Best Practices for Running HPC Applications on Microsoft Azure using MVAPICH2

Jithin Jose, Jon Shelley
Azure HPC Team
Agenda

✓ Overview of Microsoft Azure
✓ Azure HPC Offerings
✓ HPC Software Ecosystem
✓ HPC Deployment Models and Demo
✓ Performance Characteristics
✓ Best Practice Recommendations
Microsoft Azure

✓ Cost
✓ Global Scale
✓ Performance
✓ Security
✓ Speed
✓ Productivity
✓ Reliability
HPC Fleet in Azure

H-Series (InfiniBand)

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

N-Series (NVIDIA GPU + InfiniBand)*

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

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

*GPU-only sizes not listed
HB120rs_v2 VM Instances

- AMD Rome
- VM Cores: 120
- Clock Speed: 3.3 GHz
- Memory Bandwidth: 340 GB/sec
- Memory: 480 GB (4GB/core)
- Local Disk: 900 GB NVMe
- NVIDIA Mellanox InfiniBand Network: 200 Gbps HDR (SR-IOV)
ND40rs_v2 VM
Instances

- Intel Skylake
- VM Cores: 40
- Memory: 672 GB
- NVIDIA Mellanox InfiniBand Network: 100 Gbps EDR (SR-IOV)
- 8 x NVIDIA V100 NVIDIA NVLINK connected GPUs
  - 32 GB GPU memory per GPU
Network Features

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

- **HBv2:**
  - HDR 200Gb/s NVIDIA Mellanox 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

- Hardware Tag Matching

- Reliability/Congestion Control
  - SHIELD, Adaptive Routing
Outline

✓ Overview of Microsoft Azure
✓ Azure HPC Offerings
✓ **HPC Software Ecosystem**
✓ HPC Deployment Models and Demo
✓ Performance Characteristics
✓ Best Practice Recommendations
HPC Software Ecosystem

- Out-of-the Box CentOS-HPC VM Images
  - NVIDIA Mellanox OFED
  - MPI Libraries
    - Includes MVAPICH2, MVAPICH2X-Azure
  - HPC Libraries
  - Optimization Configurations
  - All recipes in GitHub repository
    - https://github.com/Azure/azhpc-images/

- Or, BYO Software Stack
  - Any Linux/Windows OS flavor
  - Build/Configure custom HPC Software stack
  - Prepare custom image
MVAPICH2-X Azure

- Available in all Azure CentOS-HPC images

- Feature Highlights:
  - Enhanced tuning for point-to-point and collectives
  - XPMEM Support
  - DC Support
  - Cooperative Protocol
  - Hybrid RC/UD Support
Outline

✓ Overview of Microsoft Azure
✓ Azure HPC Offerings
✓ HPC Software Ecosystem
✓ **HPC Deployment Models and Demo**
✓ Performance Characteristics
✓ Best Practice Recommendations
Prerequisites:

- Azure Account
- Azure Subscription
- Sufficient Quota
  - # Cores
  - Specific to Region/ SKU Type
Deployment Options:

- **AzureHPC Scripts**
  - Deployment Scripts tailored for HPC needs

- **CycleCloud**
  - HPC Workload manager

- **Azure Batch**
  - Cloud scale job scheduling and Compute Management

- **ARM Templates**
  - Azure Resource Manager Templates
Setting up Azure HPC Scripts

- Prerequisites for AzureHPC
  - Azure CLI
    - [https://docs.microsoft.com/cli/azure/install-azure-cli](https://docs.microsoft.com/cli/azure/install-azure-cli)
  - Other utilities: bash, jq and ssh

- Can be invoked from:
  - Azure Cloud Shell
  - Linux VM
  - Windows Ubuntu Shell

- Detailed instructions:
  - [https://github.com/Azure/azurehpc/blob/master/README.md](https://github.com/Azure/azurehpc/blob/master/README.md)
AzureHPC for Deployment

- Install AzureHPC
  
  ```
  source ~/azurehpc/install.sh
  ```

- Initialize/Configure Cluster
  
  ```
  azhpc-init -c $azhpc_dir/examples/simple_hpc_pbs -d hbv2_cluster
  # Update config.json
  # Select SKU type, instance count, region, etc.
  ```

- Deploy Cluster
  
  ```
  azhpc-build
  ```

- Connect to your Azure Cluster
  
  ```
  azhpc-connect -u hpcadmin headnode
  ```
Demo: Deploy an HPC Cluster on Azure
Overview of Microsoft Azure

Azure HPC Offerings

HPC Software Ecosystem

HPC Deployment Models and Demo

Performance Characteristics

Best Practice Recommendations
Experiment Setup

- HBv2 VM Instances
- CentOS 7.7 HPC Image
- MPI Libraries
  - MVAPICHR2 2.3.4
  - MVAPICHR2-X 2.3
- NVIDIA Mellanox OFED 5.1
• MVAPICH2, MVAPICH2-X achieves < 2us latencies
• MVAPICH2-X offers better large message latencies for intra-node transfers (XPMEM)
• MVAPICH2, MVAPICH2-X close to line rates
• Both uses same inter-node protocols
• RPUT Rendezvous protocol (MV2_RNDV_PROTOCOL=RPUT)
• MVAPICH2-X XPMEM Collectives offers better large message allreduce latencies
• 16 HBv2 nodes, 120 PPN
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
Outline

- Overview of Microsoft Azure
- Azure HPC Offerings
- What's unique
- HPC Software Ecosystem
- HPC Deployment Models and Demo
- Performance Characteristics
- Best Practice Recommendations
Prerequisite for InfiniBand support

- If using VMs:
  - Use single Availability Set for all VMs
    - Logical Grouping of Virtual Machines
  - All VMs in Availability Set will have same PKEY (InfiniBand partition key)

- If using Virtual Machine Scale Set (VMSS):
  - All VMs in VMSS will have same PKEY
  - VMSS:
    - Set of VM instances
    - Supports flexible scale up/scale down

- Check PKEY
  
  ```
  $ cat /sys/class/infiniband/mlx5_0/ports/1/pkeys/0
  $ 0x801d
  ```
Best Practices: Guest Agent Configuration

- Minimal Guest Agent Configuration
  - "Extensions.GoalStatePeriod": 300
  - "OS.EnableFirewallPeriod": 300
  - "OS.RemovePersistentNetRulesPeriod": 300
  - "OS.RootDeviceScsiTimeoutPeriod": 300
  - "OS.MonitorDhcpClientRestartPeriod": 60
  - "Provisioning.MonitorHostNamePeriod": 60

- For extremely sensitive workloads:
  - eg:
    ```
    sudo systemctl disable waagent
    <run hpc job>
    sudo systemctl enable waagent
    ```
Best Practices: Large Scale Jobs

- Use Scalable Transports
  - **Dynamic Connected Transport (DCT)**
    - Highly scalable, and supports all features of RC
    - Lesser memory footprint
    - Eg: MV2_USE_DC=1
  - **Hybrid RC/UD Transports**
    - RC for frequently communicating pairs
    - Lesser memory footprint, Avoids QP Thrashing
    - Eg: MV2_USE_UD_HYBRID

- Enable Adaptive Routing (AR)
  - AR is enabled in all non-zero Service Levels (SL)
  - To make use of AR, specify SL during job launch
    - Eg: MV2_DEFAULT_SL=1
Best Practices: NUMA Awareness

- NUMA Affinity
  - SKU/Workload Specific
  - Bind to NUMA node closer to NIC
  - Eg: `MV2_CPU_MAPPING=X`

- NUMA Binding
  - Workload specific (MV2_CPU_BINDING=numanode)

- NUMA Aware Collectives
  - NUMA Hierarchy
Best Practices: MVAPICH2 Protocols/Thresholds

- Internode:
  - RPUT protocol for Rndv Transfers
    - `MV2_RNDV_PROTOCOL=RPUT`

- Intra-node
  - Enable XPMEM (MVAPICH2-X)
    - `MV2_SMP_USE_XPMEM=1`
  - Enable XPMEM for Collectives
    - `MV2_SMP_USE_XPMEM=1 MV2_USE_XPMEM_COLL=1`
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

- MVAPICH2 on Azure

- Adaptive Routing on Azure HPC
Thank You!

jijos@microsoft.com, joshelle@microsoft.com
Microsoft