MVAPICH2 at Azure: Enabling High Performance on Cloud

10th Annual MVAPICH User Group (MUG) Meeting, August 2022

Jithin Jose, Microsoft
jjjos@microsoft.com
Agenda

• Overview of Azure HPC SKUs
  • Azure HBv3, NDv4
• Feature Highlights
• MVAPICH2 on HBv3
• MVAPICH2 GDR on NDv4
• Azure HPC VM Image
• Performance Highlights
• Conclusion
Azure HPC/AI VM Series

- **Standard HPC VMs**
  - Standard HPC Applications
  - High Compute/Memory + InfiniBand
  - HPC SKUs: HB, HC, HBv2, HBv3

- **GPU VMs**
  - Deep Learning, AI workloads
  - Visualization SKUs: NV series
  - Deep Learning/AI SKUs: NC, ND series

- “r” in VM type indicates RDMA support (InfiniBand)
- InfiniBand/RDMA enabled VMs: One VM per Host
- InfiniBand exposed to VMs using SR-IOV, offers full host bypass with full feature support
- Partition Key (P-key) based isolation
Azure HBv3

- VM Specs:
  - AMD Milan-X (NPS = 2)
  - VM Cores: 120
  - L3 Cache: 1.5 GB per VM
  - 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)

Ideal for traditional HPC/MPI workloads
Azure NDv4

• VM Specs:
  • AMD Rome (NPS=2)
  • VM Cores: 96 (48 per socket)
  • Memory: 900 GB
  • 8 x NVIDIA A100 GPUs (NVLink 3.0)
  • 8 x HDR 200Gbps InfiniBand
  • Local Disk: 6.4 TB local NVMe SSD

Ideal for AI/Deep learning workloads
Agenda

- Overview of Azure HPC SKUs
  - Azure HBv3, NDv4
- Feature Highlights
- MVAPICH2 on HBv3
- MVAPICH2 GDR on NDv4
- Azure HPC VM Image
- Performance Highlights
- Conclusion
InfiniBand Features in Azure

- **HB, HC, NDv2:**
  - EDR 100 Gb/s InfiniBand
  - Up to 200 M messages/second

- **HBv2, HBv3, NDv4:**
  - HDR 200 Gb/s InfiniBand
  - Up to 215 M messages/second

- **Dynamically Connected Transport (DCT)**
  - Reliable and scalable transport
  - Lesser Memory footprint

- **Hardware offload**
  - Collectives offload framework
  - Hardware tag matching

- **UD multicast (MCAST)**
  - Unreliable datagram (UD) based multicast

- **SHARP**
  - Switch based collectives

- **Dynamic Routing**
  - Advanced Congestion Control
  - Adaptive Routing

- **Better Reliability**
  - SHIELD detects link failures and reroutes
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

![NCCL AllReduce (64 ranks)](image)

NCCL AllReduce (64 ranks)

- SHARP-Disabled
- SHARP-Enabled

227 GB/s
Congestion Control

- Available on all VM Series with HDR200
- Transparent to customer applications
- Improve tail latencies
- Critical in public multi-customer environments
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
Agenda

• Overview of Azure HPC SKUs
  • Azure HBv3, NDv4
• Feature Highlights
• MVAPICH2 on HBv3
• MVAPICH2 GDR on NDv4
• Azure HPC VM Image
• Performance Highlights
• Conclusion
MVAPICH2 on HBv3 (inter-node)

Software Configuration:
- VM Image: Azure CentOS-HPC 8.1 VM Image
- MPI Libraries: MVAPICH2 2.3.7-1, MVAPICH2 3.0a + UCX (RC)
- UCX: 1.10.0
MVAPICH2 on HBv3 (intra-node)

Software Configuration:
- VM Image: Azure CentOS-HPC 8.1 VM image
- MPI Libraries: MVAPICH2 2.3.7-1, MVAPICH2 3.0a + UCX (sm)
- UCX: 1.10.0
Agenda

• Overview of Azure HPC SKUs
  • Azure HBv3, NDv4

• Feature Highlights

• MVAPICH2 on HBv3
  • MVAPICH2 GDR on NDv4

• Azure HPC VM Image

• Performance Highlights

• Conclusion
Software Configuration: MVAPICH2 2.3.7–GDR on Azure Ubuntu-HPC 18.04 VM Image

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
Agenda

- Overview of Azure HPC SKUs
  - Azure HBv3, NDv4
- Feature Highlights
- MVAPICH2 on HBv3
- MVAPICH2 GDR on NDv4
- Azure HPC VM Images
- Performance Highlights
- Conclusion
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 libraries
  - Blis, FFTW, Flame, MKL
- Recommended Compilers
- GPU Drivers
- NCCL, NCCL RDMA Sharp Plugin, SharpD
- Other optimizations
Agenda

- Overview of Azure HPC SKUs
  - Azure HBv3, NDv4
- Feature Highlights
- MVAPICH2 on HBv3
- MVAPICH2 GDR on NDv4
- Azure HPC VM Images
- Performance Highlights
- Conclusion
NCCL at Scale on NDv4

NCCL AllReduce (w/o SHARP)
- 565 nodes (4,520 ranks)
- 1,571 nodes (12,568 ranks)

NCCL AllReduce w/ SHARP
- 128 nodes (1,024 ranks)
- 512 nodes (4,096 ranks)
- 998 nodes (7,984 ranks)
NCCL AllReduce Bandwidth Distribution

- Azure InfiniBand Clusters deploy Non-blocking (under-subscribed) fat-tree topology
- Evaluation using all-pair NCCL AllReduce benchmark
- Cluster size = ~470 NDv4 (8 x A100, 8 x 200 Gbps HDR) nodes
- Multiple pairs (N/2) communicating at the same time
- 100% pairs achieve > 186 GB/s
Tutel: Adaptive MoE at Scale

- New AlltoAll algorithm optimized for NDv4/NDmv4 cluster
  - Larger slice through IB => 8x slice size in large scale
  - Only 1-to-1 IB interconnection required in inter-node aggregation phase
  - Open-source on github.com/microsoft/msccl
  - Achieve >6.7x gain on 256MiB and >1.9x gain on 512MiB with 256 NDmv4 nodes

- New AlltoAll algorithm + Other framework optimizations: > 40% E2E performance improvement
Scaling Efficiency on HBv3 (Milan-X)

Ansys Fluent 2021 R1
f1_racecar_140m

156% scaling efficiency

Ansys Fluent 2021 R1
f1_combustor_830m

106% scaling efficiency

OpenFOAM v. 1912
Motorbike 28m

119% scaling efficiency

https://aka.ms/MilanXPerf
Agenda

- Overview of Azure HPC SKUs
  - Azure HBv3, NDv4
- Feature Highlights
- MVAPICH2 on HBv3
- MVAPICH2 GDR on NDv4
- Azure HPC VM Images
- Performance Highlights
- Conclusion
Conclusion

• Supercomputer on Cloud is real!
• Azure HPC Cloud made into Top500, Graph500
• High Performance middleware such as MVAPICH2 enables cutting edge technology
  • Deliver High Scalability and Performance
<table>
<thead>
<tr>
<th>Pointers</th>
<th>Getting Started</th>
<th>HPC VM Images</th>
<th>HPC VM Series</th>
<th>HPC VM Deployment</th>
<th>GPU VM Series</th>
<th>Azure HPC Blogs</th>
</tr>
</thead>
</table>
|          | * High Performance Computing (HPC) on Azure | * Azure HPC VM Images  
* GitHub Repository | * Sample HPC VM deployment scripts  
* Azure CycleCloud  
* MUG ‘20 Tutorial | * Azure VM sizes - HPC - Azure Virtual Machines | * Azure VM sizes - GPU - Azure Virtual Machines | * Azure Compute - Microsoft Tech Community |