MVAPICH2 at Azure: Enabling High Performance on Cloud

Jithin Jose, Microsoft

11th Annual MVAPICH User Group Conference, Aug 2023
• Introduce latest HPC and AI VMs at Azure
  • Azure HBv4
  • Azure NDv5
• MVAPICH2 for HPC workloads (HBv4)
• MVPIAHC2 for AI workloads (NDv5)
• Performance and Scaling Highlights
Azure HPC/AI VM series – H and N series

HPC VMs

- Standard HPC Applications
- High Compute, Memory + InfiniBand
- HPC SKUs: HB, HC, HBv2, HBv3, HBv4

AI, Visualization VMs

- Deep Learning, AI workloads, Visualization

Visualization SKUs:
- NV series

Deep Learning/AI SKUs:
- NC, ND series

- 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
VM Specs:
AMD Genoa-X (NPS = 2)
VM Cores: 176
Memory: 704 GB (HBv4), 1408 GB (HXv1)
Local Disk: 2 x 1.8 TB NVMe SSD
Network: 400 Gbps NDR (SR-IOV)
VM Specs:
- Intel Saphire Rapids
- VM Cores: 96
- Memory: 1900 GB
- 8 x Nvidia H100 GPUs (NVLink 4.0)
- 8 x NDR 400 Gbps InfiniBand

NDv5 VM Series – Detailed Spec
HPC/AI Interconnect on 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

**HBv4, HXv1, NDv5:**
- NDR 400 Gb/s InfiniBand
- Up to 330 M messages/second

**ConnectX-5**
- Dynamically Connected Transport (DCT)
- Reliable and scalable transport
- Lesser Memory footprint

**ConnectX-6**
- Hardware offload
  - Collectives offload framework
  - Hardware tag matching
- UD multicast (MCAST)
  - Unreliable datagram (UD) based multicast
  - Create a mcast group and broadcast

**ConnectX-7**
- SHARP
  - Switch based collectives
- Dynamic Routing
  - Advanced Congestion Control
  - Adaptive Routing
- Better Reliability
  - SHIELD detects link failures and reroutes
Azure HPC/AI 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
• Similar latency results for HBv3 vs. HBv4 (~2us)
• 2X improvement for Bandwidth and Bi-Bandwidth with HBv4
• Line rate for both HBv3 and HBv4
Latency Improvement on large messages with NDv5

2X improvement for Bandwidth and Bi-Bandwidth

Line rate for both NDv4 and NDv5 with GPUDirect RDMA

Commercial MPI library used on NDv5 (build constraints)
HBv4 Fabric (NDR) Performance Distribution

- ~N/2 pairs run in parallel
- Latency
  - 1.6-1.7 us (Hops = 2)
  - 2.2-2.3 us (Hops = 4)
- Bandwidth
  - 375-385 Gbps
- Bi-Bandwidth
  - 730-770 Gbps

Uniform distribution of bandwidth/bi-bandwidth with all-pairs communication model
Two bands in latency distribution corresponds to hops = 2, 4
NCCL Performance on NDv5

NCCL AllReduce: Impact of NVLink SHARP (256 GPUs)

NCCL AllReduce Performance Comparison
NCCL Scalability Analysis on NDv5

- NCCL AllReduce Scaling
  - 256 to 3,168 GPUs
  - Compare NCCL, NCCL-Tuned, NCCL+SHARP
~2X Improvements with HBv4 compared to HBv3
HBv4: Fluent, NAMD Scaling

**F1 Racecar 140M**
- Fluent – 64 VMs (11,264 ranks)

**STMV (224M atoms)**
- NAMD – 128 VMs (22,528 ranks)
Tutel: Adaptive MoE at Scale

- New AlltoAll algorithm optimized for NDv4/NDmv4 cluster
  - Larger slice through IB => 8x slice size in large scale
  - Only 1-1 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

**AlltoAll Bus Bandwidth (Linear vs 2D Hierarchical Algorithm)**

**Meta GPT-3 MoE Model - E2E Throughput**
## Pointers

### Getting Started
- High Performance Computing (HPC) on Azure

### HPC VM Images
- Azure HPC VM Images
- GitHub Repository

### HPC VM Series
- Azure VM sizes - HPC - Azure Virtual Machines
- Azure HBv4 VMs

### HPC VM Deployment
- Sample HPC VM deployment scripts
- Azure CycleCloud
- MUG ‘20 Tutorial

### GPU VM Series
- Azure VM sizes - GPU - Azure Virtual Machines
- Azure NDv5 VMs

### Azure HPC Blogs
- Azure Compute - Microsoft Tech Community
Thank You!

Email: jijos@microsoft.com