Microsoft

MVAPICH2 on Azure HPC: A seamless HPC Cloud Experience

Jithin Jose, Microsoft

MVAPICH User Group Meeting 2020

Agenda

- ✓ **Overview of Azure HPC**
- ✓ What's unique in HPC Cloud
- ✓ HPC Software Ecosystem
  - ✓ MVAPICH2-X Azure
- ✓ Performance Characteristics
- ✓ Conclusion

Microsoft Azure HPC

High Speed Networking

- InfiniBand Network
- Supports OFA verbs and all IB-based MPI Libraries
- Only public cloud to offer IB

Powerful Compute

- Compute Optimized SKUs
- GPU SKUs

Seamless Integration

- Seamless integration with existing HPC environments
- Scale out to Cloud

HPC Offerings in Azure

H-Series (InfiniBand)

- H16r (FDR)
- HB60rs (EDR)
- HC44rs (EDR)
- **HB120rs_v2** (HDR)

N-Series (GPU + InfiniBand)*

- NC24r (2 x Tesla K80 + FDR)
- NC24rs_v2 (4 x Tesla P100 + FDR)
- NC24rs_v3 (4 x Tesla V100 + FDR)
- ND24rs (4 x Tesla P40 + FDR)
- **ND40rs_v2** (8 x Tesla V100, EDR)

- SKU Name indicates core count
- "r" indicates RDMA support
- "s" indicates Premium Storage support

*GPU-only sizes not listed

Outline

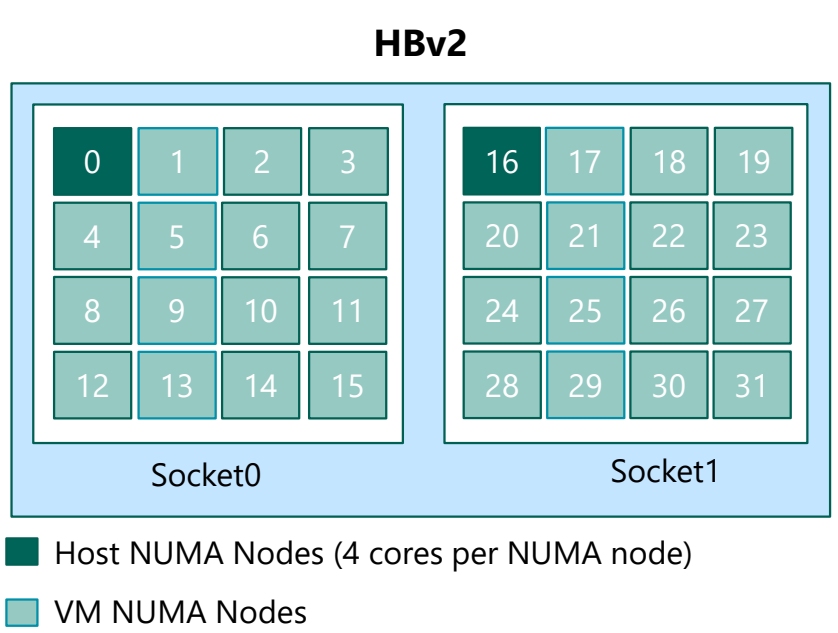
- ✓ Overview of Azure HPC
- ✓ **What's unique in HPC Cloud**
- ✓ HPC Software Ecosystem
- ✓ Performance Characteristics
- ✓ Conclusion

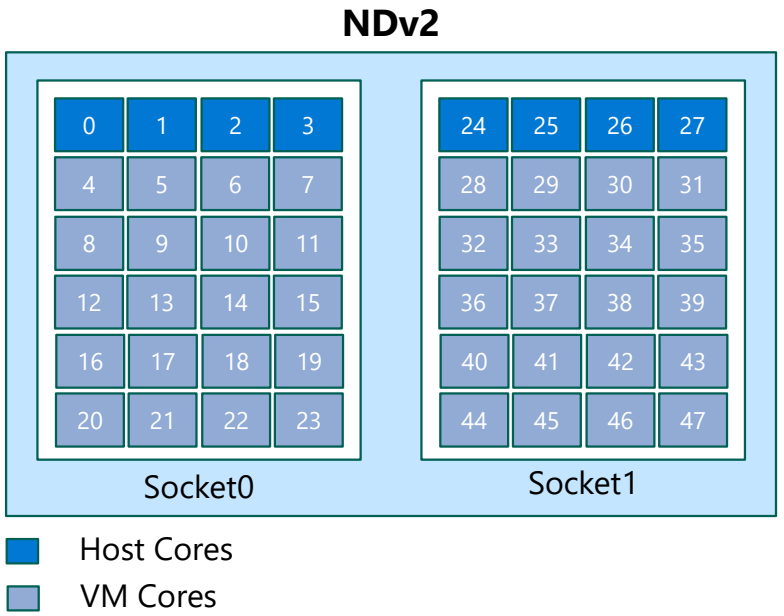
Unique Challenges for HPC in Cloud

Performance/Scalability Challenges

- Noise from Host OS / Host Agents
- Host interrupts
- Core/NUMA mapping
- Traffic from other customers
- Guest Agents

Host / VM Partitioning





- Host: AMD Rome with 128 cores
  - 32 NUMA nodes, 4 cores per NUMA node
- HBv2 VM
  - 120 cores for VM
  - 8 cores reserved for host
  - L3 = NUMA
  - No cache pollution

- Intel Xeon (Skylake) – 48 cores
  - 40 cores for VM
  - 8 cores reserved for host

✔ MinRoot: Shapes up the Host cores ✔ CPUGroups: Shapes up the VM cores ✔ No Hyperthreading ✔ One VM per Host

Efficient Network Virtualization

- Single Root I/O Virtualization (SR-IOV)
  - Expose all NIC features w/o any host intervention
  - Offers bare-metal network performance
- Single VM per host
  - One VF per Host

Network Features:

- Dynamically Connected Transport (DCT)
  - Reliable and scalable transport
  - Lesser Memory footprint
- Hardware collectives (hcoll)
  - Collectives offload framework
  - Asynchronous execution
  - Supports blocking/non-blocking collectives
- UD multicast (MCAST)
  - Unreliable datagram (UD) based multicast
  - Create a mcast group and broadcast
- Hardware Tag Matching
- Reliability/Congestion Control
  - SHIELD, Adaptive Routing

Network Security

- InfiniBand Partition Keys
  - Only VMs with same partition keys can communicate w/ each other
  - Isolates customer traffic
  - Multiple SL's possible within same PKEY

- Partition Keys in Azure
  - Single PKEY for all VMs in a VMSS (Virtual Machine Scale Set)
  - Single PKEY for all VMs associated with an Availability Set

  - Check PKEY:

```
$ cat /sys/class/infiniband/mlx5_0/ports/1/pkeys/0
$ 0x801a
```

Congestion Control

- Azure HPC InfiniBand Network is non-blocking
  - No oversubscription
- Static Routing may still cause bottlenecks
- Solution: Adaptive Routing
  - Available on CX5 and later generation NICs
  - Configured per SL, AR enabled on all SLs > 0

NUMA Mapping

- Deterministic pNUMA-vNUMA mapping
- Distance map shows 1:1 mapping
- Enables NUMA aware designs
  - Efficient Process mapping
  - NUMA aware MPI collectives

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
0	83.8	83.9	84	122	122	122	122	133	133	133	133	134	134	134	135	229	229	229	236	236	236	236	230	230	230	230	231	231	231	231
1	84.7	84.6	84.5	124	124	124	124	131	131	131	131	135	135	135	135	229	228	230	237	236	237	236	231	231	231	231	232	232	232	232
2	82.9	82.8	82.9	121	121	121	121	132	132	132	132	134	134	134	134	229	228	228	236	235	235	236	230	230	230	230	231	231	230	230
3	120	120	120	84.7	84.8	84.7	84.7	130	135	135	135	131	131	131	131	235	237	236	242	242	241	241	235	235	236	234	237	238	237	237
4	123	123	123	85.9	85.9	85.9	86	136	135	136	136	130	124	130	129	234	235	234	242	242	242	241	236	235	236	235	237	237	236	237
5	120	120	120	84.7	84.8	84.6	84.7	130	135	135	135	131	131	131	119	235	235	235	242	242	241	241	235	236	236	236	237	237	237	237
6	113	120	120	84.2	84.3	84.2	84.2	135	134	130	134	130	129	130	130	237	237	237	242	242	242	241	236	236	236	236	237	236	236	238
7	131	131	131	134	135	135	135	85	85	84.9	84.9	121	121	121	121	223	222	223	231	231	230	231	233	233	233	233	234	234	233	234
8	131	131	131	135	135	135	135	85.2	85.2	85.3	85.4	121	121	121	121	223	223	222	230	230	232	231	231	233	232	232	234	233	233	233
9	132	132	131	135	135	135	135	87.4	87.4	87.3	87.3	122	122	118	122	223	208	222	231	230	231	231	234	233	233	234	229	228	235	235
10	130	129	124	137	133	137	137	85.5	85.6	85.5	85.7	123	123	123	123	223	223	224	232	232	232	233	233	234	234	233	234	235	235	234
11	135	135	135	132	132	132	132	124	124	124	124	85.6	85.6	85.5	85.6	231	230	230	238	239	239	238	241	240	240	240	242	242	242	241
12	135	135	135	132	132	132	132	124	124	124	124	85	85	85.1	85.2	231	228	229	238	237	237	239	240	240	240	240	242	241	241	241
13	135	135	135	132	132	132	132	124	124	124	124	85.1	85.1	85	85.1	229	231	230	237	238	237	238	239	240	240	240	241	241	242	241
14	136	136	136	132	132	132	132	124	124	124	124	85.6	85.6	85.6	85.6	231	230	231	238	239	239	238	240	241	241	240	242	243	241	243
15	229	229	228	235	237	235	234	230	229	229	229	230	230	230	231	84.2	84.3	84.2	124	124	124	124	131	131	131	131	135	135	135	135
16	228	228	227	235	235	236	236	229	229	229	230	229	231	231	230	84.7	84.7	84.8	125	125	125	124	131	132	132	132	136	136	135	136
17	228	229	228	236	235	234	236	229	229	229	229	231	229	230	231	84.8	84.8	84.8	125	124	125	124	131	131	131	131	135	135	135	135
18	240	239	240	244	245	244	245	236	235	236	236	237	237	237	238	124	124	123	86.8	86.7	86.7	86.7	138	138	138	138	131	131	131	131
19	239	239	240	246	245	245	245	236	237	235	236	238	237	237	238	123	123	123	86.6	86.7	86.7	86.7	138	138	138	138	131	131	131	131
20	239	239	238	243	243	244	244	235	235	235	236	237	237	237	236	123	123	123	86.2	86.1	86.2	86.2	137	137	137	137	130	130	130	130
21	240	238	238	243	243	243	244	235	235	235	234	237	236	236	236	123	123	123	86.2	86.3	86.3	86.2	137	137	137	137	130	130	130	130
22	223	223	223	227	227	227	228	233	232	233	233	233	234	234	234	131	131	131	135	135	135	135	84.9	84.7	84.7	84.7	120	120	120	120
23	224	223	224	229	228	228	228	234	234	233	232	235	236	235	234	131	131	131	135	135	135	135	85.5	85.6	85.6	85.5	121	120	121	121
24	224	224	223	229	219	229	229	234	227	234	234	236	235	235	235	131	131	131	138	138	138	138	86.9	86.9	86.9	86.9	124	124	124	124
25	223	224	224	229	228	228	229	234	233	233	233	234	234	234	227	124	130	130	137	137	137	137	86	86.1	86.1	86	123	118	123	114
26	229	229	230	239	239	239	237	239	238	239	239	241	241	241	242	134	133	133	132	132	132	132	122	122	121	122	83.7	83.6	83.6	83.6
27	226	230	228	238	237	239	238	239	239	239	239	241	241	241	241	133	133	133	132	132	132	131	121	116	121	121	83.3	83.2	83.4	83.3
28	230	231	232	240	241	241	242	241	240	241	241	241	242	242	243	135	135	135	132	132	132	132	125	125	125	125	85.2	85.1	85	85.1
29	233	234	234	243	242	242	241	242	241	241	240	243	242	242	242	136	136	136	133	133	133	133	125	125	125	125	86.1	86.2	86.2	86.1

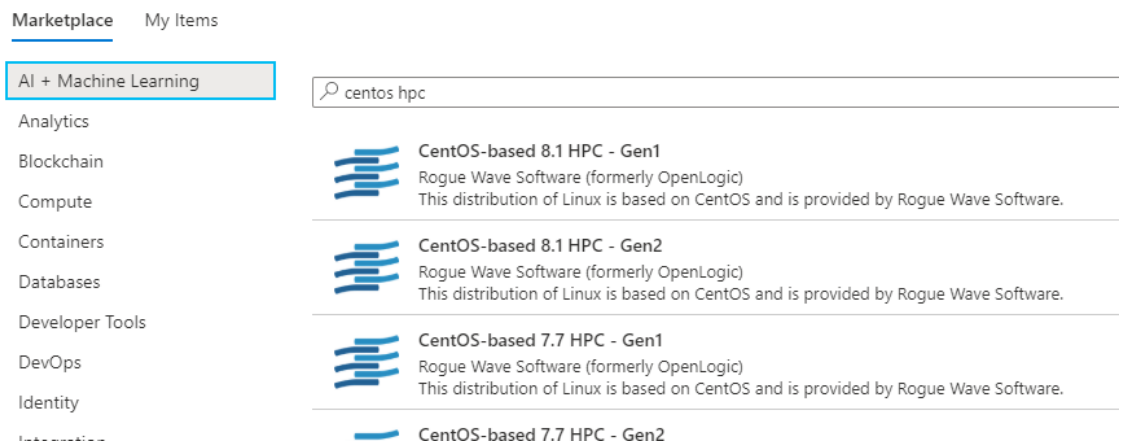
MLC Latency Matrix on HBv2 (ns)

Outline

- ✓ Overview of Azure HPC
- ✓ What's unique in HPC Cloud
- ✓ **HPC Software Ecosystem**
  - ✓ MVAPICH2-X Azure
- ✓ Performance Characteristics
- ✓ Conclusion

HPC Marketplace Images

- CentOS HPC Images
  - Mellanox OFED
  - MPI Libraries
    - Includes **MVAPICH2, MVAPICH2X-Azure**
  - HPC Libraries
  - Optimization Configurations



- OpenSource GitHub repository
  - https://github.com/Azure/azhpc-images/

- Use pre-built HPC VM images, or build custom image based on these, or BYO software stack

MVAPICH2-X Azure

- Available in all Azure CentOS-HPC images
- Targeted for Azure HB, HBv2, HC VM instances
- Feature Highlights:
  - Enhanced tuning for point-to-point and collectives
  - XPMEM Support
  - DC Support
  - Co-operative Protocol
  - Hybrid RC/UD Support

Blog Post: https://techcommunity.microsoft.com/t5/azure-compute/mvapich2-on-azure-hpc-clusters/ba-p/1404305

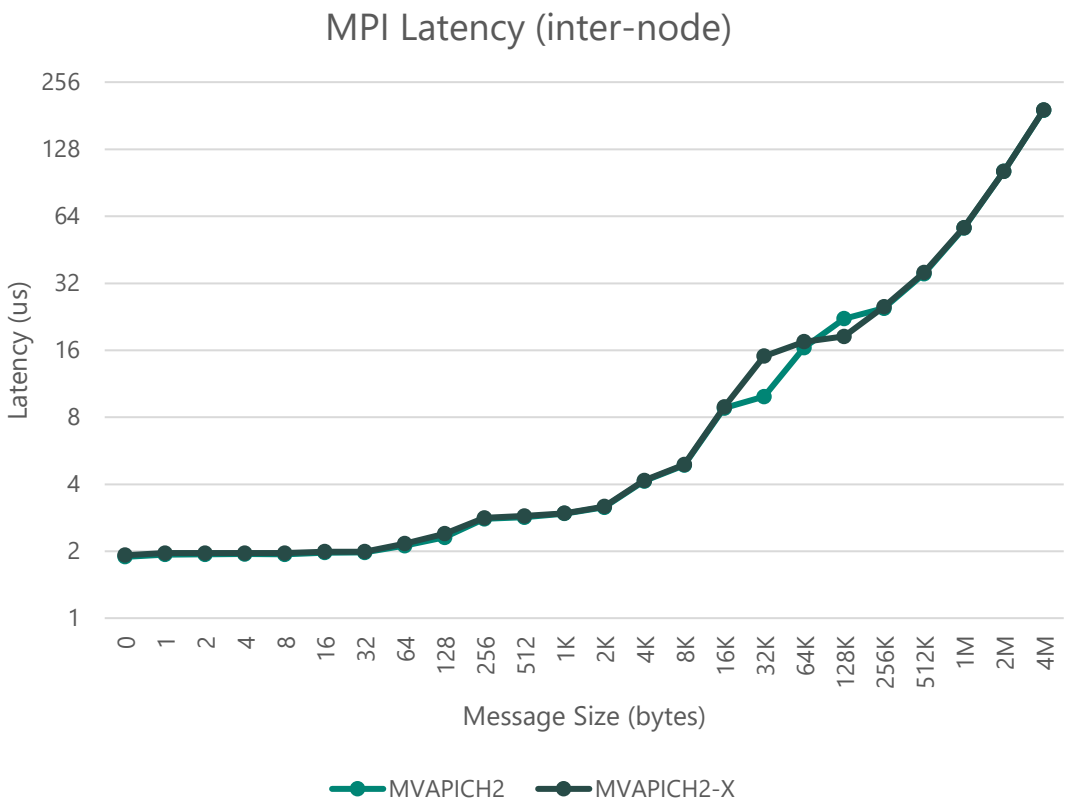
Outline

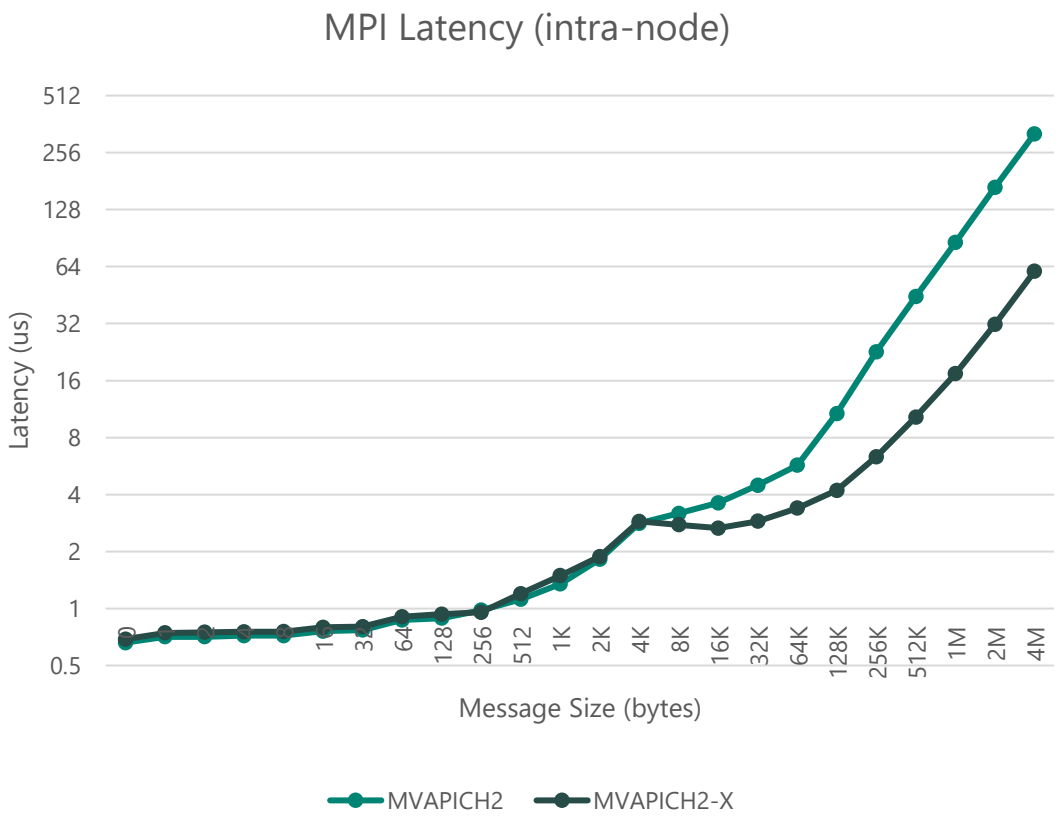
- ✓ Overview of Azure HPC
- ✓ What's unique in HPC Cloud
- ✓ HPC Software Ecosystem
- ✓ **Performance Characteristics**
- ✓ Conclusion

Experiment Setup

- HBv2 VMs
- CentOS 7.7 HPC Image
- MPI Libraries
  - MVAPICH2 2.3.4
  - MVAPICH2-X 2.3
- Mellanox OFED 5.1

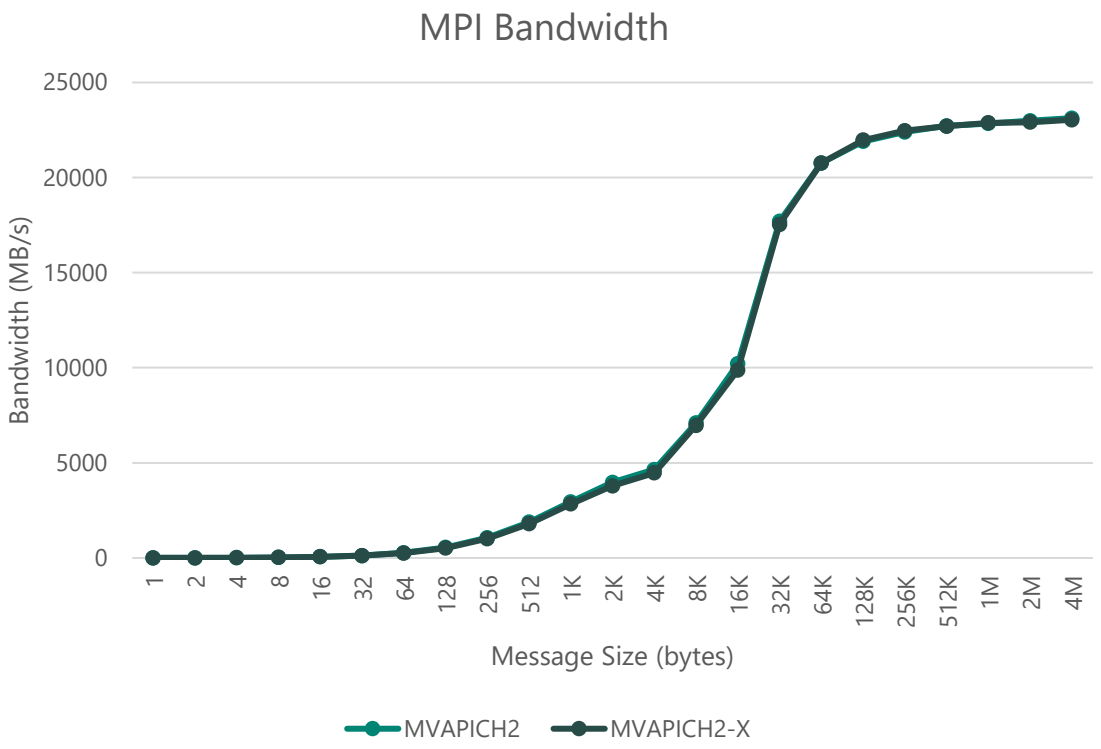
MPI Latency

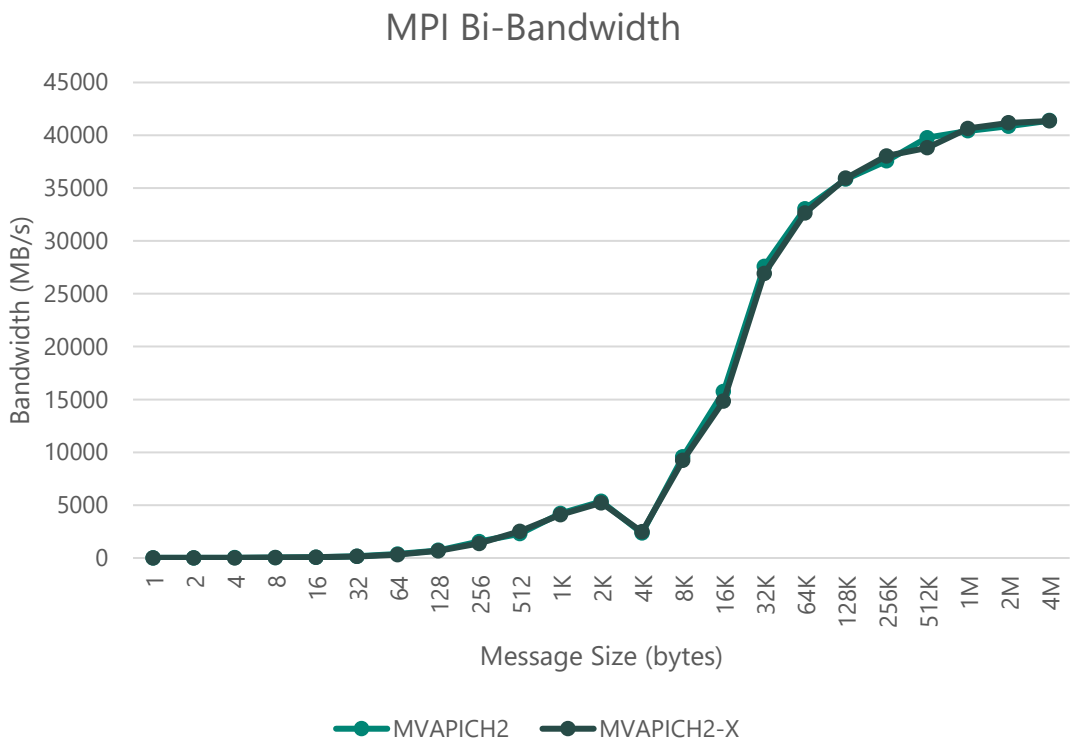




- MVAPICH2, MVAPICH2-X achieves < 2us latencies
- MVAPICH2-X offers better large message latencies for intra-node transfers (XPMEM)

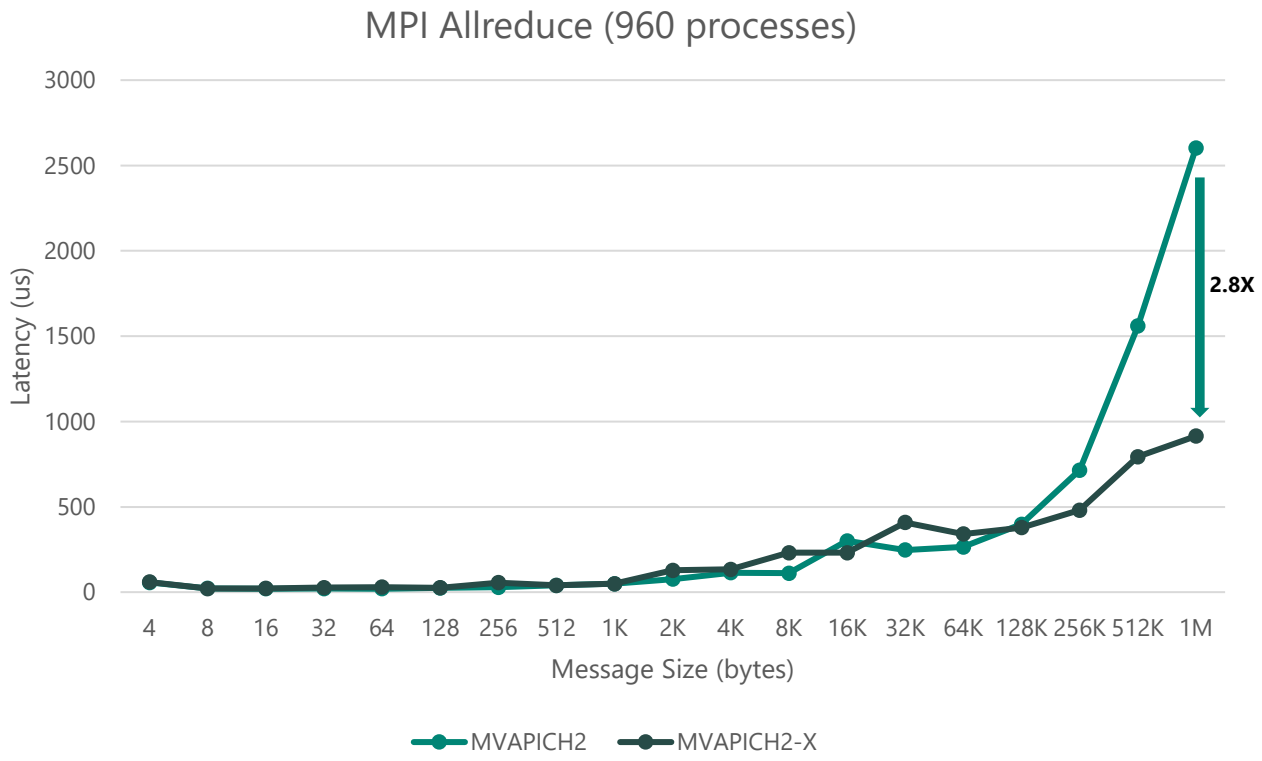
MPI Bandwidth / Bi-Bandwidth





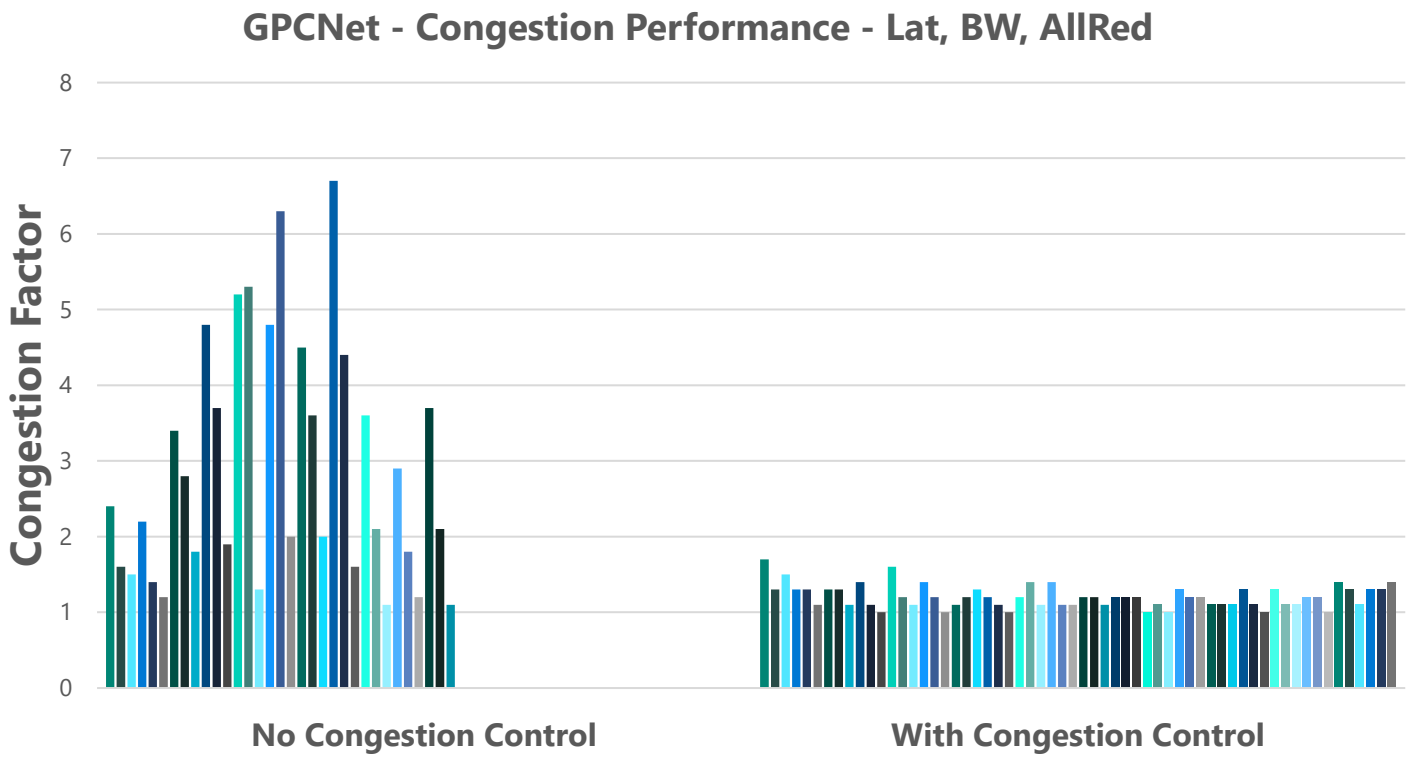
- MVAPICH2, MVAPICH2-X close to line rates
- Both versions use same protocols

MPI Allreduce



- MVAPICH2-X XPMEM Collectives offers better large message allreduce latencies
- 8 HBv2 nodes, 120 PPN

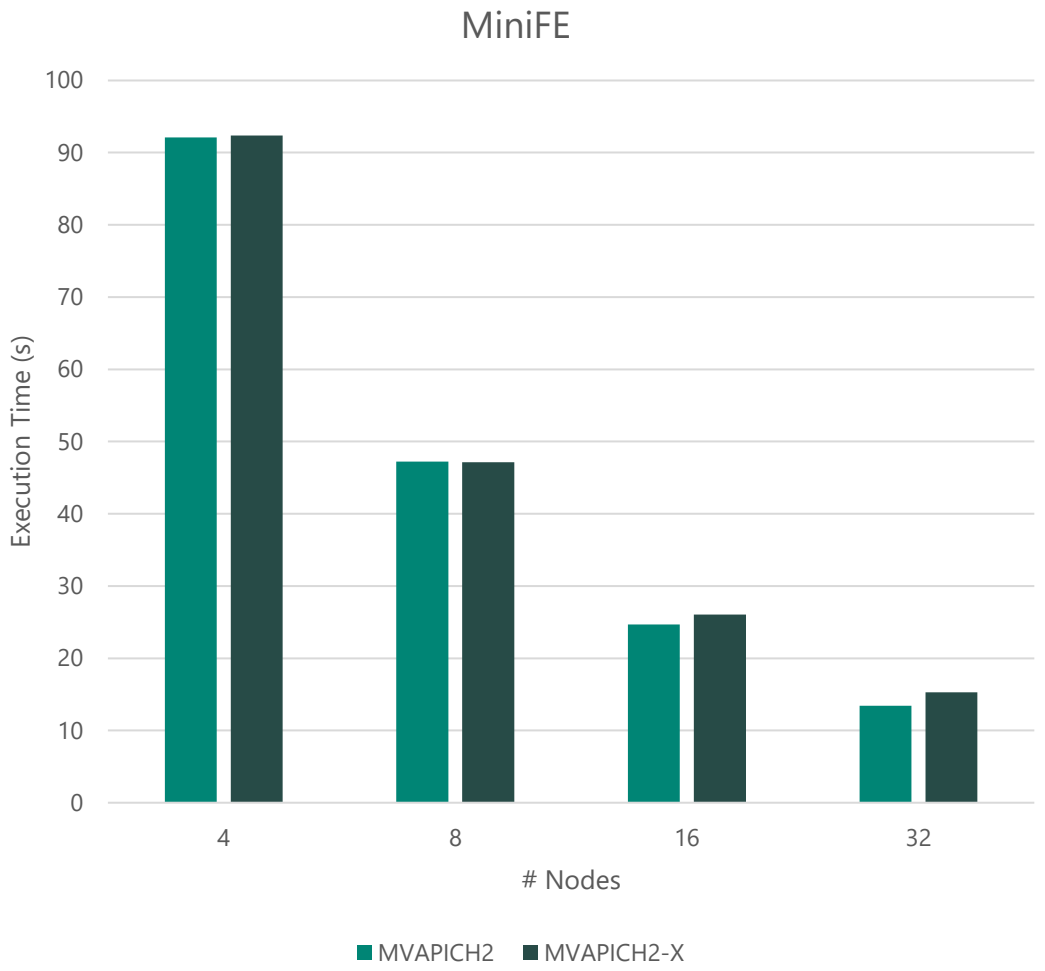
GPCNet on HBv2



- Measure Congestion Factor with and without Congestion Control (CC)
- 128 HBv2 VMs, 120 PPN (15,360 MPI ranks)
- Upcoming F/W version

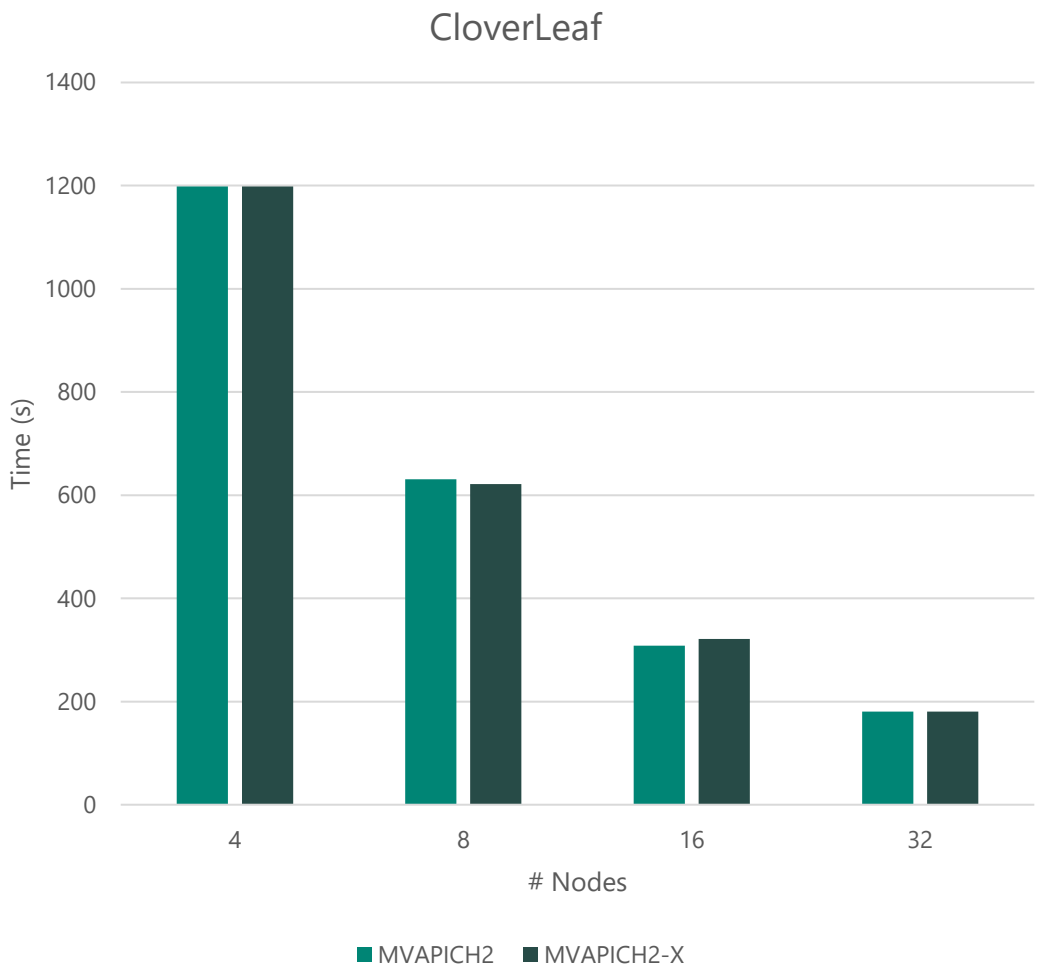
MiniFE

- Finite Element Mini-Application
- Proxy application for unstructured implicit FE codes
- Strong scaling experiment
- Version: openmp-opt
- Problem Size
  - nx=1024, ny=1024, nz=1024



CloverLeaf

- Hydrodynamics mini0app to solve compressible Euler equations in 2D
- Version: CloverLeaf_MPI
- DataSet: clover_bm256.in
  - x_cells: 15360, y_cells: 15360
  - Steps: 2955



WRF

- WRF 3.6
  - https://github.com/hanschen/WRFV3
- Benchmark: 12km resolution case over the Continental U.S. (CONUS) domain
  - https://www2.mmm.ucar.edu/wrf/WG2/benchv3/#_Toc212961288
- Update io_form_history in namelist.input to 102
  - https://www2.mmm.ucar.edu/wrf/users/namelist_best_prac_wrf.html#io_form_history

* Courtesy: MVAPICH Team

Outline

- ✓ Overview of Azure HPC
- ✓ What's unique in HPC Cloud
- ✓ HPC Software Ecosystem
- ✓ Performance Characteristics
- ✓ **Conclusion**

Conclusion

- Azure HPC design offers bare-metal performance
- SR-IOV efficiently exposes network features
- Out-of-the box HPC VM Images
  - MVAPICH2, MVAPICH2-X
- MVAPICH2/MVAPICH2-X offers great performance and scalability on Azure

Pointers

- AzureHPC Deployment Scripts
  - https://github.com/Azure/azurehpc
- Azure HPC/GPU VM Sizes
  - https://docs.microsoft.com/azure/virtual-machines/sizes-hpc
  - https://docs.microsoft.com/azure/virtual-machines/sizes-gpu
- HPC Marketplace Images
  - https://techcommunity.microsoft.com/t5/azure-compute/azure-hpc-vm-images/ba-p/977094
- MVAPICH2 on Azure
  - https://techcommunity.microsoft.com/t5/azure-compute/mvapich2-on-azure-hpc-clusters/ba-p/1404305
- Adaptive Routing on Azure HPC
  - https://techcommunity.microsoft.com/t5/azure-compute/adaptive-routing-on-azure-hpc/ba-p/1205217

Thank You!

jijos@microsoft.com