Leveraging Microsoft Azure’s GPU N-Series for Compute and Visualization

Karan Batta
Microsoft Azure
Vision

- GPU based capabilities in cloud infrastructure
- High end performance
- Not “one size fits all” approach
- Deliver and empower developer scenarios
- Achieve true “Pipeline in the Cloud”
- Critical workloads
High Performance Computing in the Cloud

Many individual tasks

Tasks assigned to VMs

Independent or coupled tasks
Workflow

- Upload data
- Submit job
- Split job/setup execution pipeline
- Manage data
- Return results

(A) Azure
- High-scale job execution over elastic VM pools
- GPU VMs
- Outputs
  - Analytics
  - Dynamic Modelling
  - Virtual Desktops

(B) APP
- {REST API}
- Seemlessly manage data dependencies for jobs
- Authentication

(C) GPU Visualization
Where?

Finance
• FX Options
• Risk Management
• Hedge Fund Management

Manufacturing & Oil/Gas
• Automotive design
• Reservoir modelling
• Manipulation of models & parts

Media
• Streaming games/video
• Transcoding
• Social media analysis

Rendering
• VFX/Ray-Tracing rendering
• CAD applications
• Simulations
Technology

- DDA (Discrete Device Assignment)
- Introduced in Windows Server 2016
- Pass-through PCIe devices
- Allows for close to bare-metal performance
Performance (CNTK)

The chart shows the average samples per second for different resources with varying numbers of GPUs. The resources are labeled as follows:

- CPU
- 1 GPU
- 2 GPUs
- 3 GPUs
- 4 GPUs

The y-axis represents the number of samples per second, ranging from 0 to 40,000. The x-axis represents the resource type. The chart indicates an increase in average samples per second as the number of GPUs increases, with a linear trend line showing the expected average samples per second for each resource type.
Architecture

- Custom Applications
- Data and Applications from the Azure Marketplace
- Bring your own Image
- Azure VM Marketplace Images

- Hyper-V
- DDA

- NVIDIA M60 GPU (Viz SKU)
- NVIDIA K80 GPU (Compute SKU)
## Visualization VMs

<table>
<thead>
<tr>
<th></th>
<th>NV6</th>
<th>NV12</th>
<th>NV24</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cores</strong></td>
<td>6 (E5-2690v3)</td>
<td>12 (E5-2690v3)</td>
<td>24 (E5-2690v3)</td>
</tr>
<tr>
<td><strong>GPU</strong></td>
<td>1 x M60 GPU (1/2 Physical Card)</td>
<td>2 x M60 GPU (1 Physical Card)</td>
<td>4 x M60 GPU (2 Physical Cards)</td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>56 GB</td>
<td>112 GB</td>
<td>224 GB</td>
</tr>
<tr>
<td><strong>Disk</strong></td>
<td>~380 GB SSD</td>
<td>~680 GB SSD</td>
<td>~1.5 TB SSD</td>
</tr>
<tr>
<td><strong>Network</strong></td>
<td>Azure Network</td>
<td>Azure Network</td>
<td>Azure Network</td>
</tr>
</tbody>
</table>


## Compute VMs

<table>
<thead>
<tr>
<th></th>
<th>NC6</th>
<th>NC12</th>
<th>NC24</th>
<th>NC24r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cores</td>
<td>6 (E5-2690v3)</td>
<td>12 (E5-2690v3)</td>
<td>24 (E5-2690v3)</td>
<td>24 (E5-2690v3)</td>
</tr>
<tr>
<td>GPU</td>
<td>1 x K80 GPU (1/2 Physical Card)</td>
<td>2 x K80 GPU (1 Physical Card)</td>
<td>4 x K80 GPU (2 Physical Cards)</td>
<td>4 x K80 GPU (2 Physical Cards)</td>
</tr>
<tr>
<td>Memory</td>
<td>56 GB</td>
<td>112 GB</td>
<td>224 GB</td>
<td>224 GB</td>
</tr>
<tr>
<td>Disk</td>
<td>~380 GB SSD</td>
<td>~680 GB SSD</td>
<td>~1.5 TB SSD</td>
<td>~1.5 TB SSD</td>
</tr>
<tr>
<td>Network</td>
<td>Azure Network</td>
<td>Azure Network</td>
<td>Azure Network</td>
<td>Azure Network + RDMA (RoCE)</td>
</tr>
</tbody>
</table>
Demo
Thank you!