Microsoft Azure and NVIDIA:
Bringing AI, high performance computing, and graphics visualization to enterprises

Vijay Kanchanahalli
vikancha@Microsoft.com
Senior Program Manager
Azure HPC Program Management

Visit us @ Booth #1351
Azure is Microsoft’s cloud computing platform
Microsoft Global Network

Highly Available network
✓ Own & Operate L1 (Optical) and L3 (IP) Network
✓ 100k+ mi of optical fiber
✓ Connects 54 Azure regions
✓ 50 ER Sites, 160+ Partners
✓ 140+ Edge nodes and expanding – Continuously designing to have customers <25ms latency

Same network supporting all Microsoft businesses
✓ Azure, Exchange, Dynamics, Bing, Xbox, ...

*Network sites not exhaustive
HPC on Azure
One Cloud for all Workloads

Azure supports multiple stages of the cloud adoption lifecycle

1. **Existing apps**
   - Start using the benefits of a cloud platform without rewriting or rearchitecting your application

2. **Clone to cloud**
   - Optimize your application workflows to benefit from both on and off premises resources

3. **Hybrid workflows**
   - Port your entire application environment and resources to the cloud

4. **Cloud workflows**
   - Create new differentiating services and modernize your missing critical applications

5. **Cloud-native apps**
Specialized Compute Fleet

- **A**: Entry Level VMs
  - Dev/Test Workloads

- **D**: General Purpose VMs
  - Common Applications, Web servers etc

- **F**: Compute Optimized VMs
  - Gaming, Analytics

- **G**: Large Memory VMs
  - Large Databases

- **InfiniBand**

- **L**: Storage optimized VMs
  - No SQL Databases (Cassandra, MongoDB), Data warehousing

- **H**: High Performance VMs
  - Batch processing, fluid dynamics, monte carlo simulation

- **N**: GPU-enabled VMs
  - NV - Graphic based applications
  - NC – Advanced Simulation
  - ND – AI

- **FPGA***: Virtual Machines – HPC
  - FPGA Microservices – AI/Edge

- **Cray Services in Azure**
  - IB Connected
  - CPU/GPU/Storage available in cloud

- **>80,000 IOPs**
  - Premium Storage
  - Low latency, high throughput apps
GPU Acceleration in Azure

- **Visualization & Gaming**
  - Workload centric High Memory and Clock speed workstations
  - Small configuration GPU workstations for content consumption

- **Rendering**

- **HPC/Simulation**
  - Scale out using IB for multimode HPC and ML workloads on any MPI stack
  - Scale-up multi GPU VMs with fast NVLINK interconnect for high-density single box training and HPC workloads

- **Deep-Learning/AI**

---

- **NV**
- **NC**
- **ND**
Azure HPC VMs

No-compromise CPU and GPU based resources

H-Series: Most powerful CPU virtual machines with optional RDMA

- Up to 16 cores, 3.2 GHz E5-2667 V3 Haswell processor
- Up to 224 GB DDR4 memory, 14GB per core
- FDR InfiniBand @ 56 Gbps, 2.6 microsecond latency
- 2 TB of local SSD

- Up to 44 cores, Intel Xeon Platinum processor
- Up to 352 GB DDR4 memory, 8GB per core
- EDR InfiniBand @ 100 Gbps
- 700 GB NVMe

N-Series: GPU virtual machines specialized for graphic and compute-intensive workloads

- Up to 4 NVIDIA Tesla K80 GPUs
- Up to 24 cores
- Up to 224 GiB memory
- Up to 1440 GiB of local SSD
- FDR InfiniBand

- Up to 4 NVIDIA Pascal P100 GPUs
- Up to 24 cores
- Up to 448 GiB memory
- Up to 3 TB of local SSD
- FDR InfiniBand

- Up to 4 NVIDIA Volta V100 GPUs
- Up to 24 cores
- Up to 448 GiB memory
- Up to 3 TB of local SSD
- FDR InfiniBand

- Up to 4 NVIDIA Pascal P40 GPUs
- Up to 24 cores
- Up to 448 GiB memory
- Up to 2,948 GiB of local SSD
- FDR InfiniBand

- Up to 8 NVIDIA V100 GPUs
- 40 cores
- 672 GiB memory
- 3 TB of local SSD

- Up to 4 NVIDIA Tesla M60 GPUs
- Up to 48 cores
- Up to 448 GiB memory
- Up to 2,948 GiB of local SSD
GPU VM Triage

Start at the top and work down to find a GPU VM Solution

<table>
<thead>
<tr>
<th>Workload Fit:</th>
<th>VM / GPU Solution:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large Dataset (CFD / FEA):</td>
<td>NV_v3, M60</td>
</tr>
<tr>
<td>Conventional CAD / Modeling:</td>
<td>NV, M60</td>
</tr>
<tr>
<td>Large Model:</td>
<td>NC_v3, V100 PCIe</td>
</tr>
<tr>
<td>Large Batch Size:</td>
<td>ND, P40</td>
</tr>
<tr>
<td>General Purpose:</td>
<td>NC_v2, P100</td>
</tr>
<tr>
<td>Simple Models:</td>
<td>NC, K80</td>
</tr>
<tr>
<td>Cost-effective development VM:</td>
<td>NC_v2, P100, V100 PCIe</td>
</tr>
<tr>
<td>Cost-effective deployment:</td>
<td>ND_v2, V100 SXM + NVlink</td>
</tr>
<tr>
<td>Multi-GPU optimized (6-8 GPUs):</td>
<td>NCr_v3, V100 PCIe + FDR</td>
</tr>
<tr>
<td>Large jobs (8-500 GPUs):</td>
<td></td>
</tr>
<tr>
<td>Exploration &amp; Education:</td>
<td></td>
</tr>
<tr>
<td>General Purpose:</td>
<td>NC_v2, P100</td>
</tr>
<tr>
<td>Large Textures &amp; High Resolution:</td>
<td>ND, P40</td>
</tr>
</tbody>
</table>
NVIDIA in Azure
Platform of choice for Accelerated computing across Graphics, AI and HPC

- **Broad offering with the N-series.** GPUs from M60 to V100 to support ML & DL training & inference, HPC across industries and Graphics workloads

- **Integration with key Microsoft solutions** like AML and ONNX for end-to-end acceleration for the most complex workloads, reducing burden of producing world-class solutions

- **Versatility across applications and frameworks** to support engineering on one platform and solution for the entire team, regardless of the preferences of particular engineers

- All NVIDIA acceleration software housed in NGC via Azure marketplace to ensure users have the easiest experience packaging what they need to achieve best performance in no time.

- **Optional InfiniBand interconnect** enables scale-up performance
For Every User, Any App

Knowledge Worker
Give all your business users an uncompromised experience on any device.

Creative & Technical Professional
Empower your professional graphics users with greater flexibility.

Scalability

Performance
Announcing General Availability of our newest GPU visualization VM offering – NVv3
NV v3 – Updated GPU Visualization Platform

- Get faster results for your graphic intensive 2D and 3D applications with visualization optimized GPU instances featuring NVIDIA Tesla M60 GPUs
- Doubled memory from previous generation (up to 448 GB)
- Premium storage support (SSD backed)
- Hyperthreaded enabled for 2x threads/size
- Grid license included with each GPU instance
- 25 concurrent users for virtual desktops per GPU (vPC/VApps license)
- Specs:
  - 2048 NVIDIA CUDA cores per GPU
  - 36 H.264 1080p30 streams
  - GPU Memory 8 GB/GPU
NV v3 – Regional Availability

• Generally available in
  • US West
  • US East
  • US South Central
  • Europe West
  • Europe North

• Expanding to more regions in the coming months.
Azure HPC platform

Developers

- SaaS / Client Solution (Internal and external)

HPC End-users, IT Staff, Line of Business Mgr

- Cluster templates to run existing, on-prem HPC applications on Azure

Azure Batch

- VM Management & Job Scheduling

Azure CycleCloud

- Hybrid & Cluster Manager for HPC/AI

Cloud Services, VMs, VMSS

Hardware
Azure Batch

Enable applications and algorithms to easily and efficiently run in parallel at scale

- Rendering
- Media transcoding & pre-/post-processing
- Test execution
- Monte Carlo simulations
- Genomics
- Deep Learning
- OCR
- Data ingestion, processing, ETL
- R at scale
- Compiled MATLAB
- Engineering simulations
- Image analysis & processing
Azure Batch capabilities

**Access via API’s, CLI’s, and UI’s:**
- .NET, Java, Node.js, Python, REST
- PowerShell, x-plat Azure CLI
- Azure Portal, Batch Labs x-plat client UI

**Choice of VMs:**
- Windows or Linux
- Standard or custom images
- Windows pool can use AHUB
- Use low-priority VMs

**Rich app management:**
- Get apps from blobs, Batch app packages, package managers, custom VM images
- Docker container images

**VM networking:**
- Pool VMs can be in a VNET

**Job scheduling:**
- Supports both embarrassingly parallel and tightly coupled MPI jobs
- Run > 1 task in parallel per node
- Detect and retry failed tasks
- Can set max execution time for jobs and tasks
- Task dependencies
- Job prep and cleanup tasks

**VM scaling:**
- Manual or automatic

**Monitoring:**
- VM monitoring and auto-recover
- Metrics and logs available via Portal and API
Microsoft Retail Stores

Render animations during design process and render final videos for over 80 MS store video walls

- Each of > 80 stores has unique screen layout
- Adobe After Effects used to author 2D assets which need to be laid out and rendered
- Final videos rendered at extreme resolutions; e.g. 30,000 pixels wide
- ffmpeg used to split up video into files for 4x4 screens; synchronized playback in stores
- Need to render ~700 videos
- Old 4 – 5 week process reduced to 5 hours
Designers create comps/assets in Adobe After Effects and “solve” for each store layout

On Demand Cloud Rendering

Review proposed content w/ Biz

Quality Assurance Review

Playlist is Updated and Scheduled

Deploy & Playback

Daily Camera Checks

Video Wall Pipeline | Overview
1. Creative Agencies create motion content on their PCs that requires rendering.

2. Creative Agencies upload content to Azure Blob Storage (available space automatically scales as needed. Currently 12TB generated in last 3 months).

3. Azure Batch automatically divides, distributes, and manages render jobs across all projects and content.

4. Azure VMs are spun up by Azure Batch to meet rendering demand. Currently configured to 880 vCPU cores, can be upgraded with simple configuration.

5. Azure Batch automatically collects rendered output and stores in Azure Blob Storage. Content is ready for deployment to all stores.
Thank you

Visit us @ Booth #1351