BENCHMARKING GRAPHICS INTENSIVE APPLICATION ON VMWARE HORIZON VIEW USING NVIDIA GRID VGPU

Manvender Rawat, NVIDIA
Lan Vu, VMware
AGENDA

Overview of VMware Horizon 7 and NVIDIA GRID 2.0
How to Size VMs
Scalability testing and VMware View Planner
Test results and Important takeaways
Best Practices
VMWARE HORIZON VIEW OVERVIEW
VMWARE HORIZON VIEW OVERVIEW

Enhancing performance and user experience with GPUs
VMWARE HORIZON WITH NVIDIA GRID GPU
HOW DOES NVIDIA GRID WORK?

Virtual Desktop | Virtual Desktop | Virtual Desktop | Virtual Desktop | Virtual Desktop | Virtual Desktop | Virtual Desktop

Hypervisor

Hardware

CPUs

Server
HOW DOES NVIDIA GRID WORK?

Virtual PC
- NVIDIA Graphics Driver
- vGPU

Virtual PC
- NVIDIA Graphics Driver
- vGPU

Virtual PC
- NVIDIA Graphics Driver
- vGPU

Virtual Workstation
- NVIDIA Quadro Driver
- vGPU

Virtual Workstation
- NVIDIA Quadro Driver
- vGPU

Virtual Workstation
- NVIDIA Quadro Driver
- vGPU

Hypervisor
- NVIDIA GRID vGPU manager

Server
- CPUs

Hardware

Virtualization Layer

GPU TECH CON

GPU

H.264 Encode

NVIDIA

NVIDIA GRID

MEM

vGPU

vGPU

vGPU

vGPU

vGPU

vGPU
A Time Slice is the period of time for which a process is allowed to run in a preemptive multitasking system.

Time slicing is leveraged by hypervisors (vSphere, XenServer, KVM, Hyper-V) to share physical resources (CPU, Network, I/O etc.) between multiple virtual machines.

Time slicing allows the distribution of pooled resources based on actual need.

NVIDIA GRID uses time slicing to share the 3D engine between virtual machines.

Knowledge workers or engineers may be connected to virtual machines that share a physical GPU at the same time but typically don’t utilize the physical GPU the entire time because human workflows include:

- Getting lunch
- In a meeting
- Not in office
- Thinking
- Viewing information

During these times, the GPU isn’t under load and can be shared with other virtual machines/users.
## NVIDIA VIRTUAL GPU TYPES

<table>
<thead>
<tr>
<th>Card</th>
<th>Physical GPUs</th>
<th>GRID Virtual GPU</th>
<th>Intended Use Case</th>
<th>Frame Buffer (Mbytes)</th>
<th>Virtual Display Heads</th>
<th>Max Resolution per Display Head</th>
<th>Maximum vGPUs Per GPU Per Board</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRID K1</td>
<td>4</td>
<td>K180Q</td>
<td>Power User</td>
<td>4096</td>
<td>4</td>
<td>2560x1600</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>K160Q</td>
<td>Power User</td>
<td>2048</td>
<td>4</td>
<td>2560x1600</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>K140Q</td>
<td>Power User</td>
<td>1024</td>
<td>2</td>
<td>2560x1600</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>K120Q</td>
<td>Power User</td>
<td>512</td>
<td>2</td>
<td>2560x1600</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>K100</td>
<td>Knowledge Worker</td>
<td>256</td>
<td>2</td>
<td>1920x1200</td>
<td>8</td>
</tr>
<tr>
<td>GRID K2</td>
<td>2</td>
<td>K280Q</td>
<td>Designer</td>
<td>4096</td>
<td>4</td>
<td>2560x1600</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>K260Q</td>
<td>Power User, Designer</td>
<td>2048</td>
<td>4</td>
<td>2560x1600</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>K240Q</td>
<td>Power User, Designer</td>
<td>1024</td>
<td>2</td>
<td>2560x1600</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>K220Q</td>
<td>Power User, Designer</td>
<td>512</td>
<td>2</td>
<td>2560x1600</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>K200</td>
<td>Knowledge Worker</td>
<td>256</td>
<td>2</td>
<td>1920x1200</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 1 GRID K1 and K2 Virtual GPU types
## NVIDIA VIRTUAL GPU TYPES

<table>
<thead>
<tr>
<th>Card</th>
<th>Physical GPUs</th>
<th>GRID Virtual GPU</th>
<th>Intended Use Case</th>
<th>Frame Buffer (MB)</th>
<th>Virtual Display Heads</th>
<th>Max Resolution per Display Head</th>
<th>Maximum vGPUs Per GPU</th>
<th>Maximum vGPUs Per Board</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tesla M60</td>
<td>2</td>
<td>M60-8Q</td>
<td>Designer</td>
<td>8192</td>
<td>4</td>
<td>3840x2160</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M60-4Q</td>
<td>Designer</td>
<td>4096</td>
<td>4</td>
<td>3840x2160</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M60-2Q</td>
<td>Designer</td>
<td>2048</td>
<td>4</td>
<td>2560x1600</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M60-1Q</td>
<td>Power User, Designer</td>
<td>1024</td>
<td>2</td>
<td>2560x1600</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M60-0Q</td>
<td>Power User, Designer</td>
<td>512</td>
<td>2</td>
<td>2560x1600</td>
<td>16</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M60-2B</td>
<td>Power User</td>
<td>2048</td>
<td>2</td>
<td>2560x1600</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M60-1B</td>
<td>Power User</td>
<td>1024</td>
<td>2</td>
<td>2560x1600</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M60-0B</td>
<td>Power User</td>
<td>512</td>
<td>2</td>
<td>2560x1600</td>
<td>16</td>
<td>32</td>
</tr>
<tr>
<td>Tesla M6</td>
<td>1</td>
<td>M6-8Q</td>
<td>Designer</td>
<td>8192</td>
<td>4</td>
<td>3840x2160</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M6-4Q</td>
<td>Designer</td>
<td>4096</td>
<td>4</td>
<td>3840x2160</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M6-2Q</td>
<td>Designer</td>
<td>2048</td>
<td>4</td>
<td>2560x1600</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M6-1Q</td>
<td>Power User, Designer</td>
<td>1024</td>
<td>2</td>
<td>2560x1600</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M6-0Q</td>
<td>Power User, Designer</td>
<td>512</td>
<td>2</td>
<td>2560x1600</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M6-2B</td>
<td>Power User</td>
<td>2048</td>
<td>2</td>
<td>2560x1600</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M6-1B</td>
<td>Power User</td>
<td>1024</td>
<td>2</td>
<td>2560x1600</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M6-0B</td>
<td>Power User</td>
<td>512</td>
<td>2</td>
<td>2560x1600</td>
<td>16</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 2 Tesla M60 and M6 Virtual GPU types
WHY BENCHMARK?
BENCHMARKING VIRTUALIZED ENVIRONMENTS

Typical Workstation benchmarks designed to stress all the available system resources.

Multiple VMs running the same task at the same time is not realistic test scenario.

Most scalability tests can only simulate worst case real-user scenario.

Benchmark
ViewPerf12 Catia viewset

Human workflow
GPU "heavy" process (zooming)
NEED

There is a need for

- End to end hardware/architecture comparison over generations
- Platform optimization and fine tuning
- ISV Certification process
- Sizing the VMs for best performance.
- Finding the right number of VMs that can run on the Host with acceptable performance
- Defining a workflow to automate the test process as the consolidation numbers and VM sizing will be different for different applications and physical hardware.
METHODOLOGY & TOOLS
HOW TO RUN SCALABILITY TESTS?

- Ideal scenario would be testing with actual application users and monitoring the resource utilization over an extended period of time (days/weeks)

- LoginVSI Graphic workload
  
  Multimedia Workload

  Custom workload integration with LoginVSI

- VMware View Planner

  Solidworks

  3DMark

  Custom workload integration with View Planner

- In-house scripts for scalability test execution and log collection

  AutoIT, Python, Powershell, psexec
PERFORMANCE METRICS AND USER EXPERIENCE

How to define a great User Experience?

- Application FPS
- Application Response Time
- GPU statistics (nvidia-smi)
- Resource Utilization

- And more that needs to be defined
UX METRICS EXAMPLE

ESRI defined ArcGIS Pro UX based on following Performance Metrics:

- **Draw Time Sum**: 80:90 seconds for basic tests to complete
- **Frame Per Second**: 30-60 w/ 60 being optimal but ESRI admits 30 is ok, say users can’t tell the difference
- **FPS Minimum**: a big dip would mean the user saw a freeze, etc., below 5-10 FPS is an issue.
- **Standard Deviation**: shows tests were uniform, quantity of tests:
  - <2 for 2D
  - <4 for 3D
WORKLOADS

Created and provided by ISV

ESRI provided us with a 3D “heavy” workload: Philly 3D
Adobe Photoshop graphics workload

Generic workloads for the ISV apps

Revit RFO Benchmark
Catalyst for AutoCAD

NVIDIA created workload

AutoCAD workload was created by NVIDIA with help from AutoCAD

VMware View Planner workloads
SIZING
SIZING

Use what you already know

Size VM based on optimal physical workstation configurations

Select vGPU profile based on Frame buffer requirements

Apply all hypervisor recommended best practices

Monitor VM resource utilization for a single VM test

change VM resources based on the Max resource utilization

Important not to over-allocate VM resources for virtualized environment

Resource over allocation can reduce the performance of a VM as well as other VMs sharing the same host.

Disabling hardware devices (typically done in BIOS) can free interrupt resources
SIZING METHODOLOGY

Monitor

Run

Configure/Change
SIZING METHODOLOGY

“x” users

“y” users

“z” users
VIEW PLANNER WORKLOAD

- View Planer allows benchmarking:
  - A mix of regular workload and any applications you bring
    - Health care, education, 3D graphics, Office 365
  - Up to thousands of VDI desktops or more

For Windows

<table>
<thead>
<tr>
<th>Regular Workload</th>
<th>3D Workload</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS Office</td>
<td>Autodesk</td>
</tr>
<tr>
<td>Other Apps</td>
<td>SolidWorks</td>
</tr>
<tr>
<td></td>
<td>Siemens</td>
</tr>
<tr>
<td></td>
<td>3D Mark</td>
</tr>
</tbody>
</table>
VIEW PLANNER WORKLOAD

- View Planer allows benchmarking:
  - A mix of regular workload and any applications you bring
    - Health care, education, 3D graphics, Office 365
  - Up to thousands of VDI desktops or more

For Linux

Regular Workload

- LibreOffice
- Firefox
- PDF
- Eclipse
BRING YOUR OWN APPLICATIONS (BYOA)

- Add your own customized applications
  - Including 3D graphics ones
BENCHMARKING 3D GRAPHICS WORKLOAD WITH GRID vGPU

Set up Horizon View and View Planner

Set up GRID vGPU for VDI Desktop / RDSH Server Virtual Machine

Register 3D Applications with View Planner

Create Workload Run Profile & start Benchmarking Run

Collect Benchmarking Results

Run at Scale
SCALABILITY TESTS
NVIDIA TEST SETUP

Virtual Client VMs
- 64-bit Win7 (SP1)
- 4vCPU, 4 GB RAM
- View Client 4.0

Virtual VDI desktop VMs
- 64-bit Win7 (SP1)
- 6vCPU, 14 GB RAM, 50GB HD
- Horizon View 7.0 agent

Remote Display Protocol
Blast Extreme / PCoIP

SuperMicro SYS-2027GR-TRFH
Intel Xeon E5-2690 v2 @ 3.00GHz + 2 x Nvidia GRID K1
20 cores (2 x 10-core socket) Intel IvyBridge
256 GB RAM

SuperMicro SYS-2028GR-TRT
Intel Xeon E5-2698 v3 @ 2.30GHz + 2 x Nvidia GRID M60
32 cores (2 x 16-core socket) Intel Haswell
256 GB RAM
SPEC VIEWPERF 12 OVERVIEW

Workstation benchmark that can be used to simulate different workstation application workloads.

Tests the worst case scenario for virtualized and shared resource configuration.

Can be used to Compare relative performance difference between:

- vGPU boards (K2 vs M60).
- Various remoting Protocols (PCoIP vs Blast) and the affect on the application performance.
- Can still be used for ISV certification process

Cannot be used for sizing the VMs/Hosts.
SINGLE VM K2 VS M60 VIEWPERF 12

Increase in Performance
4vcpu / 14GB RAM
VIEWPERF12 SCALABILITY TESTS

16 VM M60 vs K2

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Average FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>16VM K240Q</td>
<td>10</td>
</tr>
<tr>
<td>16VM M60-1Q creo</td>
<td>15</td>
</tr>
<tr>
<td>32VM M60-1Q</td>
<td>12</td>
</tr>
<tr>
<td>16VM K2-K240Q</td>
<td>30</td>
</tr>
<tr>
<td>16VM M60-1Q SW</td>
<td>35</td>
</tr>
<tr>
<td>32VM M60-1Q</td>
<td>25</td>
</tr>
</tbody>
</table>
VIEWPERF12 SCALABILITY TEST

Comparing Remoting Protocols
16 VMs test (4vCPU/14GB RAM/M60_2Q)
# REVIT RFOBENCHMARK

<table>
<thead>
<tr>
<th>6vCPU 6GB RAM M60_1Q</th>
<th>Model creation and view export benchmark</th>
<th>CPU Render benchmark</th>
<th>Graphics benchmark w/ hardware acceleration</th>
<th>Graphics benchmark w/o hardware acceleration</th>
</tr>
</thead>
<tbody>
<tr>
<td>opening and loading the custom template</td>
<td>opening and loading the custom template</td>
<td>opening and loading the custom template</td>
<td>opening and loading the custom template</td>
<td>opening and loading the custom template</td>
</tr>
<tr>
<td>creating the floors and levels</td>
<td>creating the floors and levels</td>
<td>creating the floors and levels</td>
<td>creating the floors and levels</td>
<td>creating the floors and levels</td>
</tr>
<tr>
<td>creating a group of walls and doors</td>
<td>creating a group of walls and doors</td>
<td>creating a group of walls and doors</td>
<td>creating a group of walls and doors</td>
<td>creating a group of walls and doors</td>
</tr>
<tr>
<td>modifying the group by adding a curtain wall</td>
<td>modifying the group by adding a curtain wall</td>
<td>modifying the group by adding a curtain wall</td>
<td>modifying the group by adding a curtain wall</td>
<td>modifying the group by adding a curtain wall</td>
</tr>
<tr>
<td>creating the exterior curtain wall</td>
<td>creating the exterior curtain wall</td>
<td>creating the exterior curtain wall</td>
<td>creating the exterior curtain wall</td>
<td>creating the exterior curtain wall</td>
</tr>
<tr>
<td>creating the sections</td>
<td>creating the sections</td>
<td>creating the sections</td>
<td>creating the sections</td>
<td>creating the sections</td>
</tr>
<tr>
<td>changing the curtain wall panel type</td>
<td>changing the curtain wall panel type</td>
<td>changing the curtain wall panel type</td>
<td>changing the curtain wall panel type</td>
<td>changing the curtain wall panel type</td>
</tr>
<tr>
<td>export all views as PNGs</td>
<td>export all views as PNGs</td>
<td>export all views as PNGs</td>
<td>export all views as PNGs</td>
<td>export all views as PNGs</td>
</tr>
<tr>
<td>export some views as DWGs</td>
<td>export some views as DWGs</td>
<td>export some views as DWGs</td>
<td>export some views as DWGs</td>
<td>export some views as DWGs</td>
</tr>
<tr>
<td>Total</td>
<td>Total</td>
<td>Total</td>
<td>Total</td>
<td>Total</td>
</tr>
</tbody>
</table>

| PCoIP 1 VM | 3.73 | 12.83 | 27.08 | 54.07 | 15.10 | 9.38 | 5.07 | 33.07 | 39.71 | 200.04 | 117.68 | 6.93 | 6.65 | 8.75 | 2.25 | 27.48 | 28.29 | 31.84 | 11.04 |
| Blast 1 VM | 5.18 | 12.84 | 26.98 | 55.42 | 14.91 | 9.56 | 5.09 | 32.70 | 40.01 | 202.69 | 119.65 | 7.12 | 6.71 | 8.92 | 2.27 | 27.93 | 28.23 | 31.30 | 11.07 |
| PCoIP 16 VMs | 4.24 | 13.06 | 28.55 | 57.90 | 15.68 | 9.70 | 5.27 | 36.51 | 44.55 | 215.45 | 296.10 | 7.65 | 7.40 | 9.87 | 2.45 | 38.59 | 41.47 | 45.32 | 15.48 |
| Blast 16 VMs | 4.02 | 13.13 | 29.08 | 58.61 | 15.63 | 9.82 | 5.32 | 37.26 | 45.23 | 222.42 | 303.50 | 7.57 | 7.42 | 9.99 | 2.52 | 40.11 | 42.44 | 43.90 | 14.38 |
| PCoIP 29 VMs | 4.42 | 14.72 | 32.92 | 67.19 | 17.41 | 10.70 | 5.80 | 45.88 | 56.65 | 255.70 | 554.45 | 8.66 | 8.37 | 11.10 | 2.69 | 71.55 | 82.02 | 76.86 | 23.40 |
| Blast 29 VMs | 4.33 | 14.31 | 31.35 | 64.56 | 17.39 | 10.41 | 5.74 | 41.47 | 54.06 | 243.62 | 536.20 | 9.21 | 8.67 | 11.62 | 2.80 | 68.31 | 78.72 | 77.53 | 27.61 |

**Total time (Lower is better)**

![Total time (Lower is better)](image)
## ESRI ARCGIS PRO 10.0 TEST RESULTS

<table>
<thead>
<tr>
<th>Application Metrics</th>
<th>GPU</th>
<th>CPU %Core Util</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Philly 3D Map</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Think Time 5 Seconds</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Navigation Time 5 Second</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>VM Config</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intel Ivy Bridge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K240q 6vcpu 6GB RAM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VM count</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>01:11.2</td>
<td>62.34</td>
</tr>
<tr>
<td>8</td>
<td>01:15.9</td>
<td>53.48</td>
</tr>
<tr>
<td>12</td>
<td>01:22.6</td>
<td>45.32</td>
</tr>
<tr>
<td>16</td>
<td>01:32.8</td>
<td>40.7</td>
</tr>
<tr>
<td>Haswell</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K240q 6vcpu 6GB RAM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>01:11.4</td>
<td>65.85</td>
</tr>
<tr>
<td>8</td>
<td>01:16.5</td>
<td>60.92</td>
</tr>
<tr>
<td>12</td>
<td>01:17.3</td>
<td>55.25</td>
</tr>
<tr>
<td>16</td>
<td>01:20.4</td>
<td>47.28</td>
</tr>
<tr>
<td>Haswell</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M60-1Q 6vcpu 6GB RAM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>01:07.4</td>
<td>66.52</td>
</tr>
<tr>
<td>8</td>
<td>01:07.9</td>
<td>63.43</td>
</tr>
<tr>
<td>16</td>
<td>01:10.5</td>
<td>57.74</td>
</tr>
<tr>
<td>24</td>
<td>01:16.2</td>
<td>50.85</td>
</tr>
<tr>
<td>28</td>
<td>01:20.3</td>
<td>47.54</td>
</tr>
<tr>
<td>32</td>
<td>01:26.0</td>
<td>43.42</td>
</tr>
</tbody>
</table>
ESRI ARCGIS PRO 10.0 DRAW-TIME SUM
ESRI ArcGIS Pro 3D

UPH - Users per Host

ESRI Heavy 3D Workload

12 UPH

K240Q Users
6vCPU - 6GB RAM

2x NVIDIA GRID K2

Medium 3D Workload

16 UPH

K240Q Users
6vCPU - 6GB RAM

Lab host:
- CPU: Dual Socket 2.3Ghz / 16 core
- RAM: 256GB RAM
- GPU: 2 NVIDIA GRID K2 cards
- 10G Core network
- iSCSI SAN: ~25K max IOPS
- VMware vSphere 6
- VMware Horizon 6.1 w/ vGPU
- Tested 6/2015
ESRI ArcGIS Pro 3D

UPH - Users per Host

ESRI Heavy 3D Workload

24 UPH

M60_1Q Users
6vCPU - 6GB RAM

2x NVIDIA GRID M60

Medium 3D Workload

28 UPH

M60_1Q Users
6vCPU - 6GB RAM

Lab host:
CPU: Dual Socket 2.3Ghz / 16 core
RAM: 256GB RAM
GPU: 2 NVIDIA GRID M60 cards
10G Core network
iSCSI SAN: ~25K max IOPS
VMware vSphere 6
VMware Horizon 6.1 w/ vGPU
Tested 6/2015
ARCGIS PRO 10.2 SCALABILITY TEST

- 16 (M60_2Q) and 19 (1Q) VMs running ESRI ArcGIS Pro.
- Draw Time of around 01:20 minutes guarantees a great user experience.

Protocol acceleration increases users per host by 18% (3VMs) for ESRI ArcGIS Pro 1.1 3D users

Source: NVIDIA GRID Performance Engineering Lab
Protocol acceleration increases average FPS by 13% across VP12 subtests running 16 VMs (2Q). Dependent on subtests the performance difference varies between -2.02% and 25.27%.

Source: NVIDIA GRID Performance Engineering Lab
DECREASES CPU LOAD
Host CPU utilization of 19 VMs (1Q) running ESRI ArcGIS Pro.

Impact of protocol acceleration increases with the amount of pixels. 16% (1920x1080) -> 22% (2560x1440).

Source: NVIDIA GRID Performance Engineering Lab
VIEW PLANNER TESTBED

**Virtual Client VMs**
- 64-bit Win7 (SP1)
- 4vCPU, 4 GB RAM
- View Client 4.0

**Virtual VDI desktop VMs**
- 64-bit Win7 (SP1)
- 4vCPU, 20 GB RAM, 24GB HD
- Horizon View 7.0 agent

Remote Display Protocol
Blast Extreme / PCoIP

Dell R730 – Intel Haswell CPUs + 2 x Nvidia GRID K1
- 24 cores (2 x 12-core socket) E5-2680 V3
- 384 GB RAM

Dell R730 – Intel Haswell CPUs + 2 x Nvidia GRID M60
- 24 cores (2 x 12-core socket) E5-2680 V3
- 384 GB RAM
SPECapc for 3DSmax - CPU utilization at the client
## SPECapc for 3DSmax - CPU utilization at the client

<table>
<thead>
<tr>
<th># of VM</th>
<th>PCoIP (A)</th>
<th>Blast - HW H264 (B)</th>
<th>A / B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3 %</td>
<td>2 %</td>
<td>2.2 x</td>
</tr>
<tr>
<td>2</td>
<td>6 %</td>
<td>2 %</td>
<td>2.6 x</td>
</tr>
<tr>
<td>4</td>
<td>12 %</td>
<td>5 %</td>
<td>2.7 x</td>
</tr>
<tr>
<td>6</td>
<td>20 %</td>
<td>7 %</td>
<td>2.9 x</td>
</tr>
<tr>
<td>8</td>
<td>27 %</td>
<td>9 %</td>
<td>3.1 x</td>
</tr>
<tr>
<td>10</td>
<td>32 %</td>
<td>11 %</td>
<td>3.0 x</td>
</tr>
<tr>
<td>12</td>
<td>41 %</td>
<td>13 %</td>
<td>3.2 x</td>
</tr>
<tr>
<td>14</td>
<td>48 %</td>
<td>15 %</td>
<td>3.2 x</td>
</tr>
<tr>
<td>16</td>
<td>54 %</td>
<td>16 %</td>
<td>3.4 x</td>
</tr>
</tbody>
</table>
SPECapc for 3DSmax - CPU utilization at the server
SPECapc for 3DSmax - CPU utilization at the client

<table>
<thead>
<tr>
<th># of VM</th>
<th>PCoIP (A)</th>
<th>Blast - HW H264 (B)</th>
<th>A / B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5 %</td>
<td>5 %</td>
<td>1.1 x</td>
</tr>
<tr>
<td>2</td>
<td>10 %</td>
<td>9 %</td>
<td>1.1 x</td>
</tr>
<tr>
<td>4</td>
<td>20 %</td>
<td>18 %</td>
<td>1.1 x</td>
</tr>
<tr>
<td>6</td>
<td>30 %</td>
<td>27 %</td>
<td>1.1 x</td>
</tr>
<tr>
<td>8</td>
<td>42 %</td>
<td>37 %</td>
<td>1.1 x</td>
</tr>
<tr>
<td>10</td>
<td>53 %</td>
<td>48 %</td>
<td>1.1 x</td>
</tr>
<tr>
<td>12</td>
<td>63 %</td>
<td>61 %</td>
<td>1.03 x</td>
</tr>
<tr>
<td>14</td>
<td>80 %</td>
<td>68 %</td>
<td>1.2 x</td>
</tr>
<tr>
<td>16</td>
<td>90 %</td>
<td>74 %</td>
<td>1.2 x</td>
</tr>
</tbody>
</table>
SPECapc for 3DSMax 2015 -
Average FPS per VM delivered at the client
BEST PRACTICES
BEST PRACTICES

We have seen Host Turboboost setting greatly impact performance. Evaluate with or without this Bios setting for your use case.

For the Host CPU, We have seen that the higher number of cores impact scalability more than higher clock speed.

Consider distributing the VMs evenly across all the GPUs

Try to size the VM within the NUMA node boundaries.

Proper single VM sizing very important for higher scalability.
THANK YOU

JOIN THE CONVERSATION
#GTC16  

JOIN THE NVIDIA DEVELOPER PROGRAM AT  developer.nvidia.com/join