Microsoft

MVAPICH2 on Azure

OSU Booth, Supercomputing 2019

Jithin Jose
Microsoft (jijos@microsoft.com)

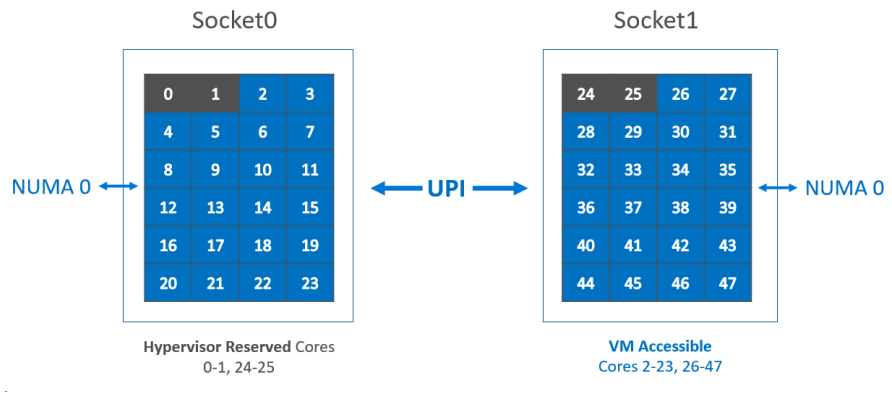
Outline

- Azure HPC Clusters – HB, HC, HBv2
  - How we shape up the SKUs
- Performance Highlights
  - Latency, Bandwidth – HB, HC, HBv2
  - Applications – MiniGhost, CloverLeaf
- BigData on Azure HPC
- MVAPICH2 Azure Release
- QuickDeploy MVAPICH2 to Azure

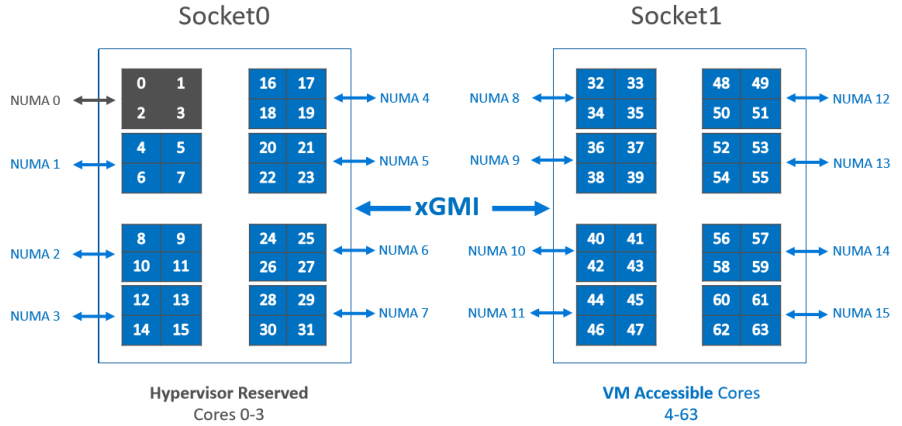
Azure HB, HC VM Series

	HB	HC	HBv2
CPU	AMD EPYC	Intel Xeon Platinum	AMD ROME
Cores / VM	60	44	120
Clock Speed	2.55 GHz	3.4 GHz	3.3 GHz
Memory Bandwidth	263 GB/sec	191 GB/sec	340 GB/sec
Memory	240 GB (4GB/core)	352 GB (8GB/core)	480 GB (4GB/core)
Local Disk	700 GB NVMe	700 GB NVMe	900 GB NVMe
InfiniBand Network	100 Gbps EDR w ConnectX-5 (SR-IOV)	100 Gbps EDR w ConnectX-5 (SR-IOV)	200 Gbps HDR w ConnectX-5 (SR-IOV)
Network	Gen2 SmartNIC (FPGA Accelerated)		

Core Partitioning in HB and HC VM Series

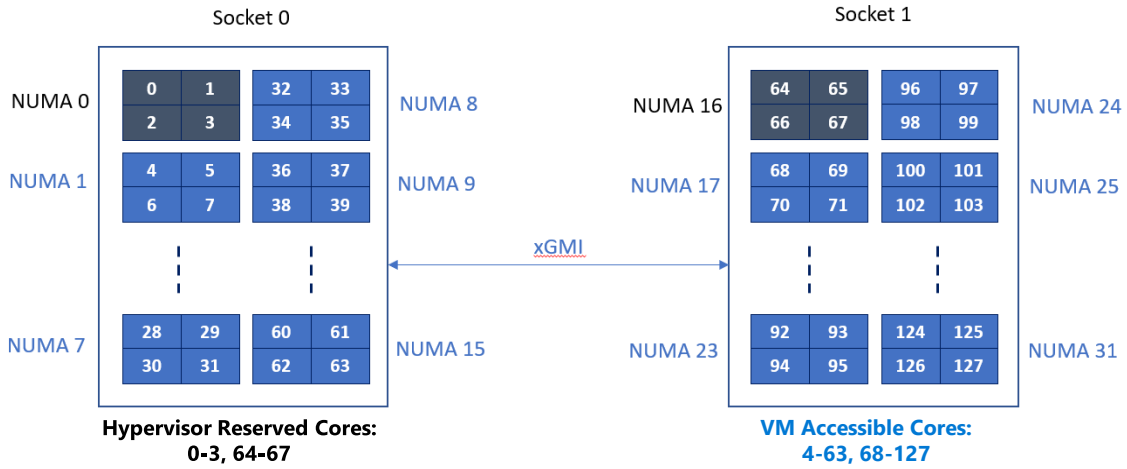


HC44rs Virtual Machines



HB60rs Virtual Machines

- Partition Host/Guest Resources
- MinRoot – Shape the Host
- CPUGroups – Shape the Guest
- No interference either direction!
- Exploring Memory Partitioning



HB120rs Virtual Machines

Network Features in HB, HC Series

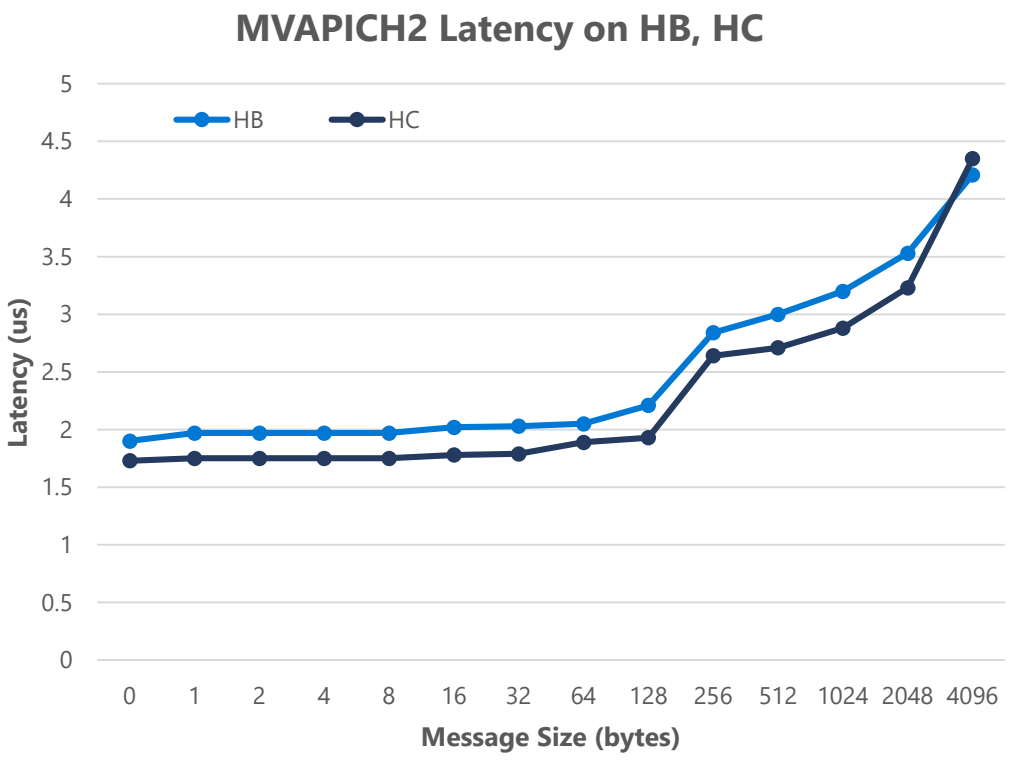
- **HB, HC:** 

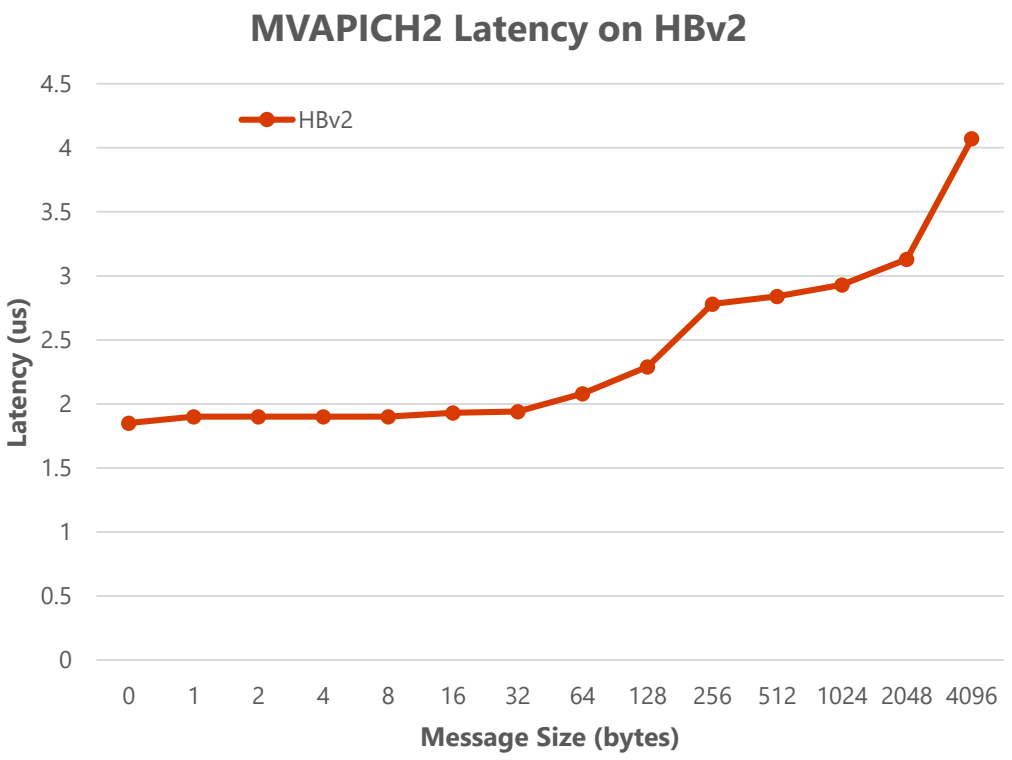
  - EDR 100Gb/s InfiniBand
  - Up to 200M messages/second
- **HBv2:** 

  - HDR 200Gb/s InfiniBand
  - Up to 215M messages/second

- 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
- Better Reliability
  - Adaptive Routing on all SLs, all transports
  - SHIELD detects link failures

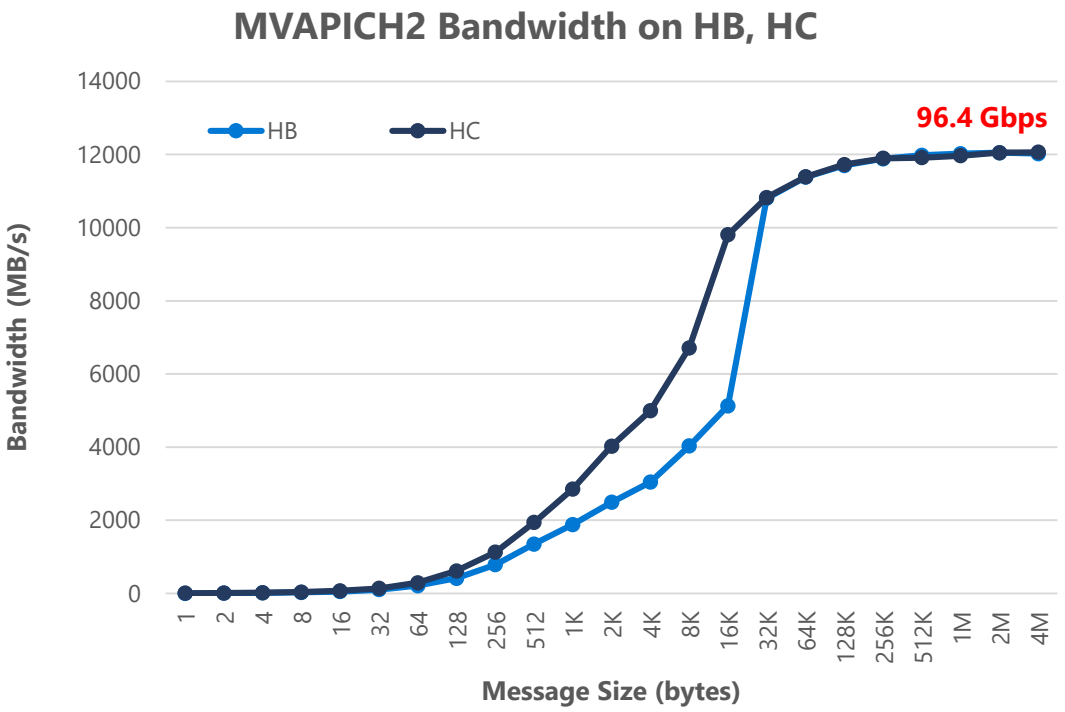
MVAPICH2 Latency

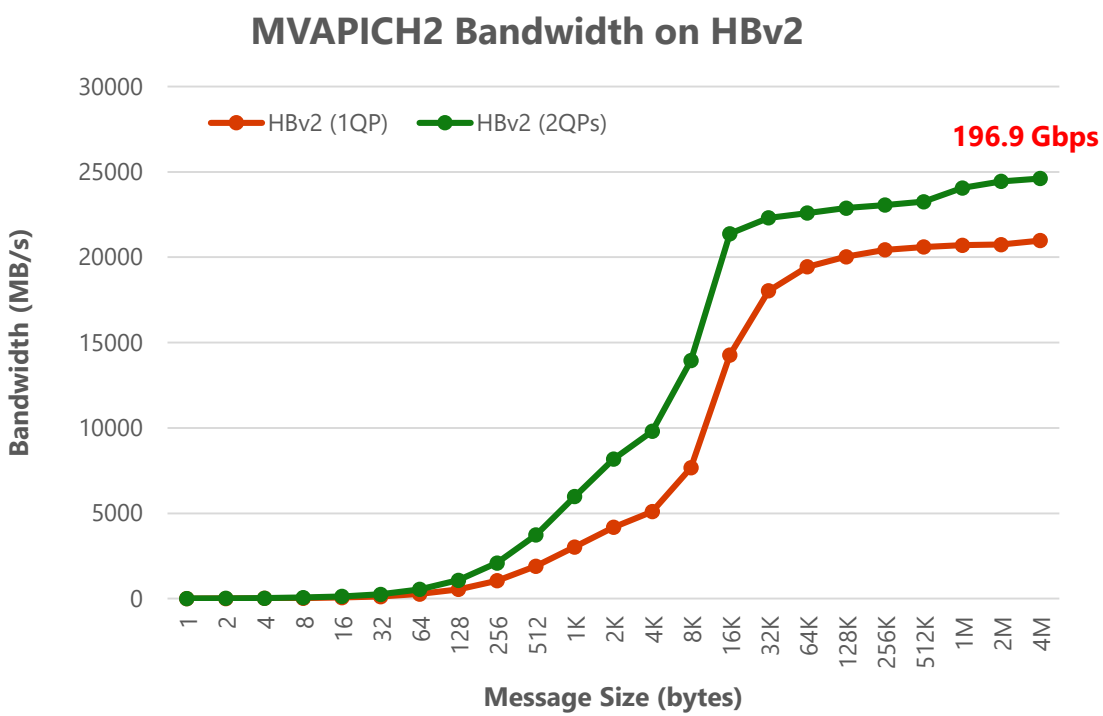




- Processes pinned closer to NIC
- 4 byte MPI level latency: HC – 1.75us, HB – 1.97us, HBv2 – 1.90us
- OSU Benchmarks 5.6.2

MVAPICH2 Bandwidth on HB/HC

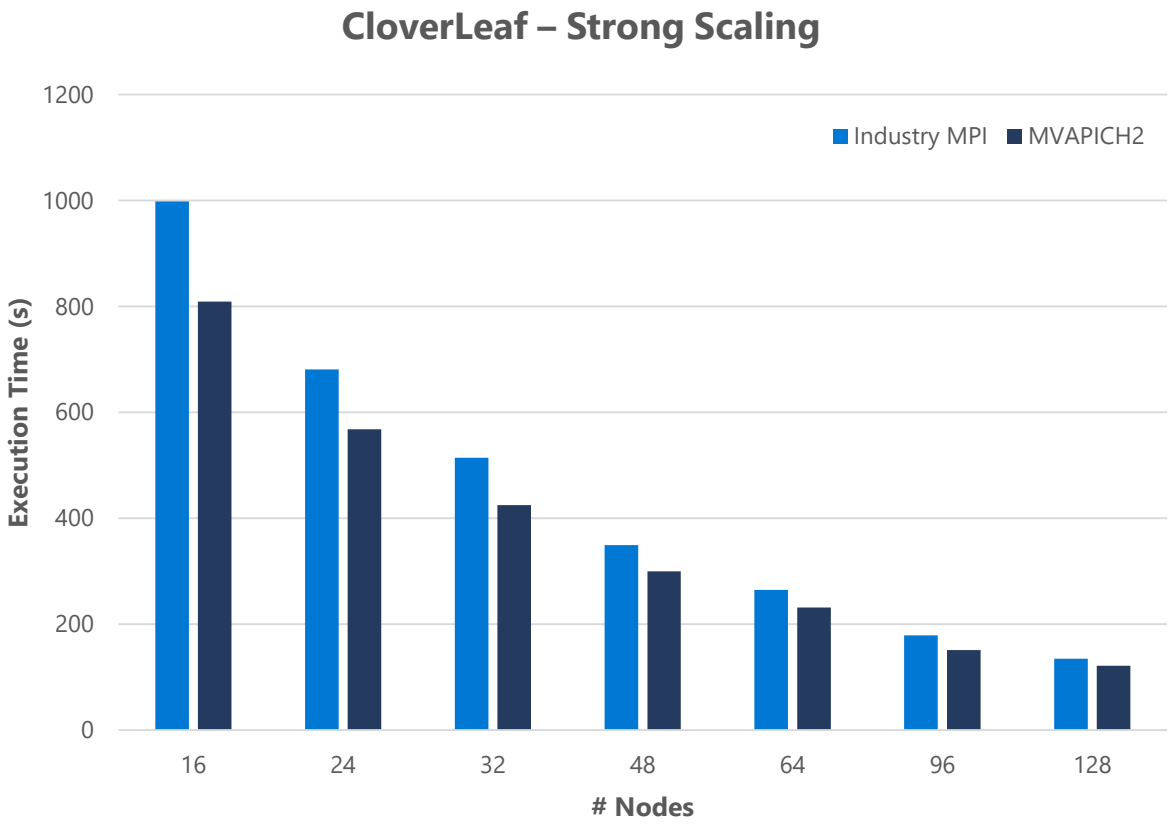




- Processes pinned closer to NIC
- MPI bandwidth: 12.05 GB/s on HB/HC, 24.6 GB/s on HBv2 (2 QPs)
- Single QP not able to saturate network bandwidth on HBv2
- OSU Benchmarks 5.6.2

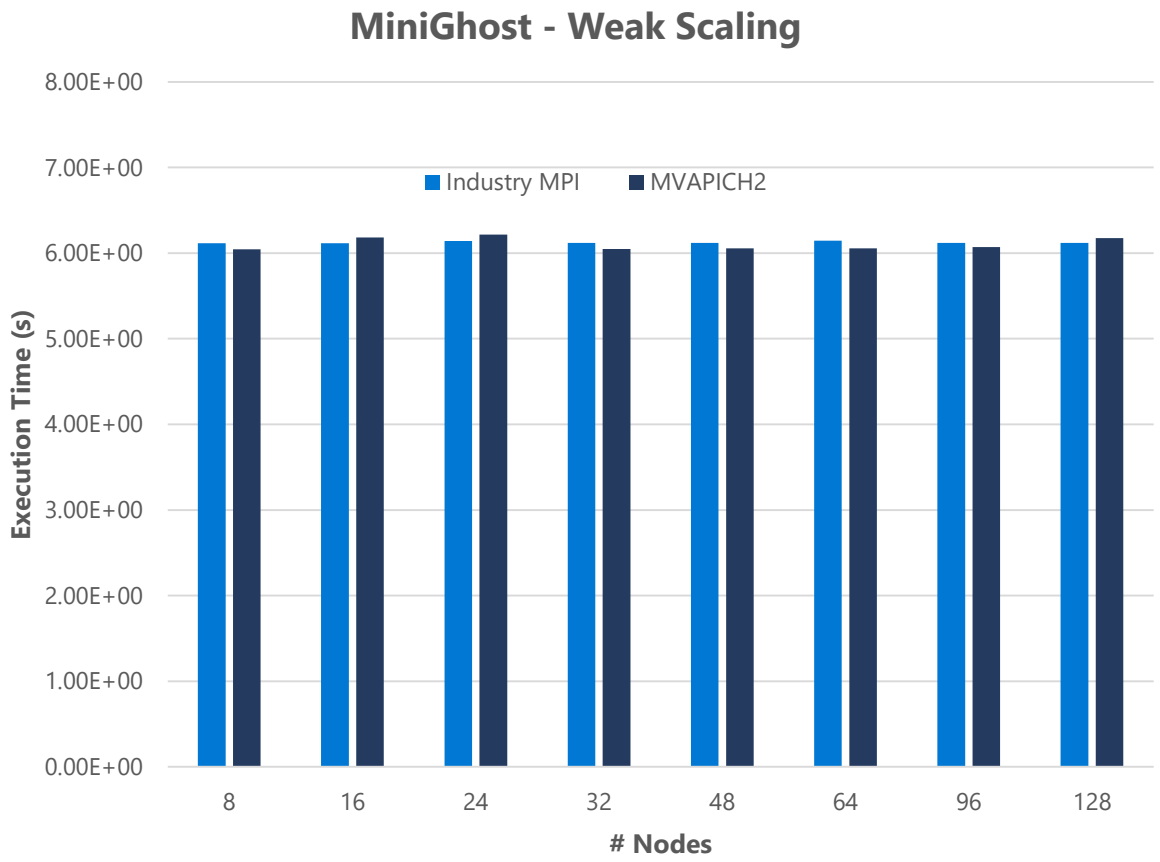
CloverLeaf Scaling Results

- CloverLeaf – opensource CFD code from UK Mini App Consortium
  - Solves compressible Euler equations on Cartesian grid
  - https://uk-mac.github.io/CloverLeaf/
- Input Size: bm256 (larger data
- On par/better performance compared to on-prem: https://intel.ly/37lYi3N



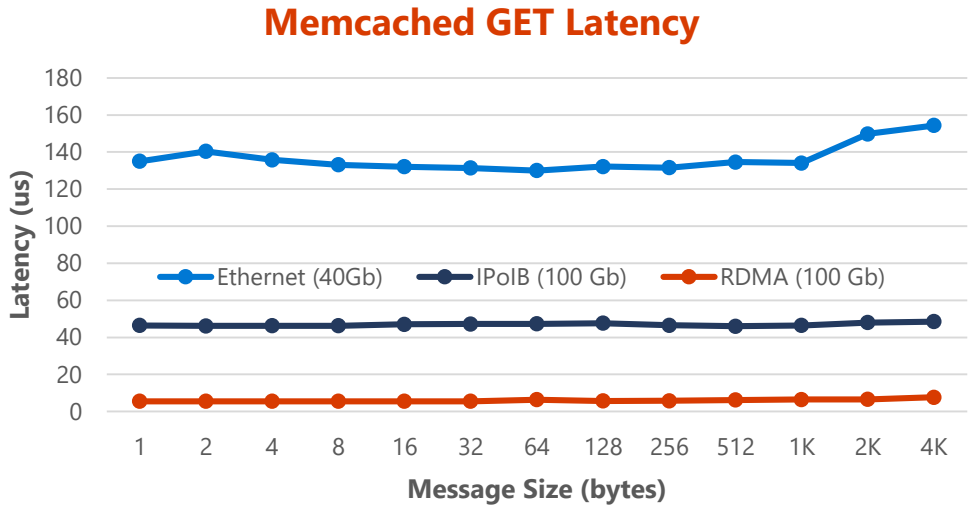
MiniGhost Scaling Results

- MiniGhost v0.9
  - Part of NERSC Trinity Mini Apps
  - 3D near neighbor halo exchange communication
- Weak Scaling: local x, y, z = 100
- 40 ppn per node
- Both MPIs show solid scaling results

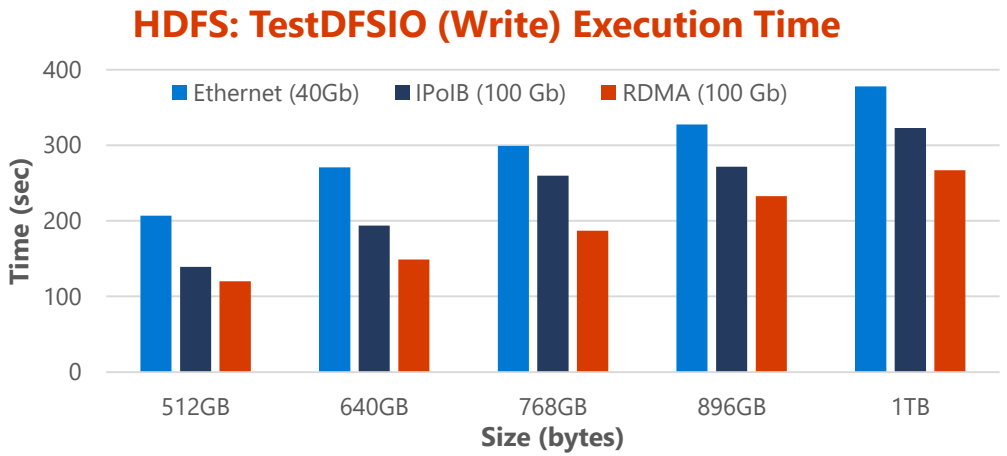


Big Data on Azure HPC

- Memcached RDMA
  - OSU RDMA Memcached 0.9.5
  - Memcached GET Latency: 5.5 us
  - Comparable performance to on-prem: https://bit.ly/2CTAGFO



- HDFS RDMA
  - OSU RDMA Hadoop 3.x 0.9.1
  - Local 700 GB SSD for data node storage
  - Comparable performance to on-prem: https://bit.ly/2NXxkIe



MVAPICH2 Azure Release!

- MVAPICH2-Azure 2.3.2
  - Based on MVAPICH2 2.3.2
  - Enhanced tuning for point-to-point and collective operations
  - Targeted for Azure HB & HC virtual machine instances
  - Flexibility for 'one-click' deployment
    - http://mvapich.cse.ohio-state.edu/downloads/

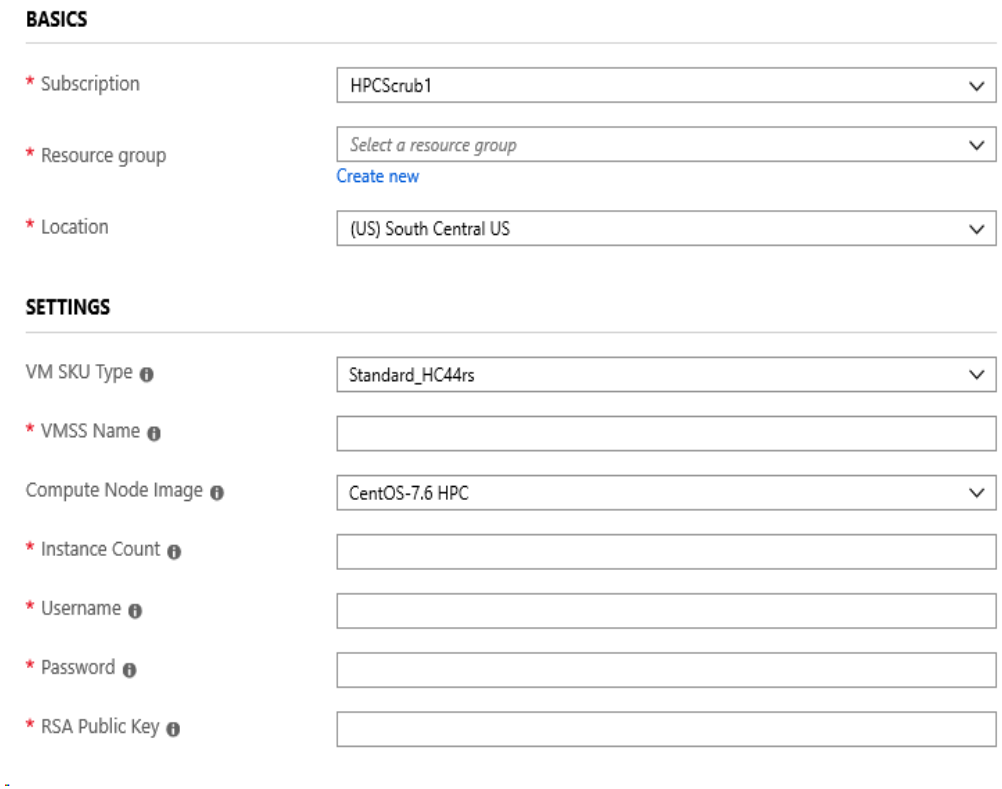
Quick-Deploy MVAICH2 to Azure

- Quick-Deploy Link
  - http://mvapich.cse.ohio-state.edu/downloads/
  - https://github.com/Azure/azhpc-templates/tree/mvapich2/create-vmss

Deploy to Azure

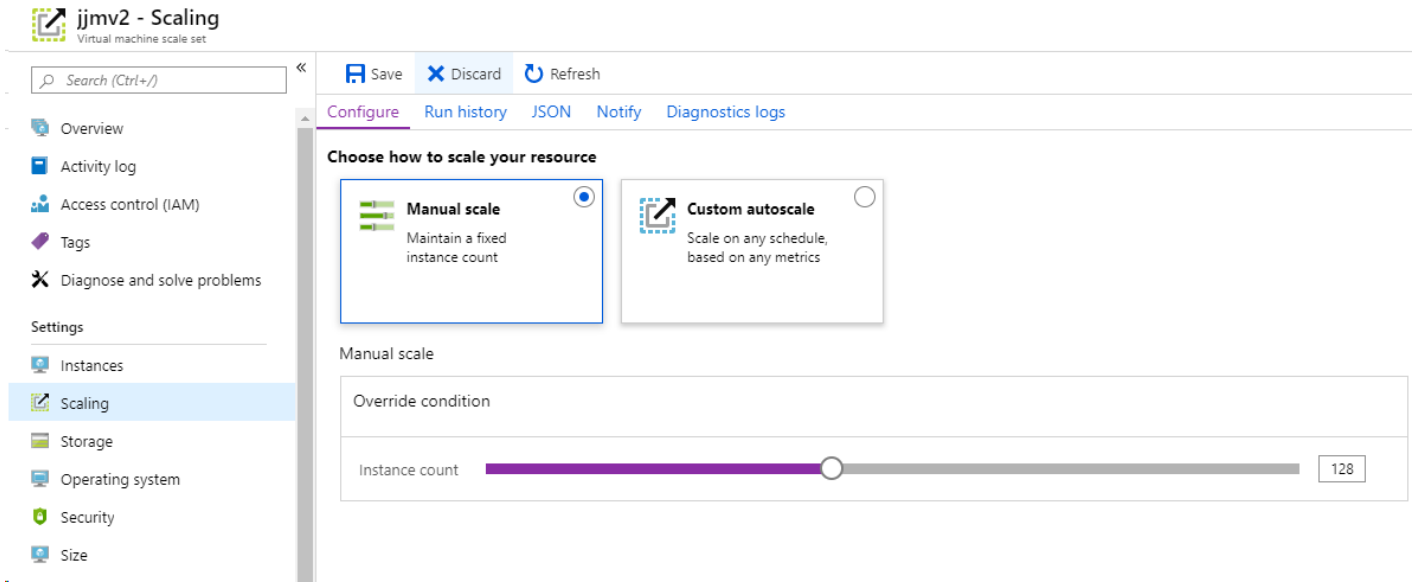
- Cluster Configurations
  - SKU type, size, user credentials

- Deployment
  - Creates a cluster in Azure
  - Head Node, and a set of Compute Nodes (VMSS)
  - NFS backed home folder, ssh ready
  - Ready to use MVAPICH2



Quick-Deploy MVAICH2 to Azure (contd.)

- MVAPICH2 environment module preloaded, and ready to use
- Generate Hostfile
  - scripts/generatehostfile
  - Generates hostfile in home folder
- Can Resize if needed



- More sophisticated deployment options available:
  - CycleCloud, Azure Batch, ARM Templates, Azure Portal, etc.

Links

- Quick Deploy Link
  - http://mvapich.cse.ohio-state.edu/downloads/
  - https://github.com/Azure/azhpc-templates/tree/mvapich2/create-vmss
- MVAPICH2 Azure Performance
  - http://mvapich.cse.ohio-state.edu/performance/mv2-azure-pt_to_pt/
- CentOS 7.6, 7.7 HPC Image
  - https://techcommunity.microsoft.com/t5/Azure-Compute/CentOS-HPC-VM-Image-for-SR-IOV-enabled-Azure-HPC-VMs/ba-p/665557
- HBv2 Preview Access:
  - https://azure.microsoft.com/blog/introducing-the-new-hbv2-azure-virtual-machines-for-high-performance-computing/
  - Initial Performance Results: https://techcommunity.microsoft.com/t5/Azure-Compute/HPC-Performance-and-Scalability-Results-with-Azure-HBv2-VMs/ba-p/1012813

Microsoft

Thank You!