Microsoft

Offering Bare-Metal Performance and Scalability on Cloud: The Azure HPC Approach

MVAPICH User Group Meeting, August 2019

Jithin Jose, Microsoft

Outline

- HPC in the Cloud
- Azure and Azure HPC
- Offering Bare-Metal Experience
- HB/HC clusters
- Performance Highlights
- MVAPICH2-Azure Release
- Demo: Deploy MVAPICH2-powered Azure Cluster
- Conclusion

- ✓ Latest Hardware
- ✓ Always available
- ✓ Instant Scaling
- ✓ Enormous Storage Capacity

- ? Performance Concerns
- ? Software Ecosystem
- ? Data Movement
- ? Security, Cost

How Azure HPC addresses these concerns?

54 regions worldwide **140** available in 140 countries

* Two Azure Government Secret region locations undisclosed

Specialized Compute Fleet in Azure

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

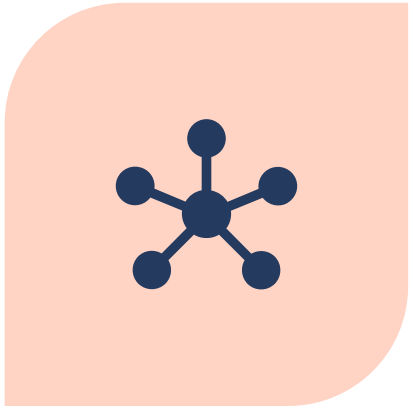
IB Connected CPU/GPU/Storage

Outline

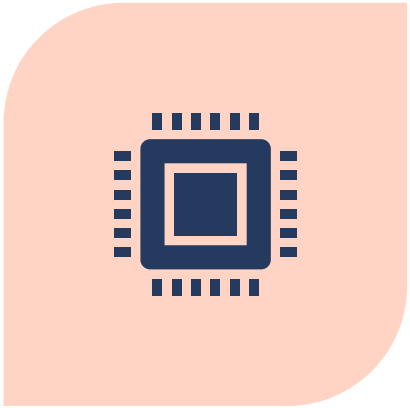
- HPC in the Cloud
- Azure and Azure HPC
- Offering Bare-Metal Experience
- HB/HC clusters
- Performance Highlights
- MVAPICH2-Azure Release
- Demo: Deploy MVAPICH2-powered Azure Cluster
- Conclusion

Offer Bare-metal Experience

Eliminate Jitter



Full Network Experience

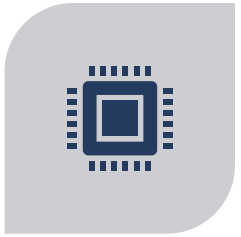


Transparent Exposure of Hardware

Offer Bare-metal Experience (contd.)

Eliminate Jitter

Full Network Experience

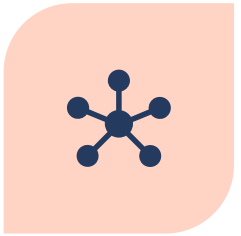


Transparent Exposure of Hardware

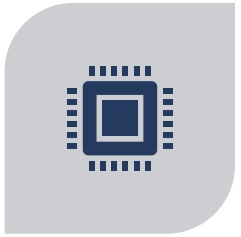
- Eliminate Jitter
  - Partition Host/Guest Resources
  - MinRoot – Shape the Host
  - CPU groups – Shape the Guest
  - No interference either direction!
  - Minimalistic Windows Server Host OS
  - Single VM per Host
  - Minimalistic Host/Guest Agent

Offer Bare-metal Experience (contd.)

Eliminate Jitter



Full Network Experience



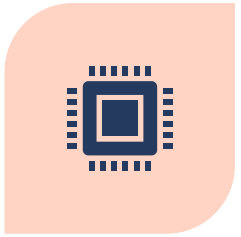
Transparent Exposure of Hardware

- Full Network Experience
  - Single Root IO Virtualization (SR-IOV)
  - Enable customers to use familiar software ecosystem
  - Expose NIC features: hcoll, MCAST, DCT*, SHARP*, etc.
  - Adaptive Routing

Offer Bare-metal Experience (contd.)

Eliminate Jitter

Full Network Experience



Transparent Exposure of Hardware

- Transparent Exposure of Hardware
  - Consistent and deterministic mapping of pNUMA - vNUMA topology*
  - Core mappings do not change over runtime

Outline

- HPC in the Cloud
- Azure and Azure HPC
- Offering Bare-Metal Experience
- HB/HC clusters
- Performance Highlights
- MVAPICH2-Azure Release
- Demo: Deploy MVAPICH2-powered Azure Cluster
- Conclusion

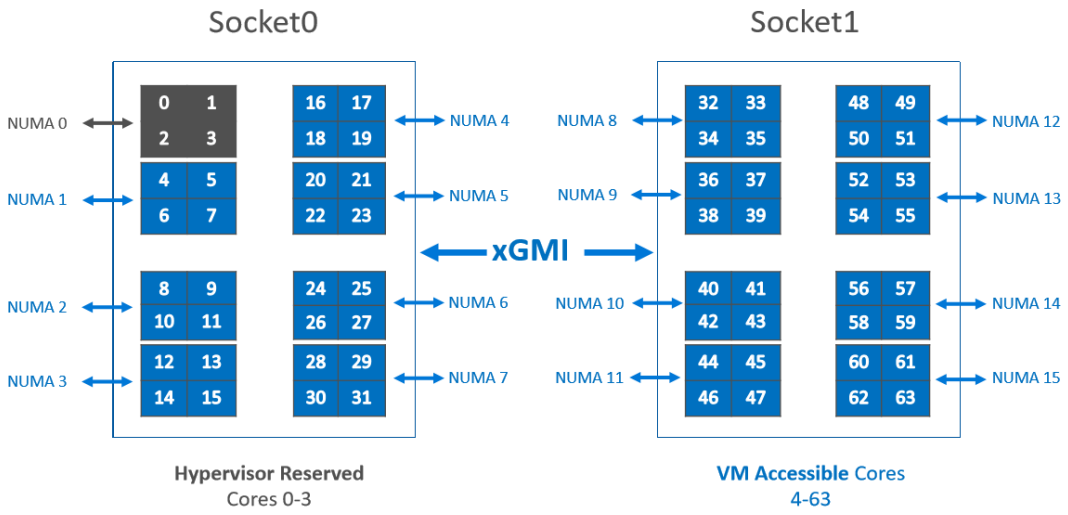
Azure HB/HC Series VMs

	HB	HC
CPU	AMD EPYC	Intel Xeon Platinum
Cores/VM	60	44
Clock Speed*	2.55 GHz	3.4 GHz
Memory Bandwidth	263 GB/sec	191 GB/sec
Memory	4 GB/core 240 total	8 GB/core 352 GB
Local Disk	700 GB NVMe	
InfiniBand	100 Gb EDR w/ ConnectX-5 (SRIOV)	
Network**	Gen 2 SmartNIC (Arria10 FPGA accelerated)	

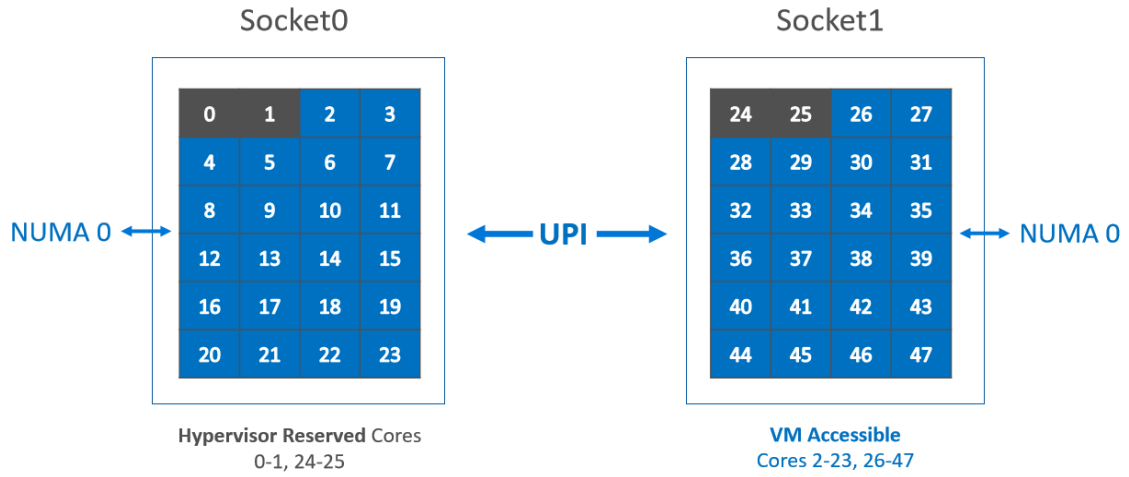
Azure HB/HC Series VMs

	HB	HC
Targets	CFD, Seismic, Weather,	FEA, MD, Chemistry
Workload Driver	Memory Bandwidth	Raw Compute
Max MPI job per VMSS (300 VMs)	18,000 cores	13,200 cores
RDMA Verbs	All	All
MPI Support	MVAPICH2, OpenMPI, MPICH, Platform MPI, Intel MPI, HPC-X	
Additional Frameworks	Unified Communication X libfabric, PGAS	
Azure Storage Support	Premium	Premium
OS Support for RDMA	CentOS/RHEL 7.5+ SLES 12 SP4+ WinServer 2016+	CentOS/RHEL 7.5+ SLES 12 SP4+ WinServer 2016+

Core Partitioning in HB and HC VMs



HB60rs Virtual Machines

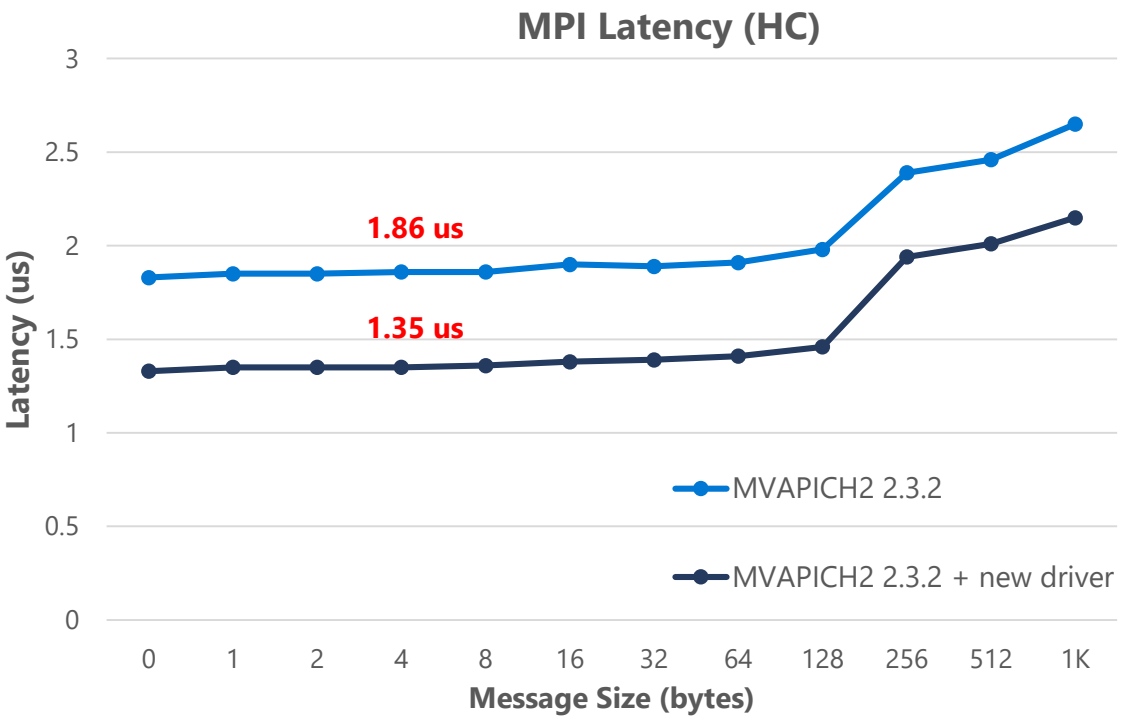


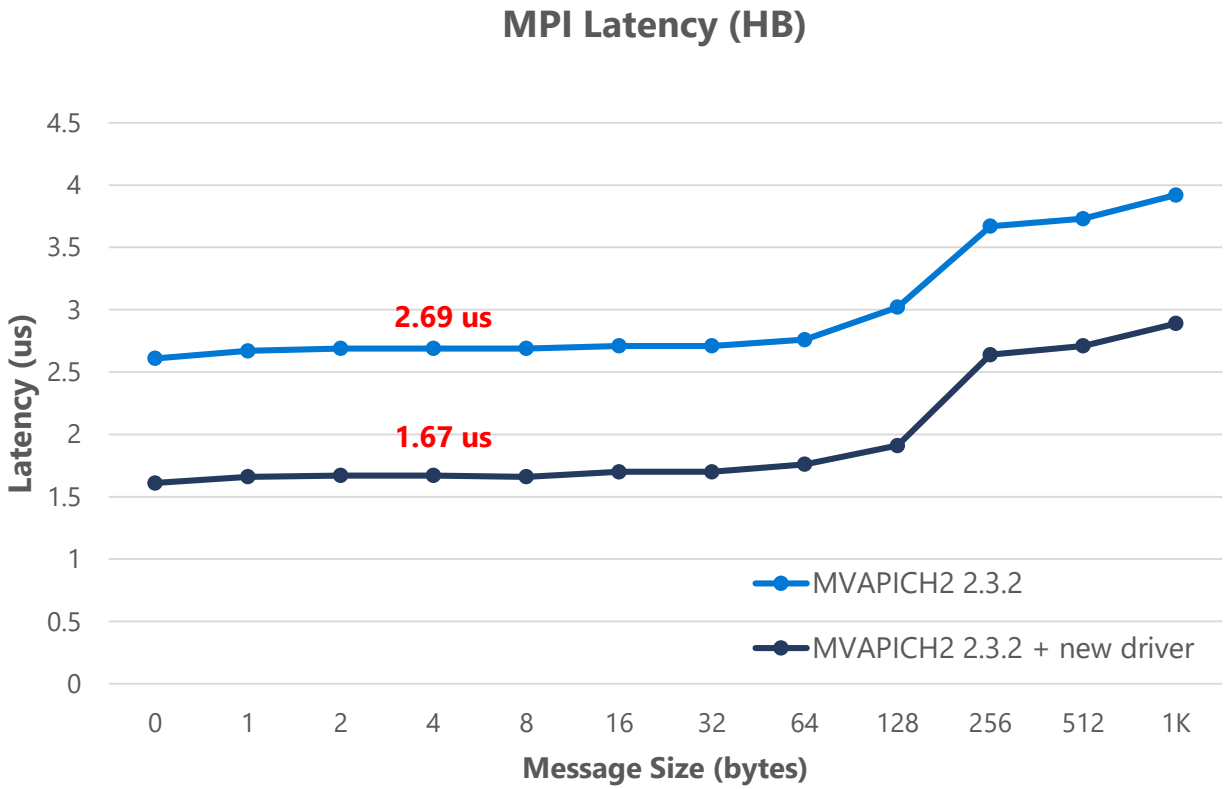
HC44rs Virtual Machines

Outline

- HPC in the Cloud
- Azure and Azure HPC
- Offering Bare-Metal Experience
- HB/HC clusters
- Performance Highlights
- MVAPICH2-Azure Release
- Demo: Deploy MVAPICH2-powered Azure Cluster
- Conclusion

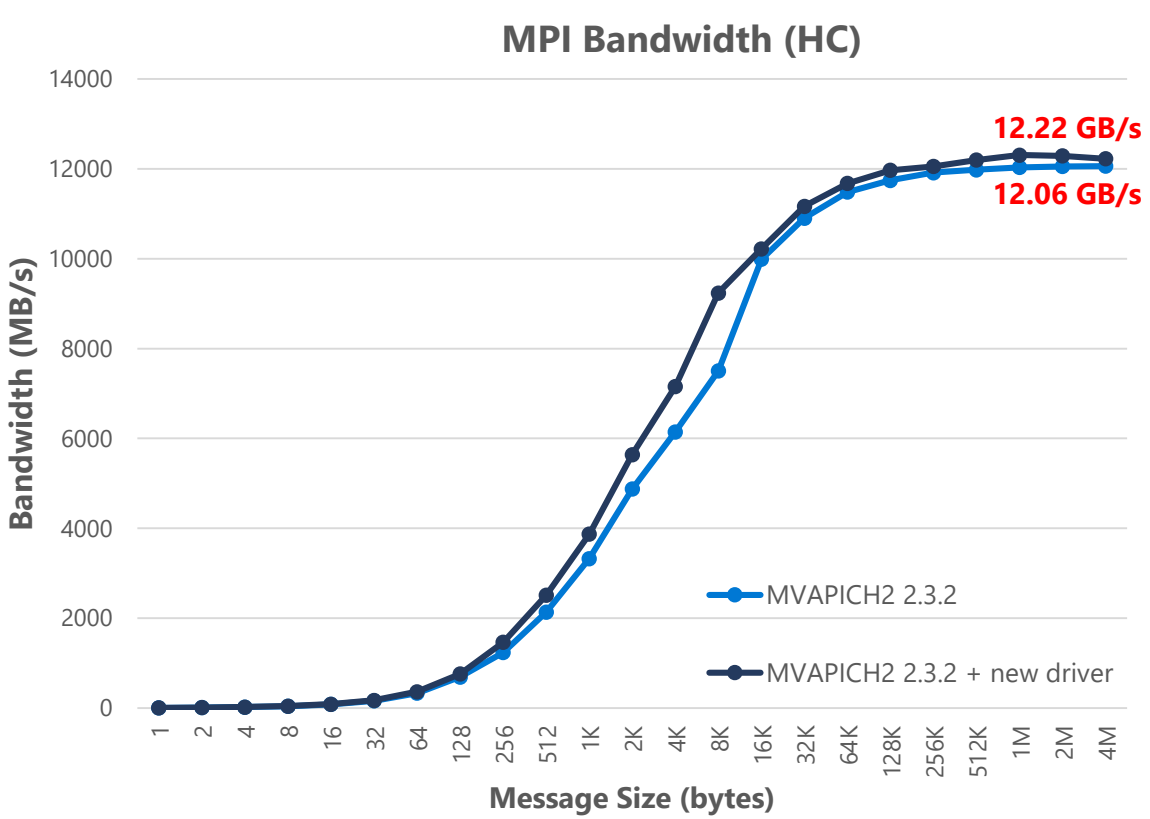
MVAPICH2 Latency on HB/HC

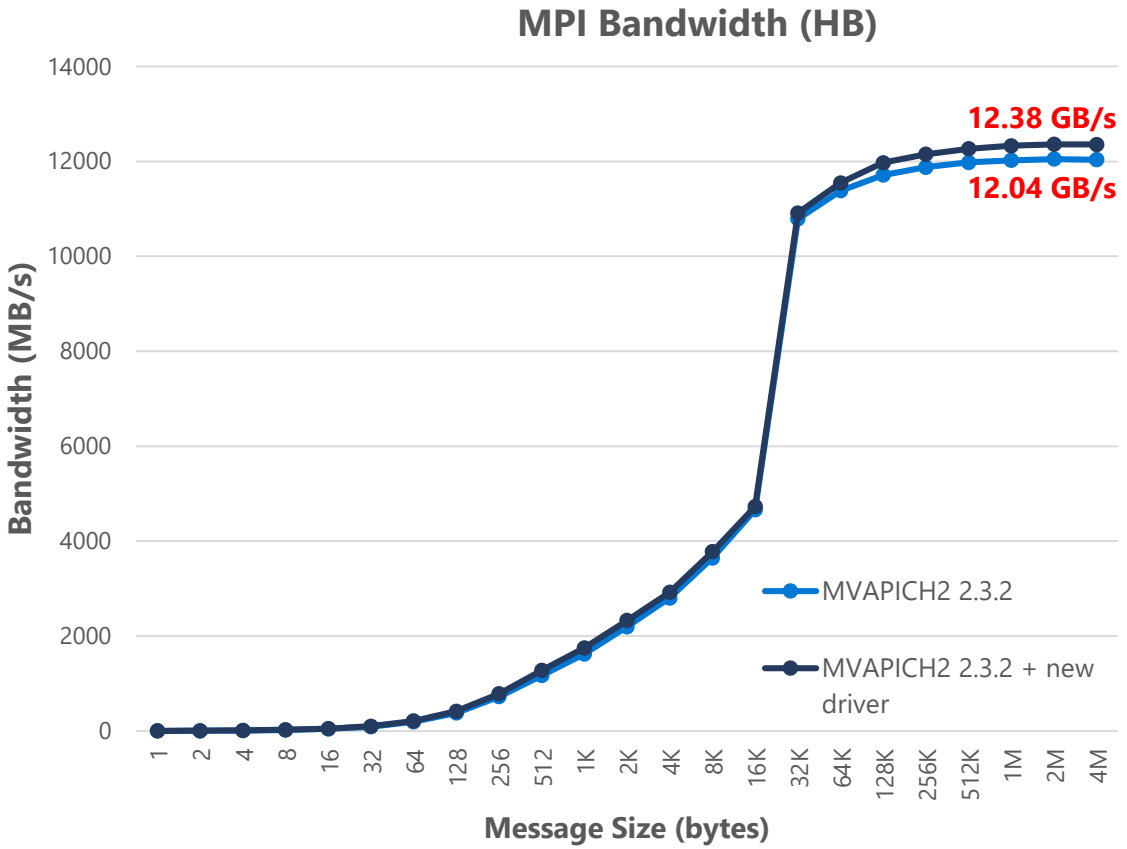




- Processes pinned closer to NIC
- 4 byte MPI level latency: **1.35 us** on HC, **1.67 us** on HB
- New driver being rolled out – improvements to Host OS, firmware, and OFED

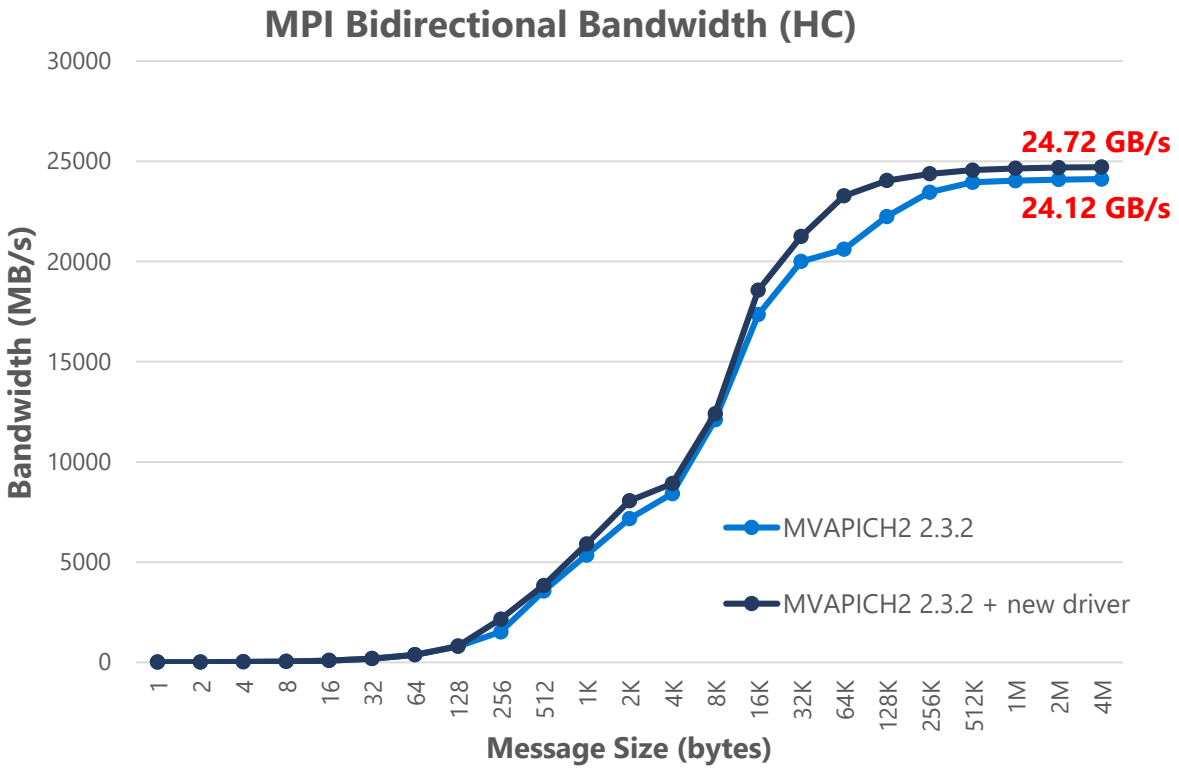
MVAPICH2 Bandwidth on HB/HC

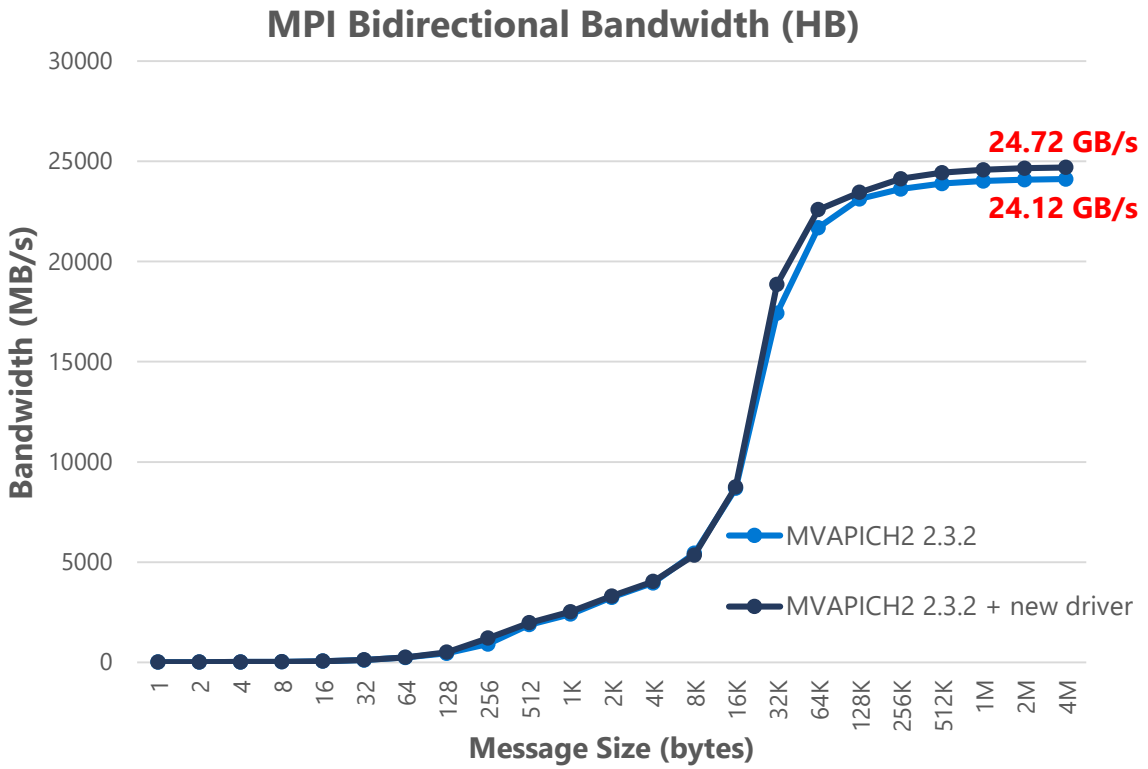




- Processes pinned closer to NIC
- 4 byte MPI level latency: **12.22 GB/s** on HC, **12.38 GB/s** on HB
- New driver being rolled out – improvements to Host OS, firmware, and OFED

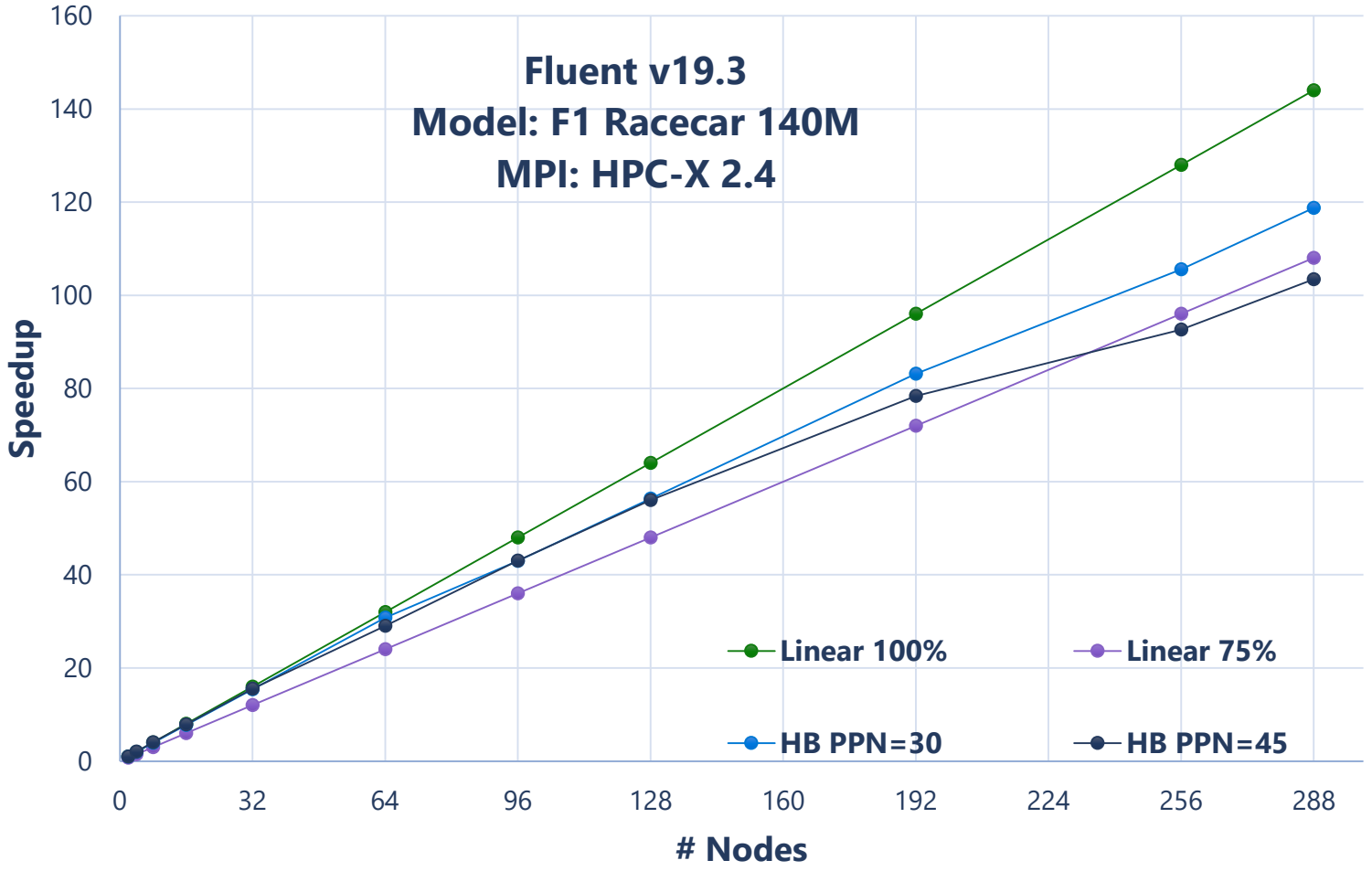
MVAPICH2 Bidirectional Bandwidth on HB/HC





- Processes pinned closer to NIC
- 4 byte MPI level latency: **24.70 GB/s** on HC, **24.72 GB/s** on HB
- New driver being rolled out – improvements to Host OS, firmware, and OFED

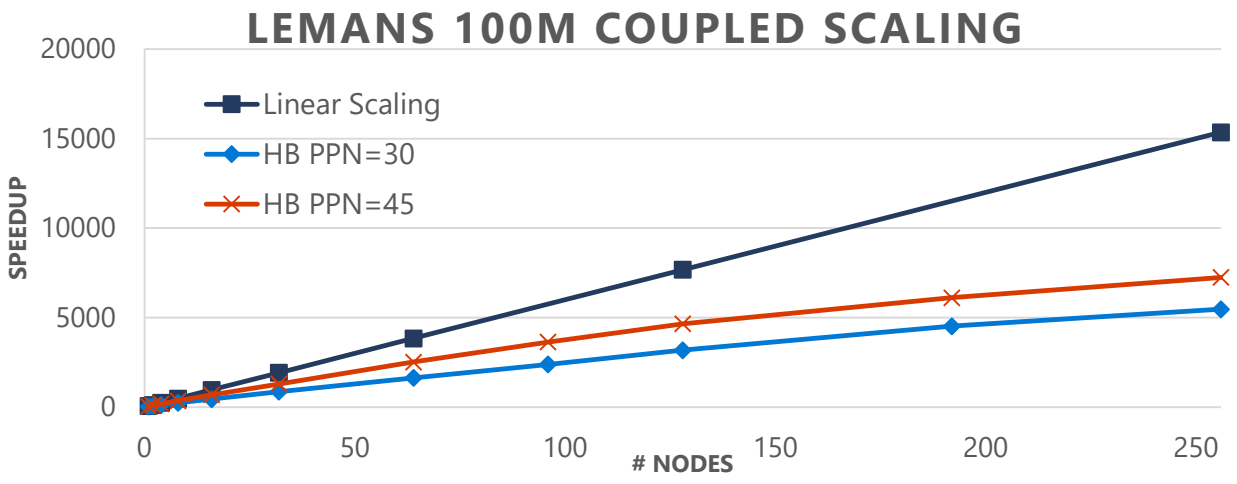
Application Scaling: Ansys Fluent

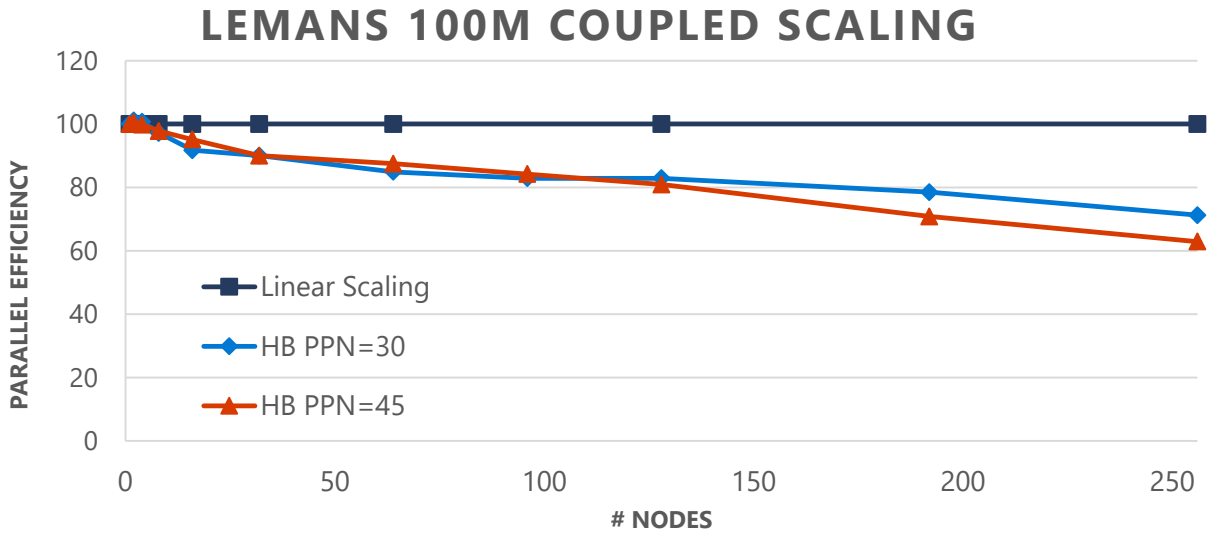


- 88% efficiency at 128 VMs, within 1% of XC50
- 83% Efficiency at 288 VMs (12,960 MPI ranks)
- Very useful learnings about scalability of EPYC

Bare Metal Results: https://www.ansys.com/solutions/solutions-by-role/it-professionals/platform-support/benchmarks-overview/ansys-fluent-benchmarks/ansys-fluent-benchmarks-release-19/external-flow-over-a-formula-1-race-car

Application Scaling: StarCCM



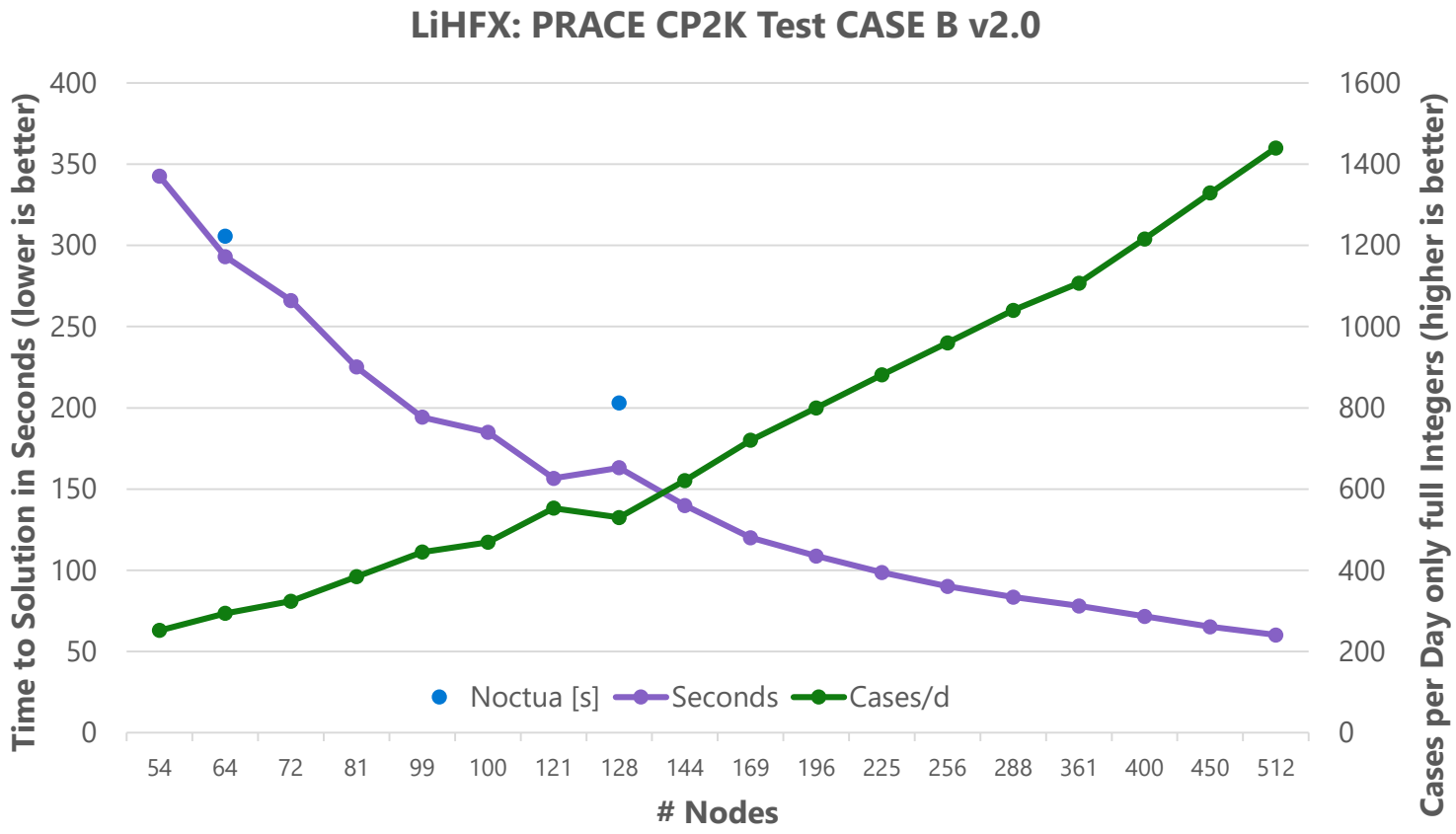


- StarCCM 14.02
- Scaling Results with 256 VMs (11,520 MPI)
- 71.3% Efficiency at 256 nodes

Host	Cores	PPN	Speedup	Efficiency
1	30	30	30	100.0
2	60	30	59.9	99.9
4	120	30	119.2	99.3
16	480	30	446.6	93.0
32	960	30	895.7	93.3
64	1920	30	1664.8	86.7
128	3840	30	3138.1	81.7
256	7680	30	5473.3	71.3

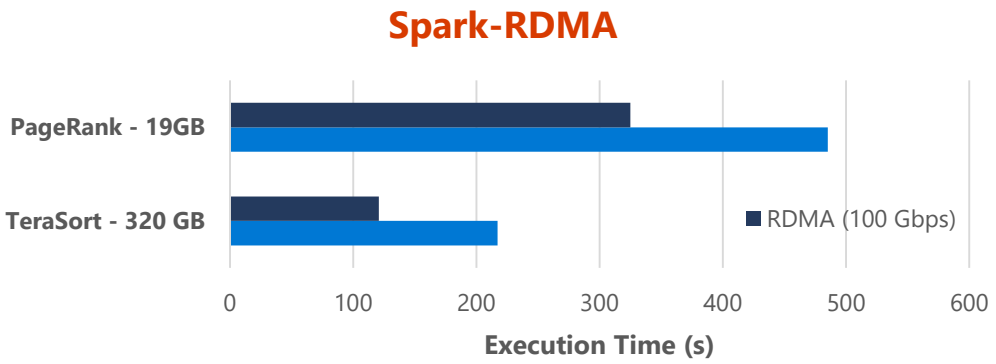
Host	Cores	PPN	Speedup	Efficiency
1	45.0	45.0	45	100.0
2	90.0	45.0	89	99.1
16	720.0	45.0	697	96.7
32	1440.0	45.0	1318	91.5
64	2880	45	2505	87.0
128	5760	45	4451	77.3
256	11520	45	7449	64.7

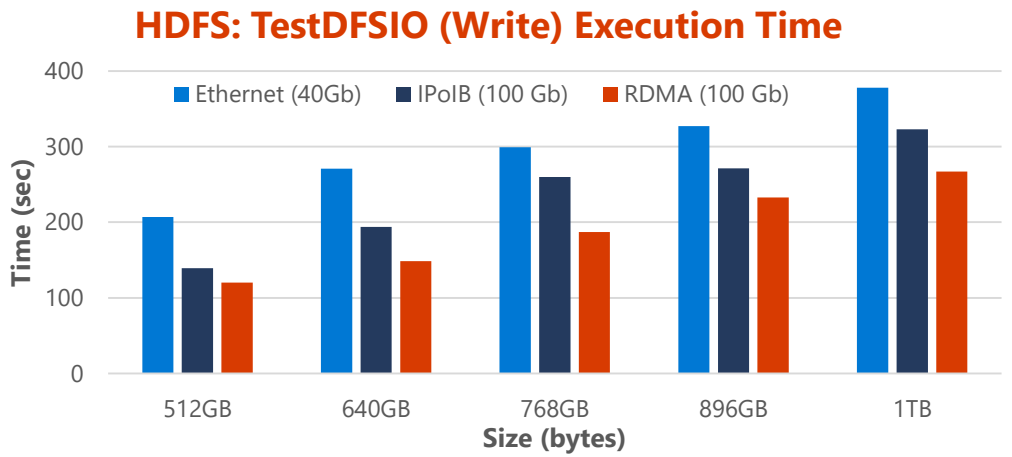
Application Scaling : CP2K at 22,528 cores

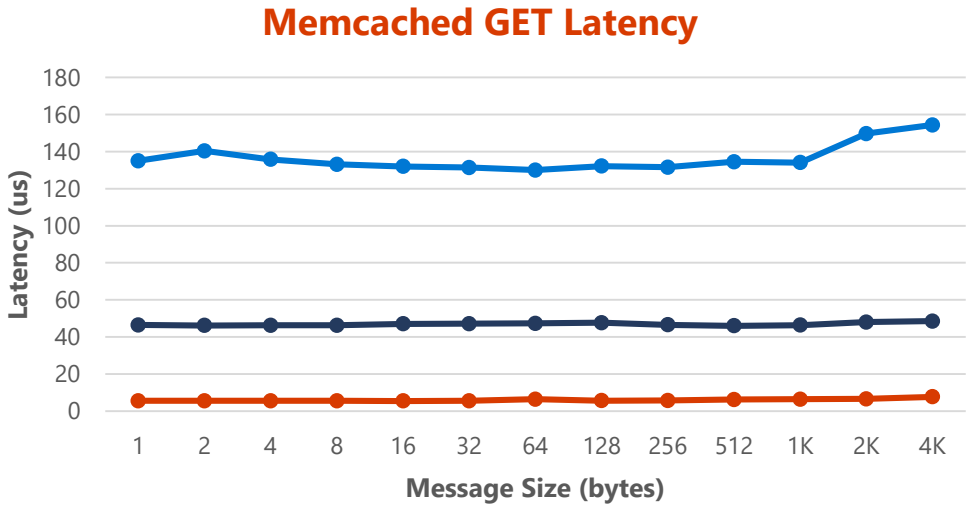


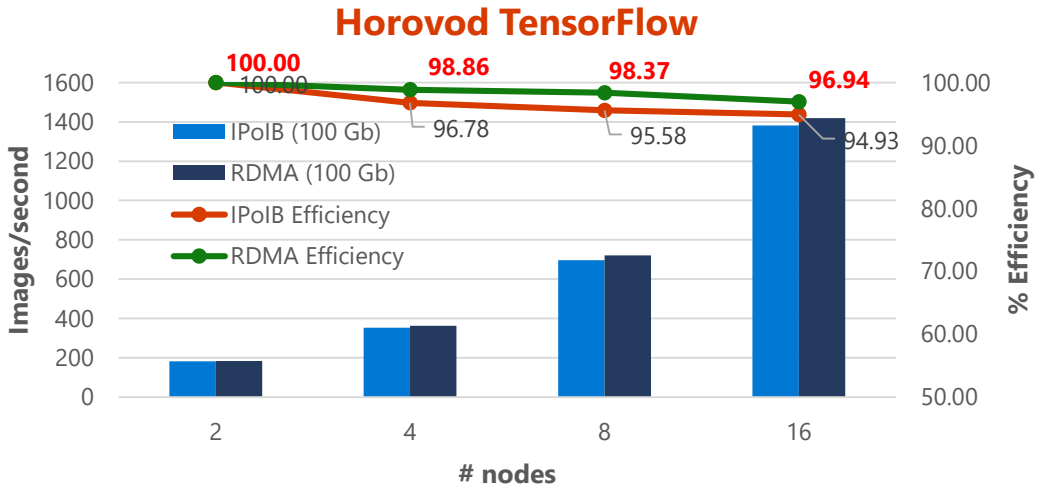
- Molecular Dynamics Application
- Part of PRACE Consortium Benchmarks
- 512 HC VMs at 22,528 cores
- 26X improvement for 21X increase in VM resources
  - 24 VMs – 55 cases/day
  - 512 VMs – 1440 cases/day
- https://www.cp2k.org/performance

Big Data on Azure HPC









Outline

- HPC in the Cloud
- Azure and Azure HPC
- Offering Bare-Metal Experience
- HB/HC clusters
- Performance Highlights
- MVAPICH2-Azure Release
- Demo: Deploy MVAPICH2-powered Azure Cluster
- Conclusion

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/

Outline

- HPC in the Cloud
- Azure and Azure HPC
- Offering Bare-Metal Experience
- HB/HC clusters
- Performance Highlights
- MVAPICH2-Azure Release
- Demo: Deploy MVAPICH2-powered Azure Cluster
- Conclusion

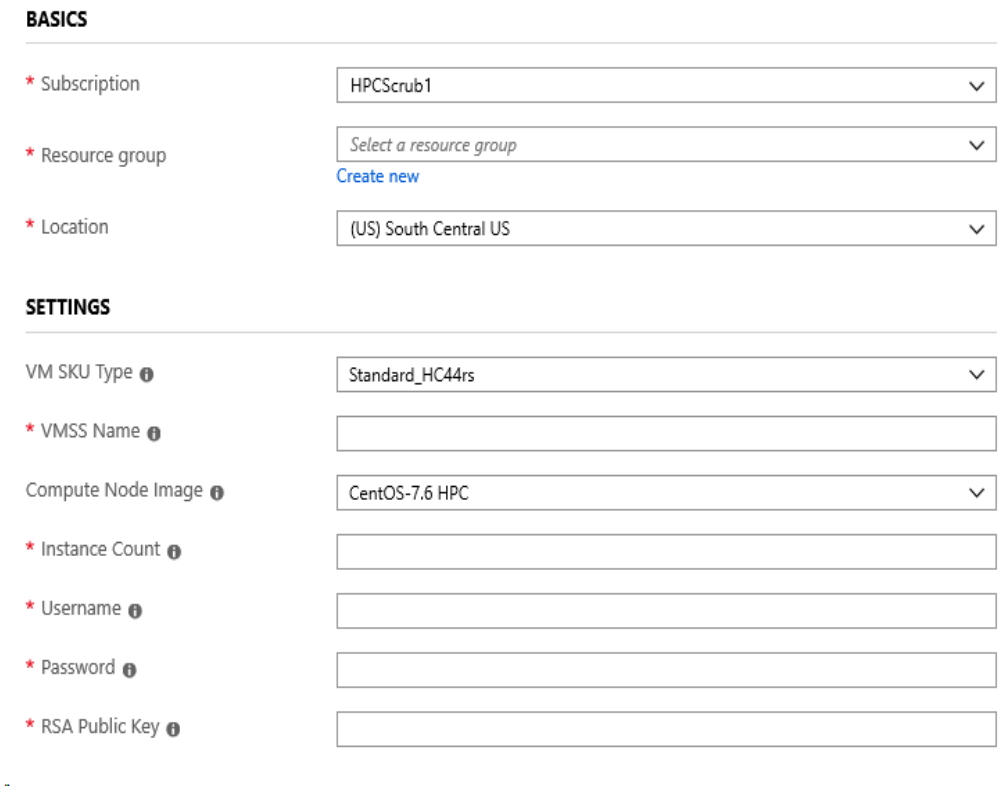
Quick-Deploy MVAICH2 to Azure

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

Deploy to Azure

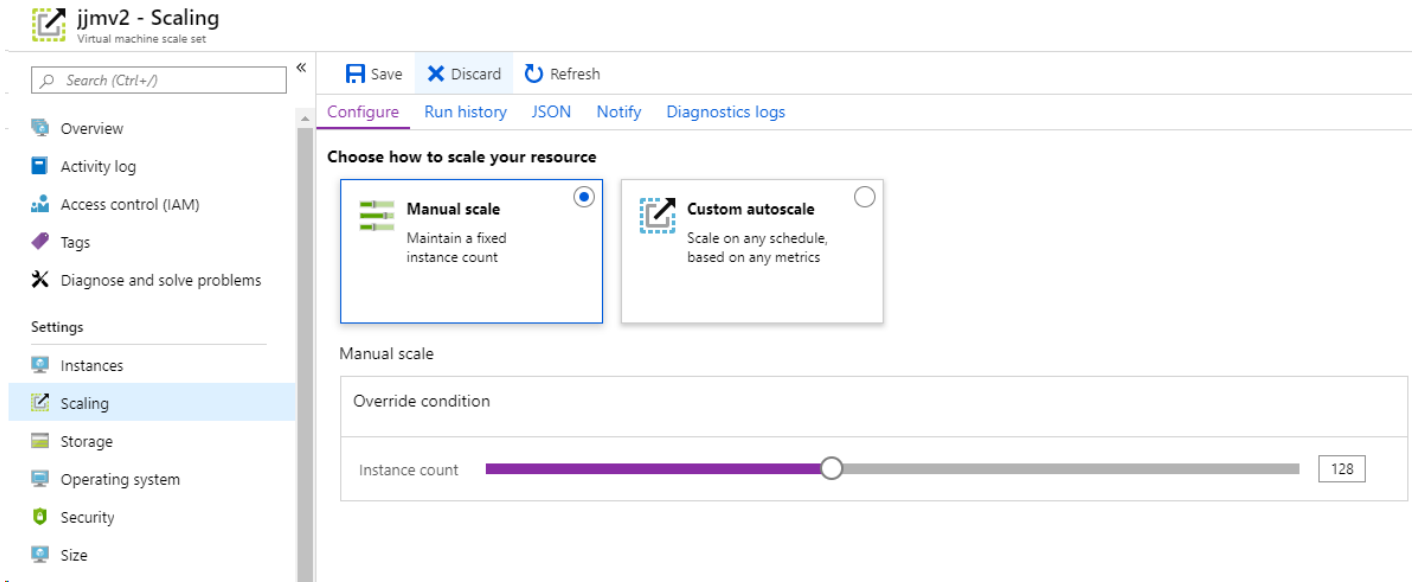
- Cluster Configurations
  - SKU type, size, user credentials

- 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



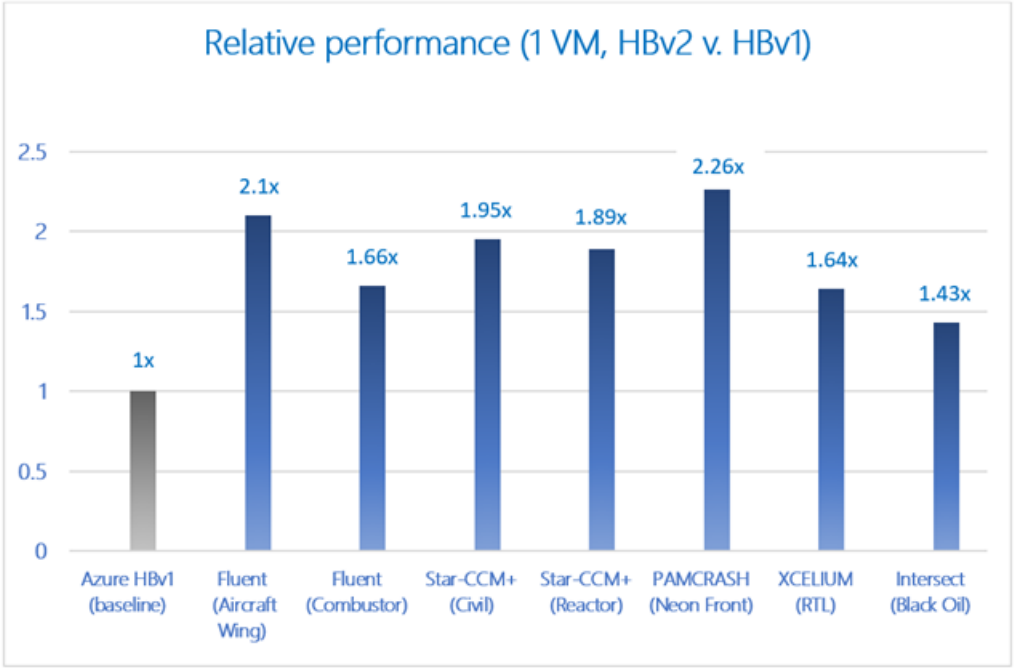
- More sophisticated deployment options available:
  - CycleCloud, Azure Batch, ARM Templates, Azure Portal, etc.

Outline

- HPC in the Cloud
- Azure and Azure HPC
- Offering Bare-Metal Experience
- HB/HC clusters
- Performance Highlights
- MVAPICH2-Azure Release
- Demo: Deploy MVAPICH2-powered Azure Cluster
- Conclusion

Upcoming HBv2 Instances

- AMD based HBv2
  - 120 AMD EPYC™ 7002-series CPU cores (no SMT)
  - 480 GB of RAM, 480 MB of L3 cache
  - Up to 350 GB/sec of memory bandwidth
  - Connect-X 6 200 Gb/s HDR InfiniBand
  - Sign up for Preview Access!



Size	CPU cores	Memory: GB	Memory per CPU Core: GB	Local SSD: GiB	RDMA network	Azure network
Standard_HB120rs	120	480 GB	4 GB	1.6 TB	200 Gbps	40 Gbps

Preview Access: https://azure.microsoft.com/en-us/blog/introducing-the-new-hbv2-azure-virtual-machines-for-high-performance-computing/

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 HPC Image
  - https://techcommunity.microsoft.com/t5/Azure-Compute/CentOS-HPC-VM-Image-for-SR-IOV-enabled-Azure-HPC-VMs/ba-p/665557
- Virtual Machine Scalesets
  - https://docs.microsoft.com/azure/virtual-machine-scale-sets/overview
- HBv2 Announcement
  - https://azure.microsoft.com/blog/introducing-the-new-hbv2-azure-virtual-machines-for-high-performance-computing/

Microsoft

Thank You!