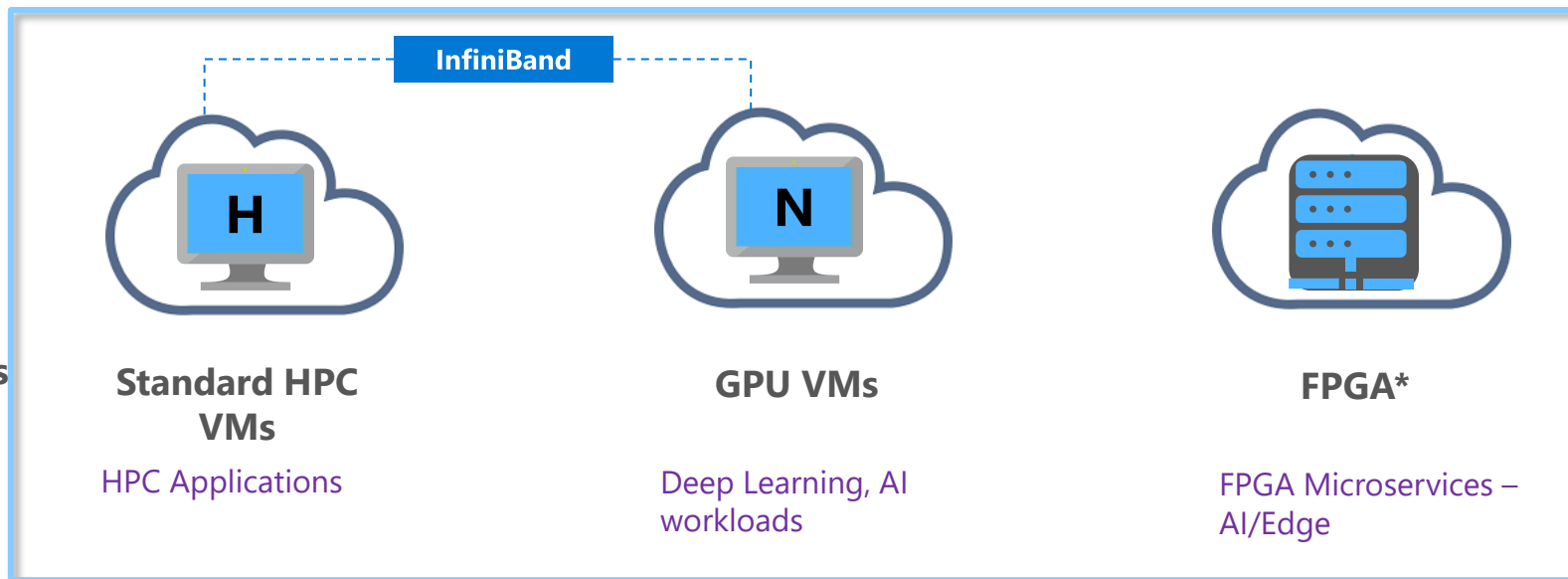


FPGA*

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



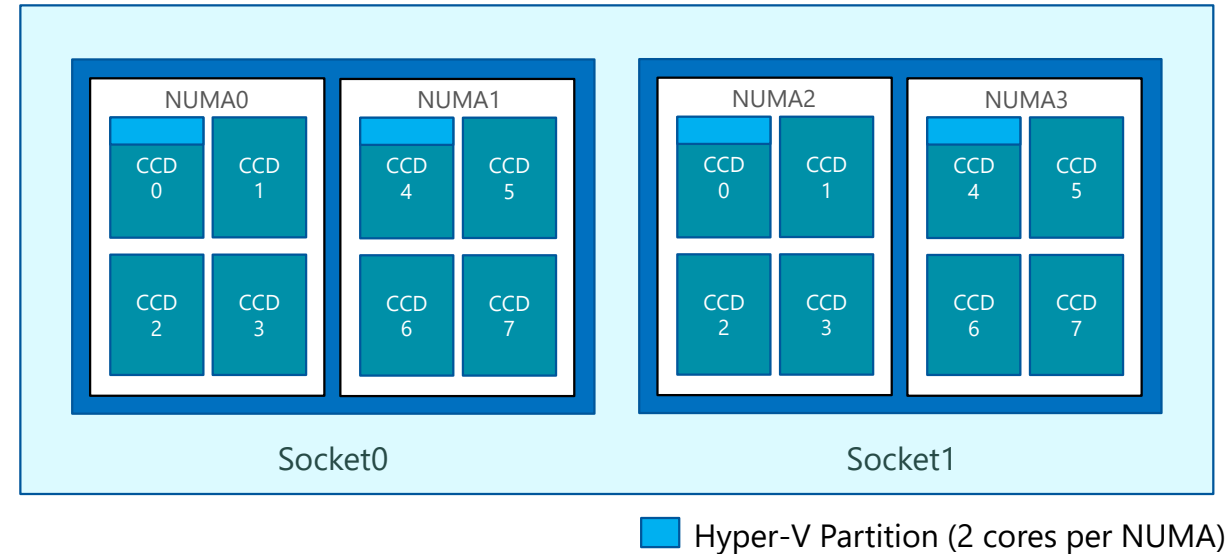
AMD EPYC
Milan



NVIDIA®
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)



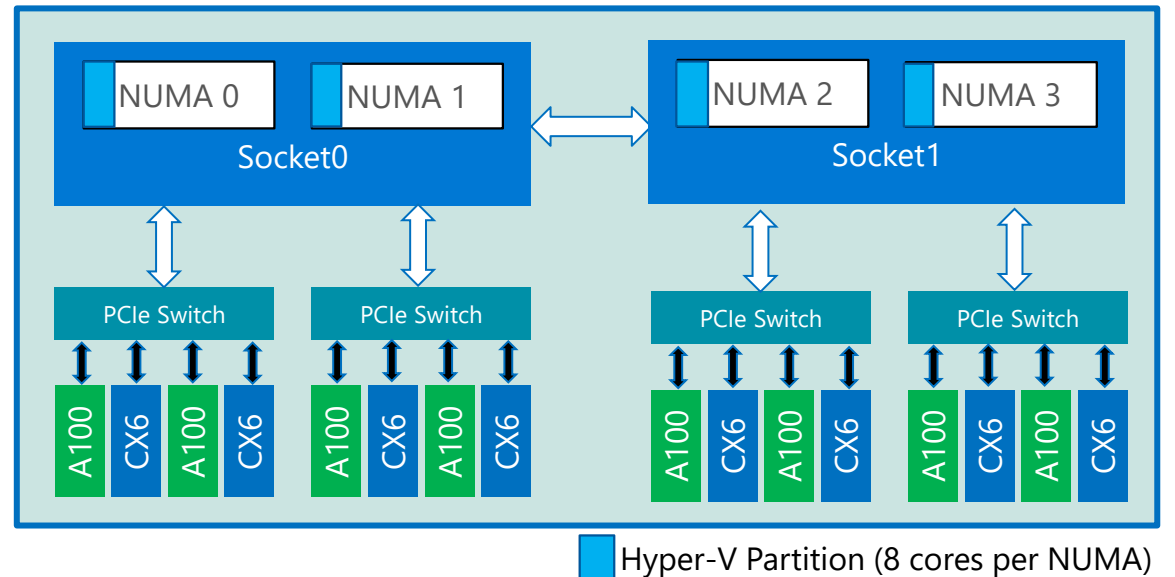
HBv3 VM Sizes (one VM per Host):

- 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)

```
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 2916 0.00 195.48 0.002913
8388608 2915 0.00 195.62 0.002915
```

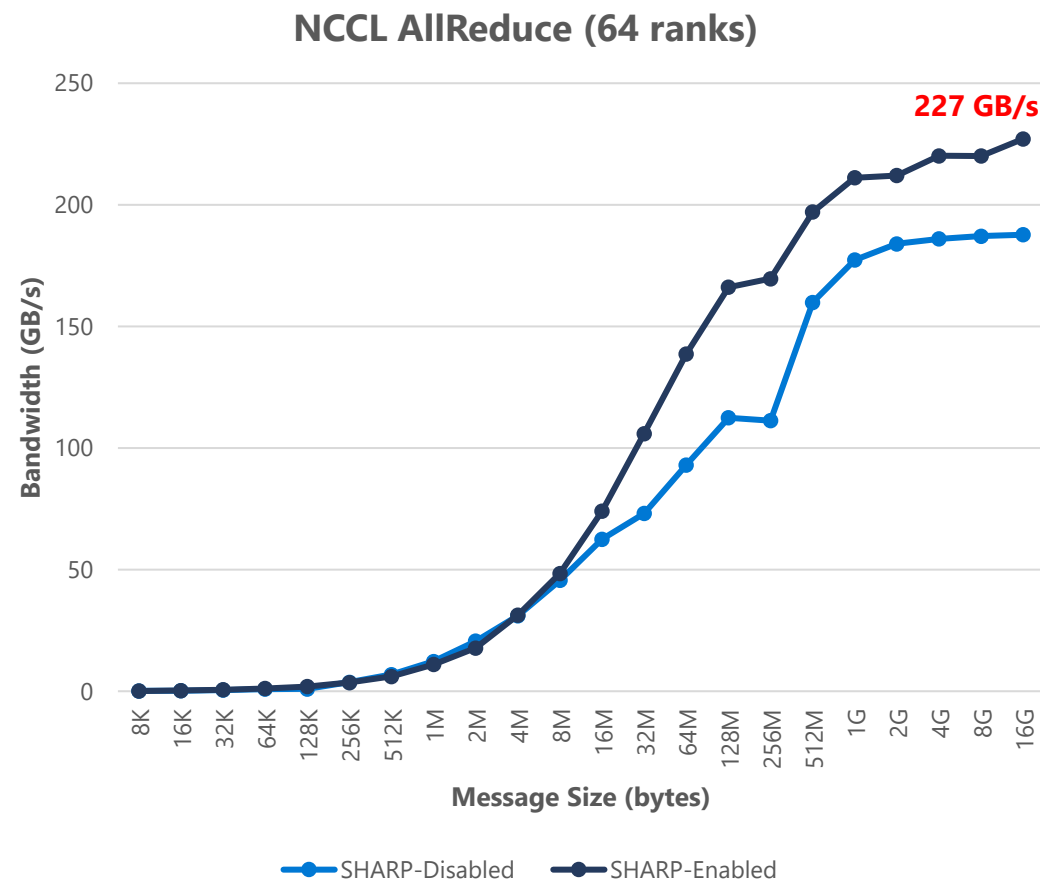
GPUDirect RDMA (Host Memory)

```
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:~$
```

GPUDirect RDMA (GPU Memory)

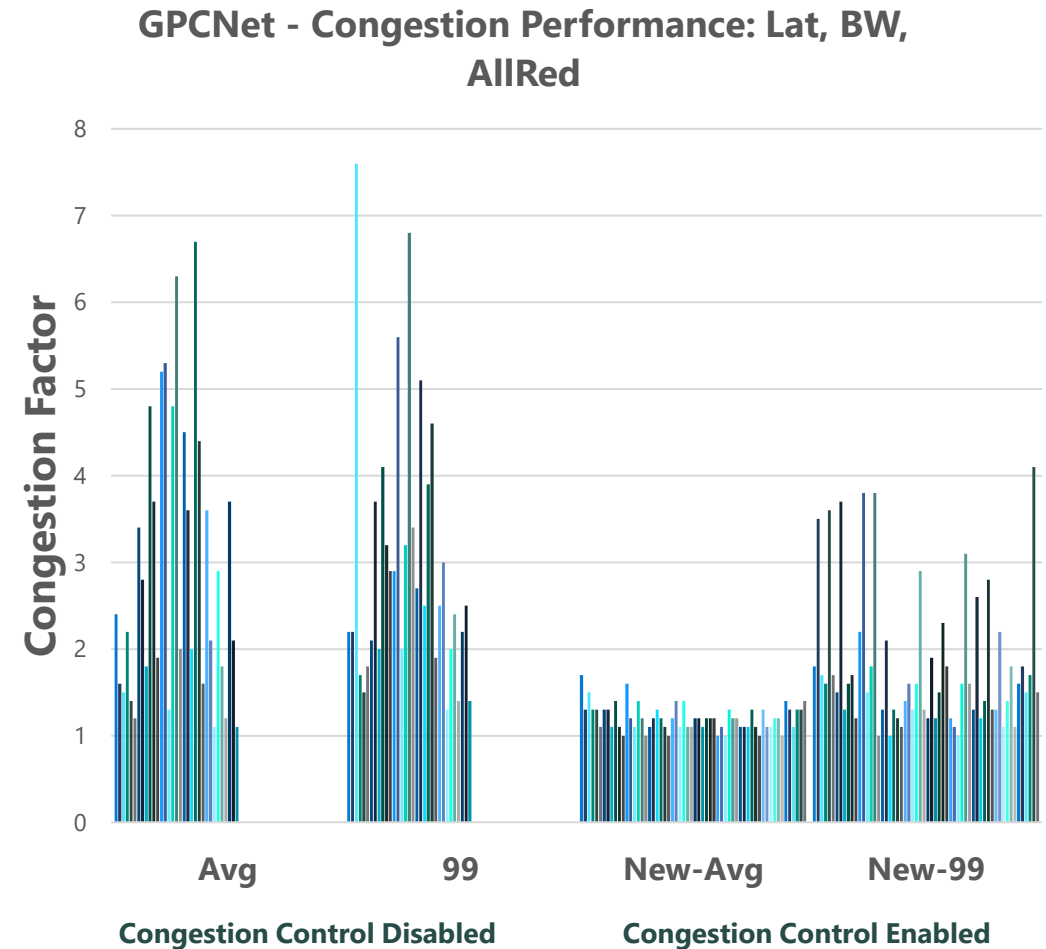
SHARP

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

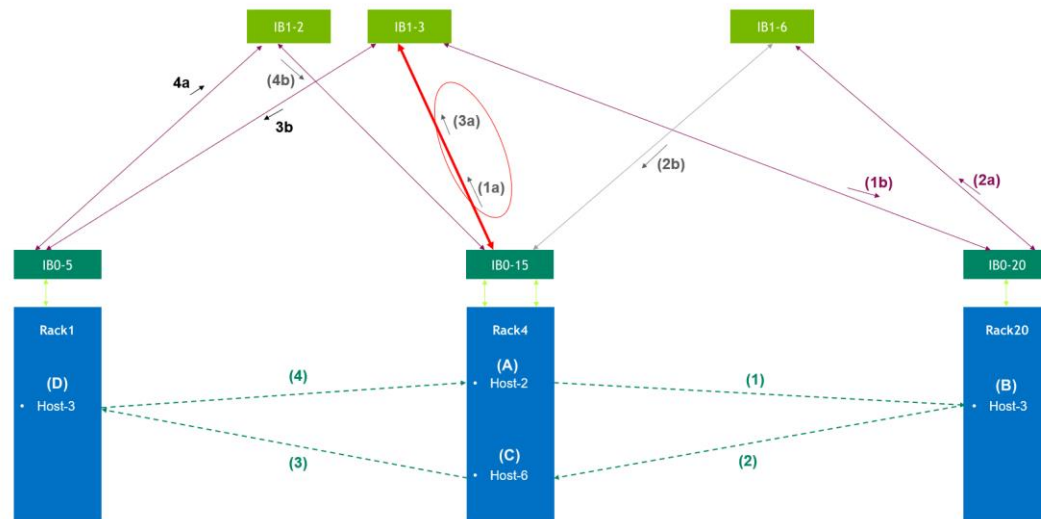


Congestion Control

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



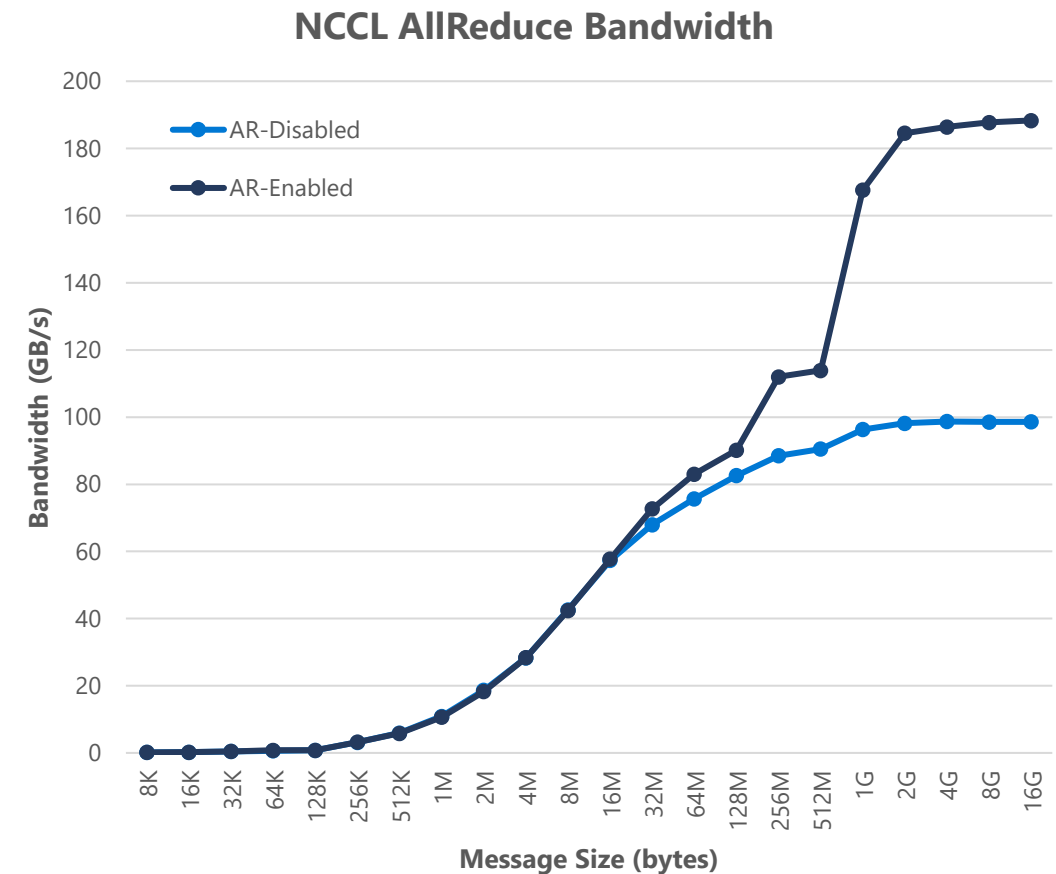
Adaptive Routing



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

```
Open 2017-11-13 14:42:00.107
NUMANode L#3 (P#3 221GB)
L3 L#18 (16MB)
L2 L#72 (512KB) + L1d L#72 (32KB) + L1i L#72 (32KB) + Core L#72 + PU L#72 (P#72)
L2 L#73 (512KB) + L1d L#73 (32KB) + L1i L#73 (32KB) + Core L#73 + PU L#73 (P#73)
L2 L#74 (512KB) + L1d L#74 (32KB) + L1i L#74 (32KB) + Core L#74 + PU L#74 (P#74)
L2 L#75 (512KB) + L1d L#75 (32KB) + L1i L#75 (32KB) + Core L#75 + PU L#75 (P#75)
L3 L#19 (16MB)
L2 L#76 (512KB) + L1d L#76 (32KB) + L1i L#76 (32KB) + Core L#76 + PU L#76 (P#76)
L2 L#77 (512KB) + L1d L#77 (32KB) + L1i L#77 (32KB) + Core L#77 + PU L#77 (P#77)
L2 L#78 (512KB) + L1d L#78 (32KB) + L1i L#78 (32KB) + Core L#78 + PU L#78 (P#78)
L2 L#79 (512KB) + L1d L#79 (32KB) + L1i L#79 (32KB) + Core L#79 + PU L#79 (P#79)
L3 L#20 (16MB)
L2 L#80 (512KB) + L1d L#80 (32KB) + L1i L#80 (32KB) + Core L#80 + PU L#80 (P#80)
L2 L#81 (512KB) + L1d L#81 (32KB) + L1i L#81 (32KB) + Core L#81 + PU L#81 (P#81)
L2 L#82 (512KB) + L1d L#82 (32KB) + L1i L#82 (32KB) + Core L#82 + PU L#82 (P#82)
L2 L#83 (512KB) + L1d L#83 (32KB) + L1i L#83 (32KB) + Core L#83 + PU L#83 (P#83)
L3 L#21 (16MB)
L2 L#84 (512KB) + L1d L#84 (32KB) + L1i L#84 (32KB) + Core L#84 + PU L#84 (P#84)
L2 L#85 (512KB) + L1d L#85 (32KB) + L1i L#85 (32KB) + Core L#85 + PU L#85 (P#85)
L2 L#86 (512KB) + L1d L#86 (32KB) + L1i L#86 (32KB) + Core L#86 + PU L#86 (P#86)
L2 L#87 (512KB) + L1d L#87 (32KB) + L1i L#87 (32KB) + Core L#87 + PU L#87 (P#87)
L3 L#22 (16MB)
L2 L#88 (512KB) + L1d L#88 (32KB) + L1i L#88 (32KB) + Core L#88 + PU L#88 (P#88)
L2 L#89 (512KB) + L1d L#89 (32KB) + L1i L#89 (32KB) + Core L#89 + PU L#89 (P#89)
L2 L#90 (512KB) + L1d L#90 (32KB) + L1i L#90 (32KB) + Core L#90 + PU L#90 (P#90)
L2 L#91 (512KB) + L1d L#91 (32KB) + L1i L#91 (32KB) + Core L#91 + PU L#91 (P#91)
L3 L#23 (16MB)
L2 L#92 (512KB) + L1d L#92 (32KB) + L1i L#92 (32KB) + Core L#92 + PU L#92 (P#92)
L2 L#93 (512KB) + L1d L#93 (32KB) + L1i L#93 (32KB) + Core L#93 + PU L#93 (P#93)
L2 L#94 (512KB) + L1d L#94 (32KB) + L1i L#94 (32KB) + Core L#94 + PU L#94 (P#94)
L2 L#95 (512KB) + L1d L#95 (32KB) + L1i L#95 (32KB) + Core L#95 + PU L#95 (P#95)
HostBridge L#20
PCI 10de:20b2
GPU L#24 "renderD132"
GPU L#25 "card4"
HostBridge L#21
PCI 10de:20b2
GPU L#26 "card5"
GPU L#27 "renderD133"
HostBridge L#22
PCI 15b3:101c
Net L#28 "ib4"
OpenFabrics L#29 "mlx5_ib4"
HostBridge L#23
PCI 15b3:101c
Net L#30 "ib5"
OpenFabrics L#31 "mlx5_ib5"
```

```
jijos@a1008bcb000001:/opt/AMDuProf_3.4-498/bin$ ./AMDuProfCLI info --list cacheline-events
```

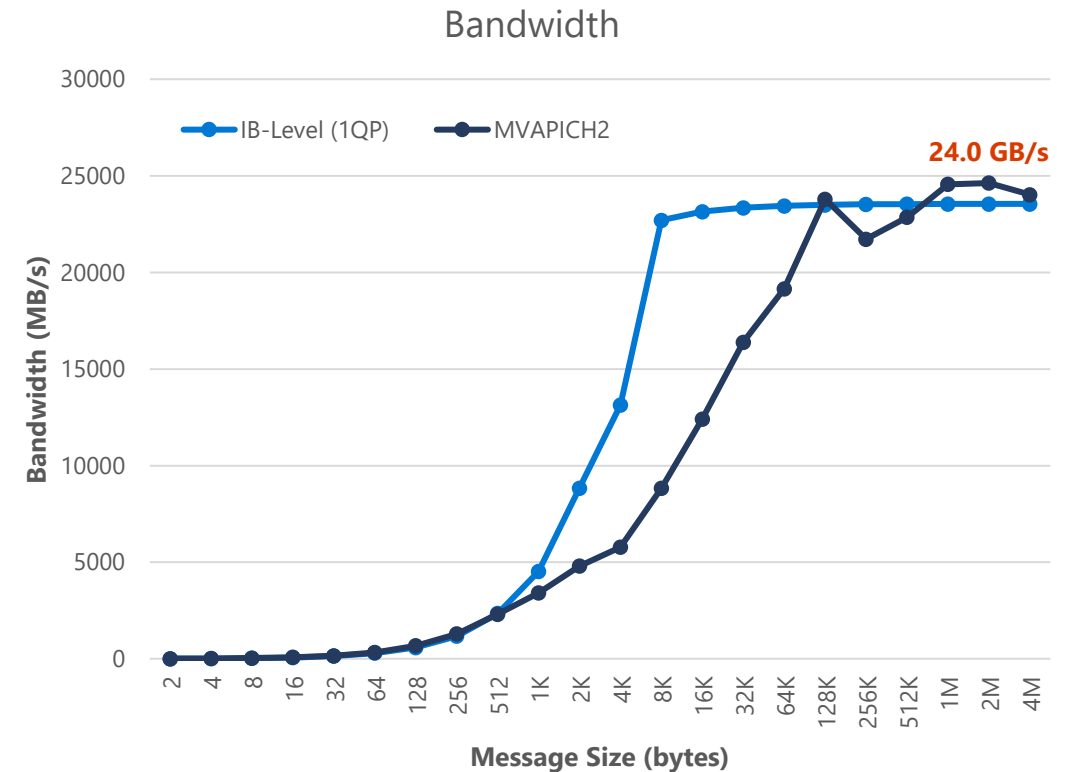
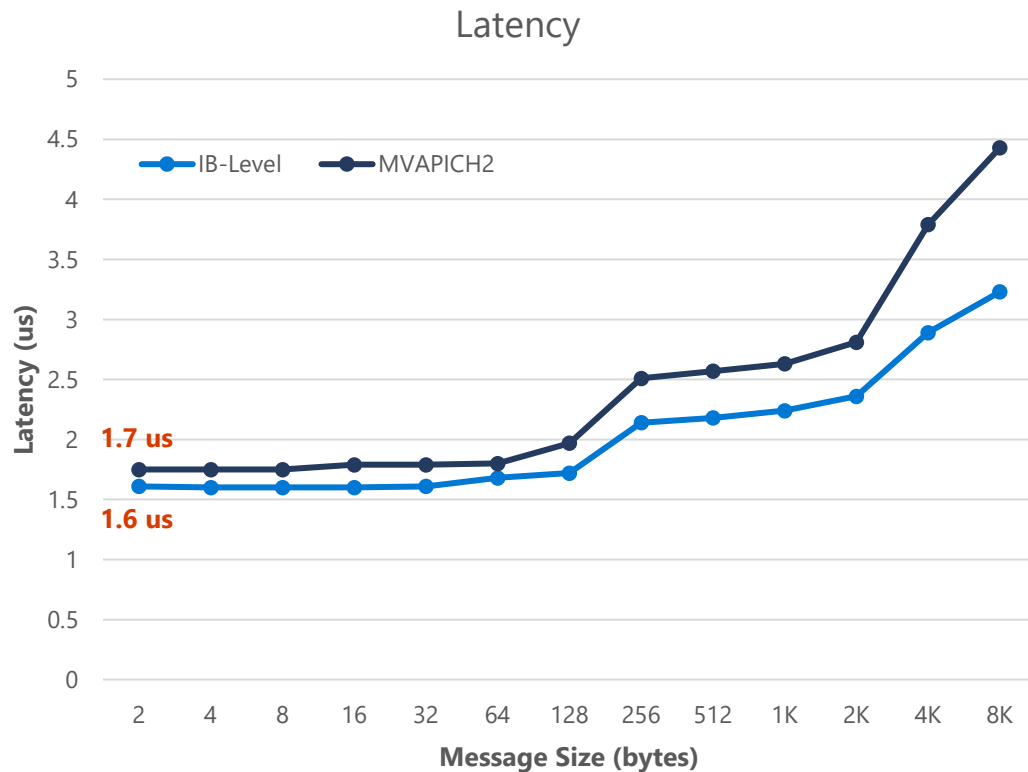
ALIAS	Description
-----	-----
ldst-count	Number of load/stores
ld-count	Number of loads
st-count	Number of stores
cache-hitm	Number of loads serviced by modified cache
lcl-cache-hitm	Number of loads serviced by modified cache on the local node
rmt-cache-hitm	Number of loads serviced by modified cache 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

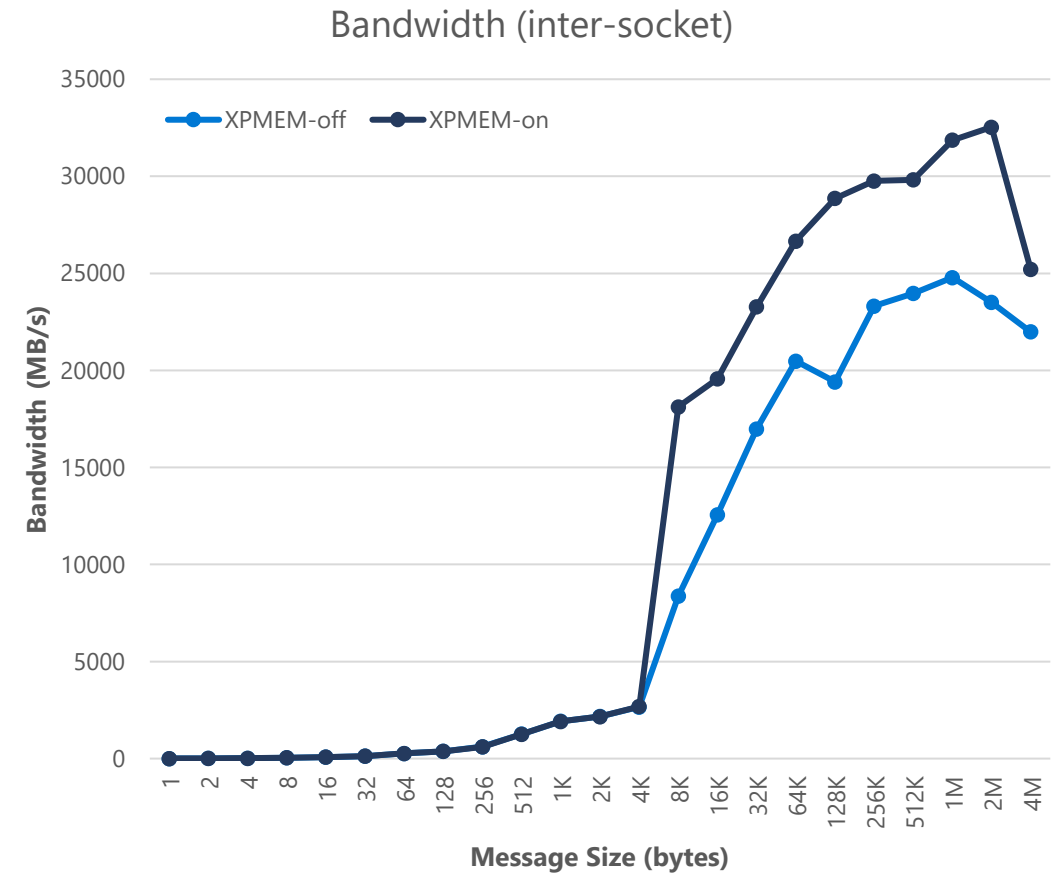
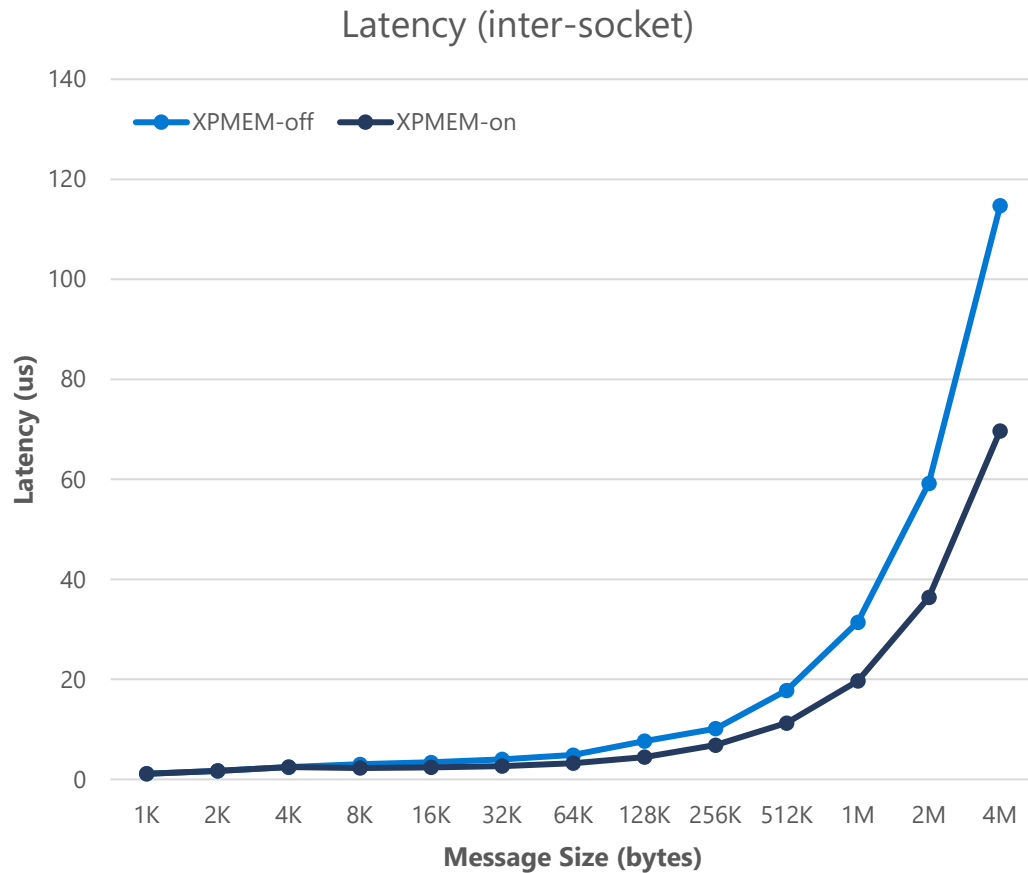


MVAPICH2-Azure on HBv3 (inter-node)



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)

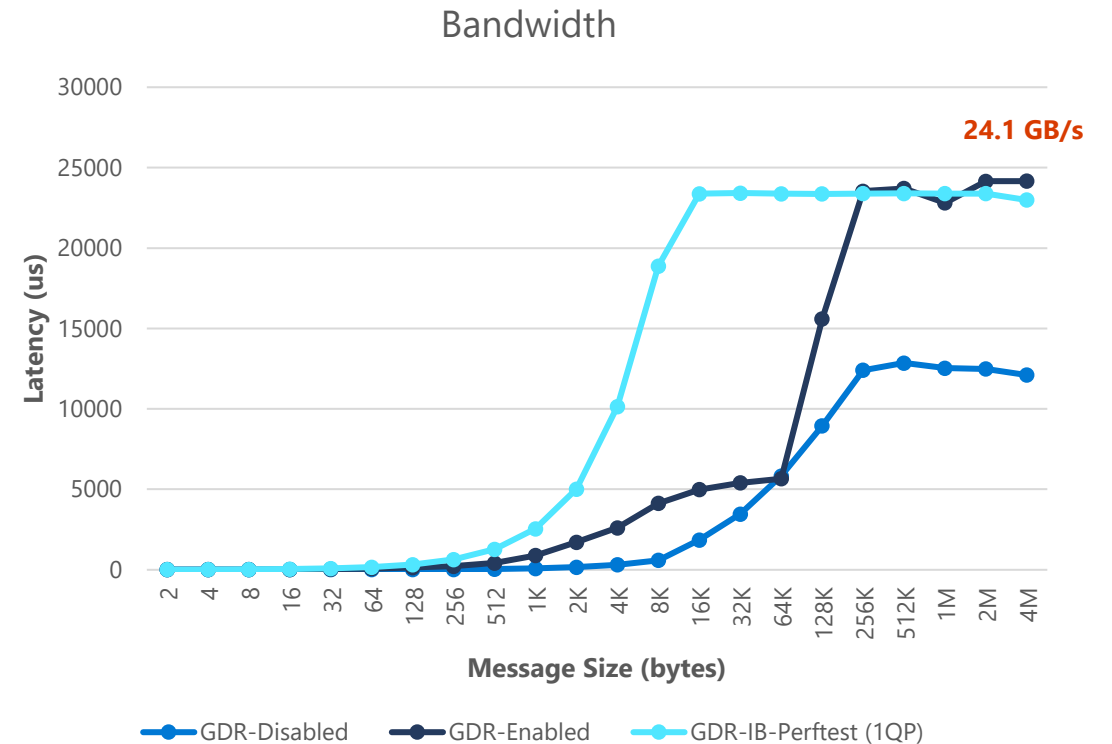
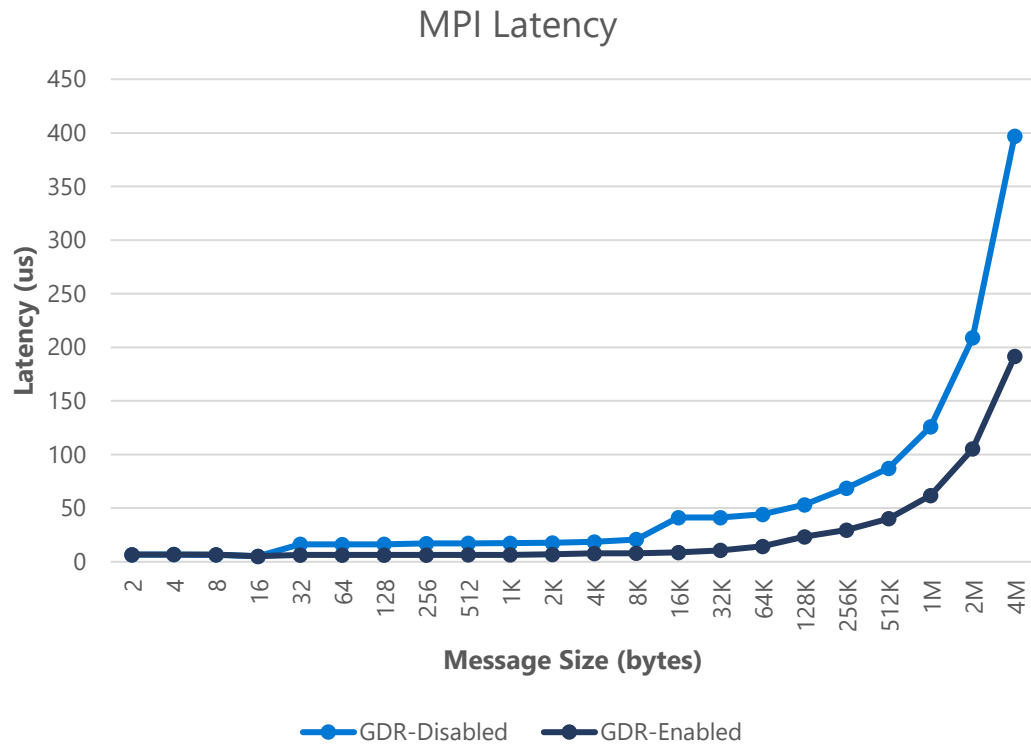


Agenda

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



MVAPICH2-GDR on NDv4



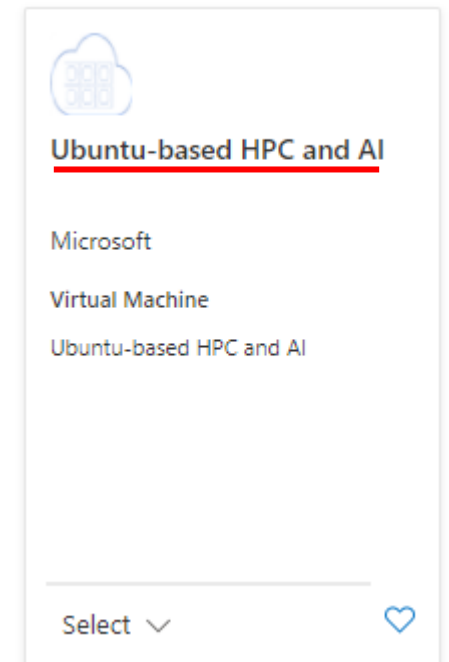
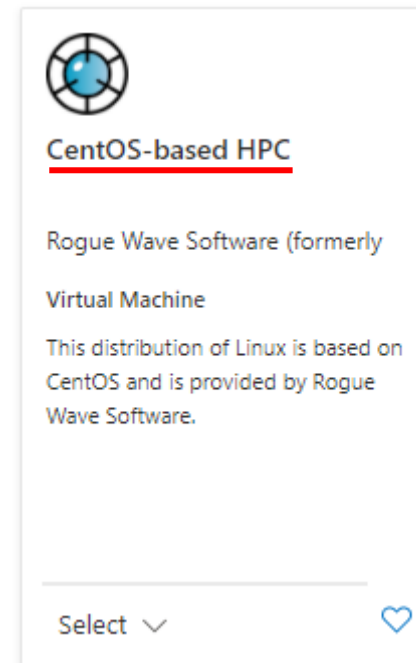
Environment parameters: MV2_NUM_QP_PER_PORT=4 MV2_IBA_EAGER_THRESHOLD=66560 MV2_VBUF_TOTAL_SIZE=66560 MV2_RNDV_PROTOCOL=RPWT

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

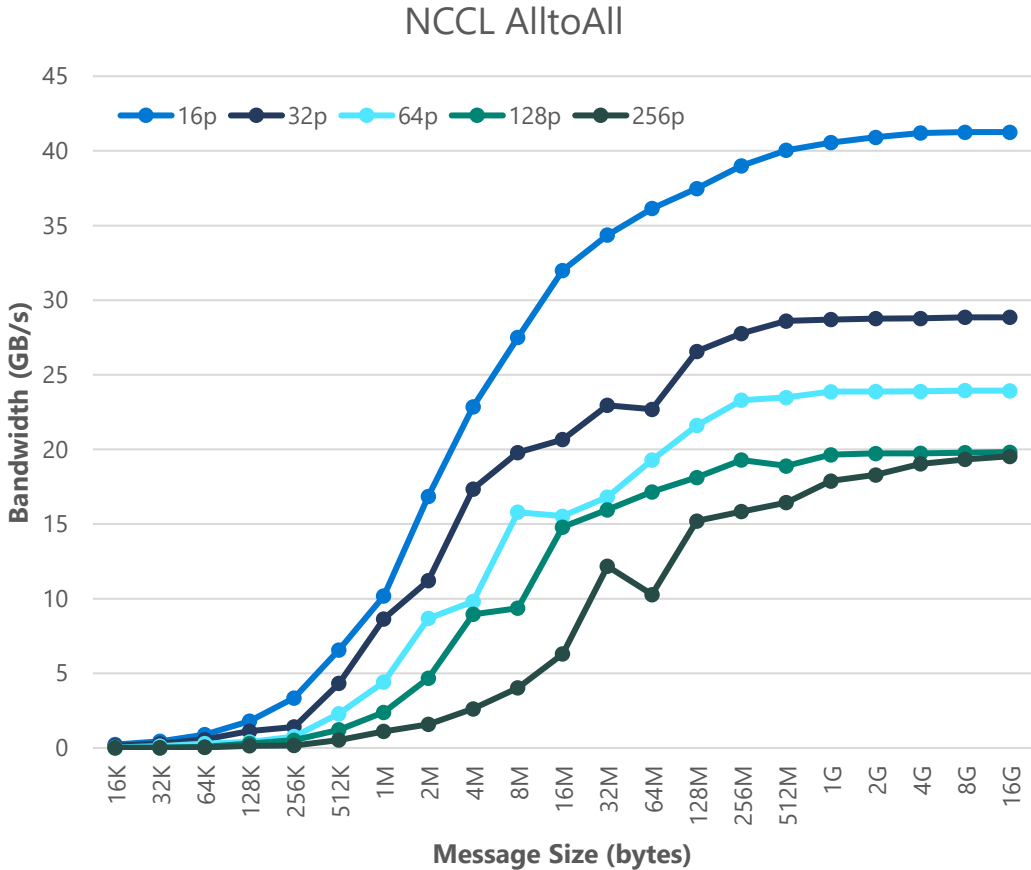
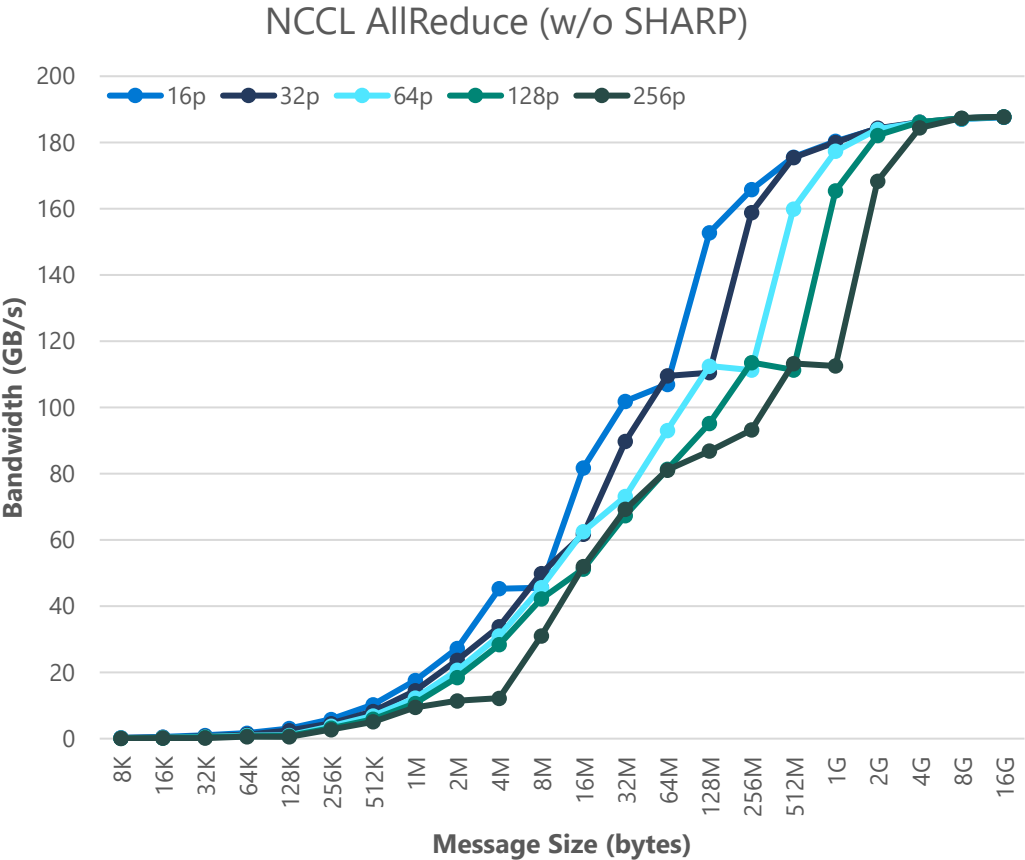


Agenda

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

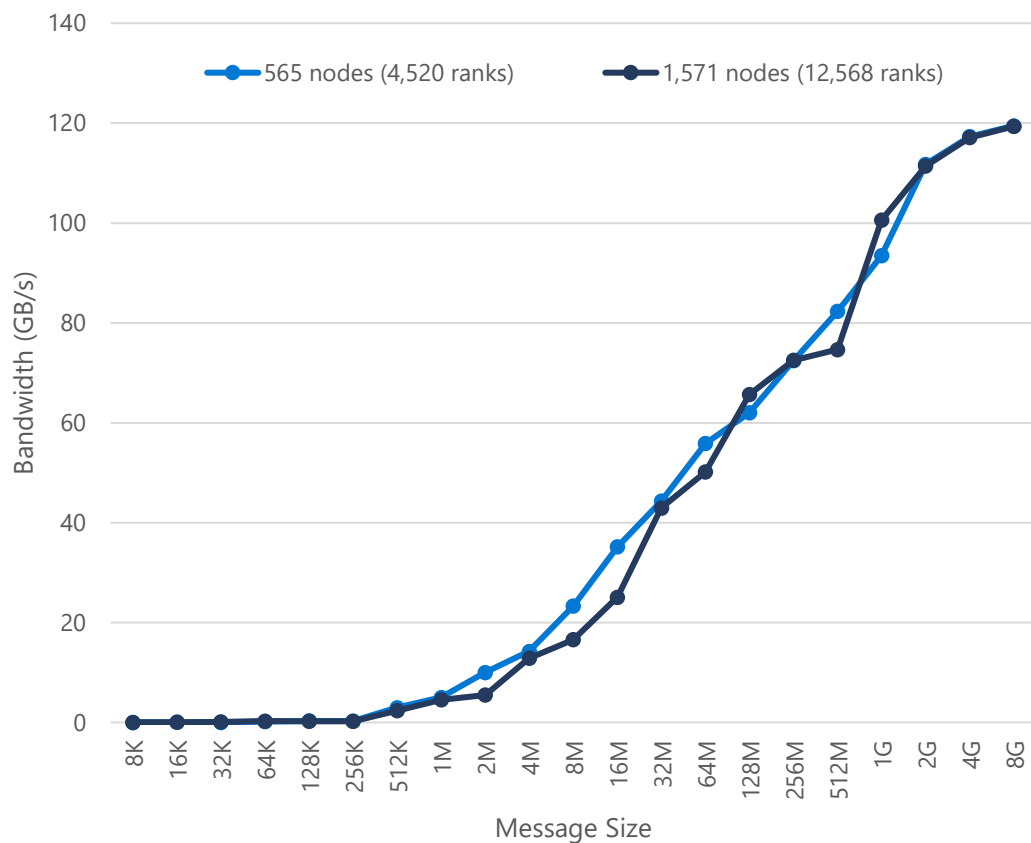


NCCL on NDv4

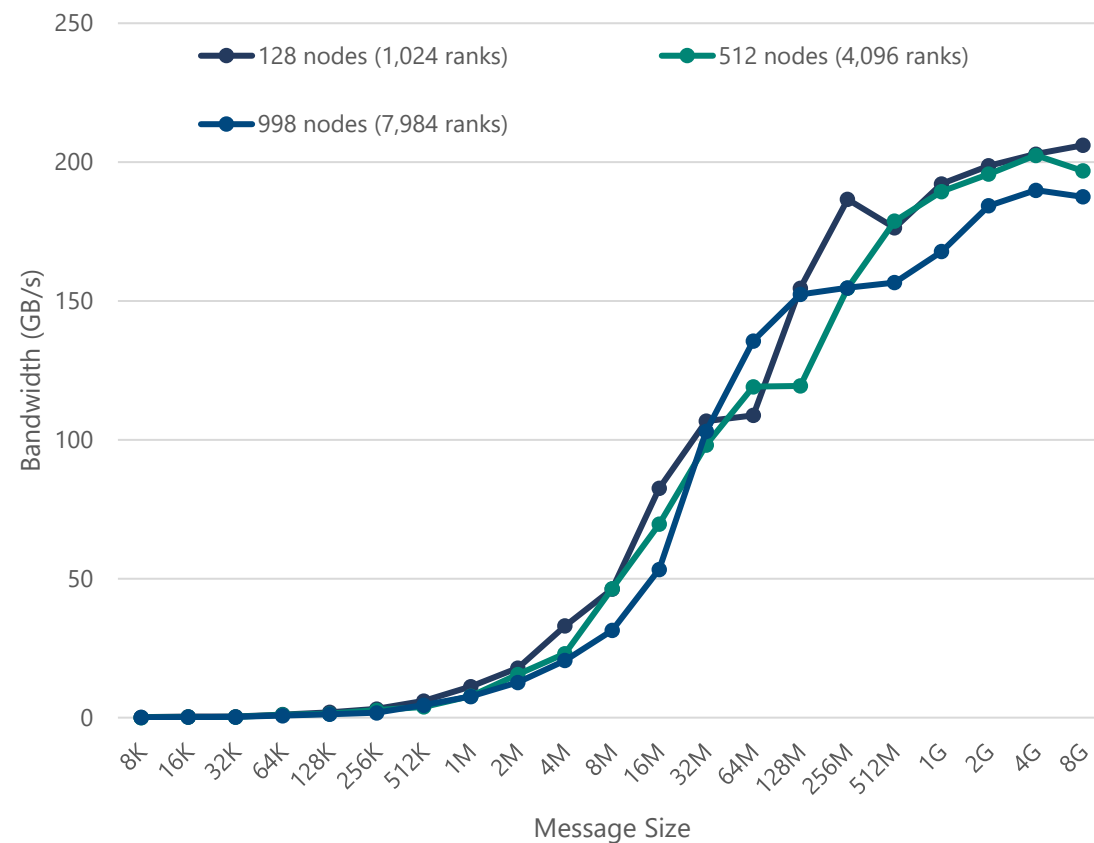


NCCL at Scale on NDv4

NCCL AllReduce (w/o SHARP)

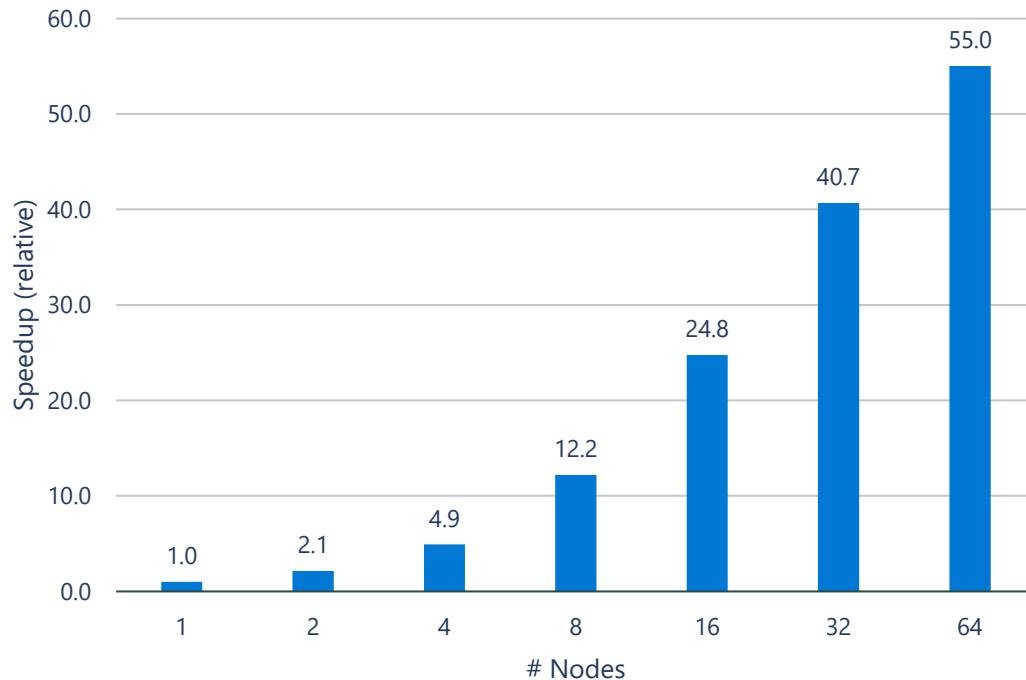


NCCL AllReduce w/ SHARP



HBv3 Performance Advantage

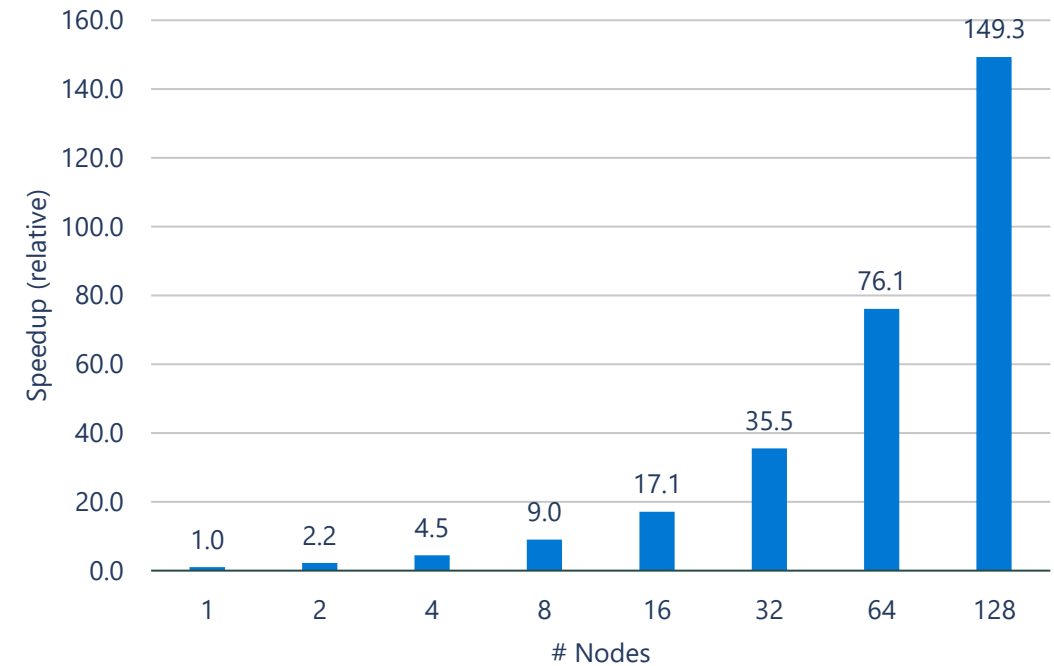
OpenFOAM



OpenFOAM

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

Fluent



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 Series

- [Azure VM sizes - HPC - Azure Virtual Machines](#)

GPU VM Series

- [Azure VM sizes - GPU - Azure Virtual Machines](#)

HPC VM Images

- [Azure HPC VM Images](#)
- [GitHub Repository](#)

HPC VM Deployment

- [Sample HPC VM deployment scripts](#)
- [Azure CycleCloud](#)
- [MUG '20 Tutorial](#)

Azure HPC Blogs

- [Azure Compute - Microsoft Tech Community](#)

We are Hiring!

jjjos@microsoft.com





Thank you