INTRO TO NVIDIA GPUS IN AZURE

In Kee Paek, 10/07/2016
Cloud Data Solution Architect
Microsoft
GPU VIRTUALIZATION VISION

- Deliver accelerated graphics & compute capabilities in Azure infrastructure
- High end performance
- Not “Swiss-army knife” offering
- Helps achieve true “HPC in the Cloud”
- Close partnership with NVIDIA
FINANCE

- FX Options
- Risk Management
- Fixed Income Analytics
- Market data analysis
- Hedge Fund management
MANUFACTURING
OIL & GAS

✓ Dynamic manipulation of Engineering models and parts
✓ Realistic material creation
✓ Reservoir modelling and seismic processing
✓ Automotive engineering
MEDIA

- Stream high fidelity video games
- Encoding and transcoding
- Image processing
- Social media sentiment analysis
RENDERING

- Visual Effects (VFX)
- Ray-Tracing rendering
- Advertising & Marketing
- CAD Applications in Architecture
- Simulations
TECHNOLOGY

- DDA - Discrete Device Assignment
- Introduced in Windows Server 2016 as part of Hyper-V
- Pass-through PCIe devices directly to a Guest VM
- Allows for close to bare-metal performance
DDA (DISCREET DEVICE ASSIGNMENT)

- Entire device is mapped into the VM just as it would be running on bare metal
- Allows for full access to capabilities of that device as well as allowing the device’s native driver to be used
- Each device may be mapped to a single VM (1 device to 1 VM), but multiple devices can be mapped into the same VM (1 VM to many devices)
ARCHITECTURE

- Custom Applications
- Services/Data from the Azure Marketplace
- Bring your own OS Image
- Azure VM Marketplace Images

- Hyper-V
- DDA

- NVIDIA M60 GPU (Viz SKU)
- NVIDIA K80 GPU (Compute SKU)
## COMPUTE VIRTUAL MACHINES

<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</td>
<td>12</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>GPU</td>
<td>1 K80 GPU (1/2 Physical Card)</td>
<td>2 K80 GPUs (1 Physical Card)</td>
<td>4 K80 GPUs (2 Physical Cards)</td>
<td>4 K80 GPUs (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>InfiniBand</td>
</tr>
</tbody>
</table>
TESLA K80 - “IT’S FAST...”

- LINPACK
- SGEMM
- DGEMM
- GROMACS
- NAMD
- LAMMPS
- AMBER14
- QMCPACK
- LSMS
- miniFE
- Cloverleaf
- MILC
- Chroma

Benchmarks
Molecular Dynamics
Quantum Chemistry
Physics
RATE OF IMPROVEMENT

Image Recognition
IMAGENET

Pedestrian Detection
CALTECH

Object Detection
KITTI
### VISUALIZATION VIRTUAL MACHINES

<table>
<thead>
<tr>
<th></th>
<th>NV6</th>
<th>NV12</th>
<th>NV24</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cores</td>
<td>6</td>
<td>12</td>
<td>24</td>
</tr>
<tr>
<td>GPU</td>
<td>1 M60 GPU (1/2 Physical Card)</td>
<td>2 M60 GPUs (1 Physical Card)</td>
<td>4 M60 GPUs (2 Physical Cards)</td>
</tr>
<tr>
<td>Memory</td>
<td>56 GB</td>
<td>112 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>
</tr>
<tr>
<td>Network</td>
<td>Azure Network</td>
<td>Azure Network</td>
<td>Azure Network</td>
</tr>
</tbody>
</table>
REMOTE DESKTOP SERVICES

- Graphics performance improvements through DDA
- Enhanced support for OpenGL and OpenCL
- Connection Broker capabilities to handle massive concurrent connections
- Edge support
- Pen support
- Windows 10 and Mac preview available
NV VM DEMO
COLLABORATION WITH CNTK

✓ First class citizen
✓ Scalability - multi-GPU-multi-VM
✓ Performance
✓ Internal use-cases across various Microsoft properties and products
✓ (DSVM) Data Science VM by Azure Machine Learning
✓ N-Series and CNTK works really well together
CNTK PERFORMANCE ON DDA

- Avg. Samples/Sec
- Linear (Avg. Samples/Sec)

Samples per Second

- CPU: 2670
- 1 GPU: 10560
- 2 GPUs: 18755
- 3 GPUs: 27575
- 4 GPUs: 35750

Resource

Microsoft
SUMMARY

 ✓ High end platform
 ✓ Support OpenGL and DirectX
 ✓ Support OpenCL and CUDA
 ✓ Support existing libraries and tools
 ✓ Can utilize Azure higher-level services on top of GPU
 ✓ Fulfill both Compute and Visualization use-cases
RESOURCES

• DDA (Direct Device Assignment)

• Preview Announcement

• Tuesdays with Corey
THANK YOU

JOIN THE CONVERSATION
#GTCxKorea2016