EVALUATING WINDOWS 10: LEARN WHY YOUR USERS NEED GPU ACCELERATION

Erik Bohnhorst, Manager, ProViz Performance Engineering, NVIDIA
Nachiket Karmarkar, Senior Performance Engineer, NVIDIA
WINDOWS 10 VDI USER TESTING

CPU only vs GPU-Accelerated VDI

Workload

CPU

GPU +30%

GPU instance supported 30% higher workload

User Rating

CPU 9%

Pretty Good/PC-Native Experience

GPU 99%

Based on side-by-side testing from 136 respondents. Testing done on WebGL, Google Earth and YouTube.
WINDOWS 10 GRAPHICS USAGE

Highest graphics requirement from any operating system to date

WINDOWS 95

WINDOWS 7

WINDOWS 10

30% Increase in CPU Consumption, compared to Windows 7*

*Percent of time consuming GPU (DirectX or OpenGL)
BENCHMARKING WITH CIRRUS

Quantifying User Experience and Scale with NVIDIA Expertise

• Data driven sizing and configuration decisions
• UNIQUELY quantifies remoted user experience
  • Measures end user latency
  • Frames remoted to end users
  • Consistency of end user experience
  • Resource consumption
• Outputs realistic sizing recommendations
**TEST TO UNDERSTAND YOUR SETUP**

<table>
<thead>
<tr>
<th>Target</th>
<th>GRID vGPU</th>
<th>Remote Protocol</th>
<th>Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Host/Cluster</td>
<td>FRL</td>
<td>Blast H.264 HW</td>
<td>Benchmark Score</td>
</tr>
<tr>
<td>vCPUs</td>
<td>Allocation Policy</td>
<td>Blast H.264 SW</td>
<td>PerfMon</td>
</tr>
<tr>
<td>vRAM</td>
<td>vGPU Profile</td>
<td>Blast JPG/PNG</td>
<td>Remoted FPS</td>
</tr>
<tr>
<td>vGPU Profile</td>
<td>Scheduling Policy</td>
<td>PCoIP*</td>
<td>ESXTOP</td>
</tr>
<tr>
<td>Datastore</td>
<td></td>
<td>NVIDIA-SMI</td>
<td></td>
</tr>
<tr>
<td>Screen Resolution</td>
<td></td>
<td>Image Quality</td>
<td></td>
</tr>
<tr>
<td>Workload</td>
<td></td>
<td>End User Latency</td>
<td></td>
</tr>
<tr>
<td>Number of VMs</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Horizon 7 with PCoIP
CIRRUS
High Level Architecture

Establish Remote Connections

Provision VMs

Start performance monitoring

Start Workload

Data Collection and Analysis

Results & Report
CIRRUS
End User Latency (Click-To-Photon)

T1 = Timer Start
Response Observed

Latency = T2 - T1

T2 = Timer Stop

MouseClicked

Response Observed
## SYSTEM UNDER TEST
### Configuration Details

<table>
<thead>
<tr>
<th>Host Configuration</th>
<th>VDI Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP ProLiant DL 380 Gen9</td>
<td>vCPU - 2</td>
</tr>
<tr>
<td>Intel Xeon E5-2697v4 @ 2.30 GHz</td>
<td>vRAM - 4096 MB</td>
</tr>
<tr>
<td>VMware ESXi 6.5</td>
<td>NIC - 1 (E1000)</td>
</tr>
<tr>
<td>Number of CPUs: 36 (2 x 18)</td>
<td>Hard Disk - 32 GB</td>
</tr>
<tr>
<td>Memory: 768 GB</td>
<td>vGPU - 1 GB</td>
</tr>
<tr>
<td>Storage: All-Flash SAN (iSCSI)</td>
<td>Virtual Hardware - vmx-11</td>
</tr>
<tr>
<td>Hyperthreading, Turbo boost</td>
<td>FRL enabled - Yes</td>
</tr>
<tr>
<td>Power Setting: High Performance</td>
<td>VDI agent - VMware Horizon 7.1</td>
</tr>
<tr>
<td>GPU: 2 x M10</td>
<td>VMware Blast H.264</td>
</tr>
</tbody>
</table>

Cirrus Knowledge Worker Workload (Excel, Word, PowerPoint, Chrome, Media Player, PDF)
BEST USER EXPERIENCE WITH NVIDIA GRID

Local like latency with NVIDIA GRID

~26% better consistency in End User Latency

~200ms decrease in End User Latency

Cirrus Knowledge Worker Workload (Excel, Word, PowerPoint, Chrome, Media Player, PDF)
BEST USER EXPERIENCE WITH NVIDIA GRID
3x frames with NVIDIA GRID

Aggregate Remoted Frames

<table>
<thead>
<tr>
<th>Total Frames sent</th>
<th>NVIDIA GRID Accelerated</th>
<th>CPU Only</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>410654</td>
<td>136333</td>
</tr>
</tbody>
</table>

Cirrus Knowledge Worker Workload (Excel, Word, PowerPoint, Chrome, Media Player, PDF)
BEST BLAST IMAGE QUALITY WITH NVIDIA GRID

Blast H.264 Encoder improves the image Quality

Structural Similarity Index (SSIM)

Cirrus Knowledge Worker Workload (Excel, Word, PowerPoint, Chrome, Media Player, PDF)
NVIDIA GRID VGPU INCREASES USER DENSITY

Up to ~28% reduction in CPU utilization with NVIDIA GRID

Cirrus Knowledge Worker Workload (Excel, Word, PowerPoint, Chrome, Media Player, PDF)
CPU REDUCTION WHILE DELIVERING BEST UX

Application Performance - ~23% drop in CPU usage
TESLA M10 MEETS THE NEEDS OF KNOWLEDGE WORKERS

Tesla M10 GPU and Encode Engine match the needs of Windows 10

Cirrus Knowledge Worker Workload (Excel, Word, PowerPoint, Chrome, Media Player, PDF)
# NVIDIA GRID VGPU FOR HIGHEST DENSITY AND BEST USER EXPERIENCE

## HIGHEST DENSITY
- Up to 28% decrease in CPU utilization

## BEST USER EXPERIENCE
- ~3x more remoted frames
- ~200 ms decrease in end user latency
- Highest consistency in end user latency
- Better image quality for Blast H.264

## TESLA M10 FOR WIN10
- Meets GPU demands at scale
- Meets Encode demands at scale
- Meets Framebuffer demands at scale
DESIGNER WORKLOADS - UNDERSTANDING GPU SCHEDULING
GPU “BEST EFFORT” SCHEDULER
HOW DOES IT WORK - SIMPLIFIED VIEW

- **Time sliced Round Robin Scheduler**
  - If VM has no task or has used up its time slice the scheduler will move to the next VM
  - Cannot guarantee share of GPU cycles per VM
  - VMs can get uneven share of the GPU cycles
EQUAL SHARE SCHEDULER

HOW DOES IT WORK

- Equal Share Round Robin Scheduler
- If VM has no tasks during its time slice the GPU will idle
- Deterministic share of GPU cycles per VM
EQUAL SHARE SCHEDULER
WHAT HAPPENS WHEN A VM EXITS

VM share of GPU Cycles is relative to the other VMs on the GPU
When a VM exits the GPU cycles are shared by remaining VMs
**FIXED SHARE SCHEDULER**

**HOW DOES IT WORK**

Fixed Share Round Robin Scheduler

If VM has no tasks during its timeslice the GPU will idle

Deterministic share of GPU cycles per VM
FIXED SHARE SCHEDULER

WHAT HAPPENS WHEN A VM EXITS

VM share of GPU Cycles is Fixed, and NOT relative to the other VMs on the GPU

When a VM exits, the GPU cycles stay unused and not redistributed
## COMPARING THE SCHEDULING MODES

A high level summary cheat sheet

<table>
<thead>
<tr>
<th>Supported HW</th>
<th>BEST EFFORT</th>
<th>EQUAL SHARE</th>
<th>FIXED SHARE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maxwell, Pascal</td>
<td></td>
<td>Pascal</td>
<td>Pascal</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Primary Use cases</th>
<th>BEST EFFORT</th>
<th>EQUAL SHARE</th>
<th>FIXED SHARE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enterprise</td>
<td></td>
<td>Enterprise</td>
<td>Cloud</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>vGPU aware</th>
<th>BEST EFFORT</th>
<th>EQUAL SHARE</th>
<th>FIXED SHARE</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Needs mixed compute/graphics</th>
<th>BEST EFFORT</th>
<th>EQUAL SHARE</th>
<th>FIXED SHARE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supported</td>
<td></td>
<td>Recommended</td>
<td>Recommended</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Idle cycle redistribution</th>
<th>BEST EFFORT</th>
<th>EQUAL SHARE</th>
<th>FIXED SHARE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td></td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Guaranteed QoS</th>
<th>BEST EFFORT</th>
<th>EQUAL SHARE</th>
<th>FIXED SHARE</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Noisy neighbor protection</th>
<th>BEST EFFORT</th>
<th>EQUAL SHARE</th>
<th>FIXED SHARE</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FRL required</th>
<th>BEST EFFORT</th>
<th>EQUAL SHARE</th>
<th>FIXED SHARE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td></td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
NVIDIA Quadro vDWS with Tesla P40 Delivers Up To 2X Performance

Note: Comparing a single VM on NVIDIA Tesla M60-8Q vs a single VM on NVIDIA Tesla P40-24Q and based on SPECviewperf 12.1 benchmark.
NVIDIA Quadro vDWS with Tesla P40 Unleashes Performance at Scale

![Bar chart comparing NVIDIA Tesla M60 and Tesla P40 performance in various applications. The chart shows that NVIDIA Tesla P40 outperforms NVIDIA Tesla M60 in all applications except Energy. In Energy, NVIDIA Tesla P40 is significantly faster than NVIDIA Tesla M60.]
NVIDIA Quadro vDWS with Tesla P40

- Up to 2X Performance
- Up to 1.5X the Framebuffer
- Compute on all GRID vGPU profiles
- Quality of Service

Note: Comparing a single VM on NVIDIA Tesla M60-8Q vs a single VM on NVIDIA Tesla P40-24Q, and based on SPECviewperf 12.1 benchmark.
THANK YOU