Visualization Capabilities on Azure’s New N-Series

SC15 - November 18, 2015
Chris Huybregts
Senior Program Manager
Who

- Senior Program Manager – Virtual GPU Team, Remote Desktop
- Representing HyperV & Remote Desktop

Previously
- PM on Cloud Game Streaming project
- Dev Lead Bing Vision – Augmented Reality Project
- Software Engineer – Gaming Hardware, Interactive Entertainment, Defense Contractor
## Azure – GPU Accelerated SKUs

<table>
<thead>
<tr>
<th>Visualization SKUs</th>
<th>N1</th>
<th>N2</th>
<th>N10</th>
<th>N11</th>
<th>N12</th>
<th>N21</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CPU Cores (E5-2690v3)</strong></td>
<td>6</td>
<td>24</td>
<td>6</td>
<td>12</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td><strong>RAM (GB)</strong></td>
<td>64</td>
<td>256</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>256</td>
</tr>
<tr>
<td><strong>SSD (TB)</strong></td>
<td>~0.5</td>
<td>~2.0</td>
<td>~0.5</td>
<td>~1.0</td>
<td>~2.0</td>
<td>~2.0</td>
</tr>
<tr>
<td><strong>Network</strong></td>
<td>Azure Network</td>
<td>Azure Network</td>
<td>Azure Network</td>
<td>Azure Network</td>
<td>Azure Network</td>
<td>Azure Network RDMA Dedicated Back End</td>
</tr>
<tr>
<td><strong>GPU Resources</strong></td>
<td>1 x M60 GPU</td>
<td>4 x M60 GPU</td>
<td>1 x K80 GPU</td>
<td>2 x K80 GPUs</td>
<td>4 x K80 GPUs</td>
<td>4 x K80 GPUs</td>
</tr>
</tbody>
</table>
Pyramid Model of Visualization User Types

1. **Power Users**
   - Full GPU Resource Utilization
     - Cad/Cam/PLM
     - Gaming
     - HPC

2. **Professional Users**
   - High Burst GPU Utilization
     - Designer apps
     - Data Visualization

3. **Knowledge Worker**
   - Minimal GPU Utilization
     - Data Entry/LOB
     - Web Browser Based

Addressable Market

Single Session
- Client/Server VM

Multi-Session Or App Remoting
Approach Roadmap

Azure's Initial Offering

PARAVIRTUALIZATION (RemoteFX - vGPU)
- MS vGPU Driver
- GPU

HARDWARE PASS-THROUGH (Discrete Device Assignment)
- IHV Driver
- GPU

HARDWARE VIRTUALIZATION (GPU PARTITIONING)
- IVH Driver
- GPU

NVIDIA GRID 2.0

NVIDIA GRID vGPU
Discrete Device Assignment (DDA)

- Hardware Pass-through
Capabilities

- Entire GPU Available to the Guest
  - Includes HW Encoders
- Full IHV Driver Feature Set Available
  - DirectX, OpenGL, OpenCL, Cuda
- Guest Support
  - Server 2012r2, Server 2016, Linux
Performance

- Near Bare Metal Performance
- Full app compatibility
Linux Visualization

- Leverage your favorite 3rd party tools
Windows Visualization

- 3rd Party Tools
- Remote Desktop Services (RDS on IaaS)
Microsoft Remote Desktop App

Multiple device platforms
Windows
Windows Phone 8.1
iOS
Mac OS X
Android

Provides access to
Session-based desktops
RemoteApp programs

Great Windows experience
Productivity with secure data
RDS on IaaS

- Simplified Deployment (as few as 2 VMs required)
- AVC 444 with Hardware Encode/Decode
- OpenGL, DirectX, Cuda, OpenCL support
- Personalized Desktops
- Investigating Multi User Session Support

Demo at the Microsoft Booth
H.264/AVC 444 Mode

- HW Encode / Decode
- Enables 4:4:4 quality with standard 4:2:0 hardware
Demo

Thursday, Nov 19th – 13:00
CNTK: Open-Source, Distributed Deep Learning Systems from Microsoft
Alexey Kamenev (Microsoft Research)

Host (Seattle):
• HP Proliant SL250
• 2x Intel Xeon E52660 v2 @ 2.2GHZ
• 64 GB RAM
• 2x Nvidia K2s - (4 GPUs total)