NVIDIA GRID™ APPLICATION SIZING GUIDE FOR: AUTODESK MAYA 2017

NVIDIA Performance Engineering
November 2016
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users per Server (UPS)</td>
<td>3</td>
</tr>
<tr>
<td>Determining users per server</td>
<td>4</td>
</tr>
<tr>
<td>Scalability Test Result Summary</td>
<td>4</td>
</tr>
<tr>
<td>About the Application: Maya 2017</td>
<td>5</td>
</tr>
<tr>
<td>About NVIDIA GRID</td>
<td>6</td>
</tr>
<tr>
<td>Testing and methodology description</td>
<td>7</td>
</tr>
<tr>
<td>Test Descriptions</td>
<td>8</td>
</tr>
<tr>
<td>Methodology Description</td>
<td>14</td>
</tr>
<tr>
<td>Appendix</td>
<td>20</td>
</tr>
</tbody>
</table>
The purpose of this guide is to give a detailed analysis of how many users organizations can expect to get per servers based on performance testing with the Autodesk Maya applications. The NVIDIA Performance Lab worked in cooperation with the Autodesk Maya team to determine the maximum recommended number of users for the reference server configuration. Autodesk provided specific test scripts for the Maya 2017 software that are not generally available to end users, which the NVIDIA GRID performance team used to test specific combinations of virtual machines for performance and scale. These test scripts have been submitted to the SPEC review committee for inclusion in a Maya 2017 SPEC benchmark. Based on extensive testing, NVIDIA GRID provides the following performance and scalability recommendation.

**NOTE:** THESE NUMBERS ARE INTENDED TO BE USED A GENERAL GUIDANCE FOR A MAXIMUM NUMBER OF USERS PER HOST. CHANGES IN SERVER CONFIGURATION AND DIFFERENCES IN USAGE OF THE SOFTWARE WILL CAUSE PERFORMANCE TO VARY.

**AutoDesk Maya 2017**

**UPS - Users per Server**

- **Maya XGen Hair (OpenCL)**
  - M60-8Q Users
  - 8vCU - 8GB RAM
  - 4\textsubscript{UPS}

- **Maya Modeling/ Tessellation usage persona**
  - M60-2Q Users
  - 4vCU - 8GB RAM
  - 16\textsubscript{UPS}

- **Maya Animation User**
  - M60-2Q
  - 8vCPU - 8GB RAM
  - 10\textsubscript{UPS}

- **Maya Bifrost Simulation**
  - M60-1Q
  - 10vCPU - 8GB RAM
  - 6\textsubscript{UPS}
DETERMINING USERS PER SERVER

Since Maya 2017 is used as both modeling and animation software, both user functions are important to consider when determining sizing of a virtual environment. Our testing reveals that these functions have differences between the amount of GPU used for each function that can be quite significant. Also, the tests have shown that the availability of vCPU resources (i.e. cores) is critical to performance for animation workloads.

These tests were written to replicate typical artist workflows which use effects and take advantage of the GPU.

The following tests were provided by Autodesk:

Table 1 Maya Test categories

<table>
<thead>
<tr>
<th>Test area</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bifröst simulation</td>
<td>Bridge</td>
</tr>
<tr>
<td>Animation</td>
<td>Jungle</td>
</tr>
<tr>
<td>Modeling/Tessellation</td>
<td>Character</td>
</tr>
<tr>
<td>XGen Hair</td>
<td>Tiger</td>
</tr>
</tbody>
</table>

For simplicity purposes, test will be often referenced by their shorter name, i.e. Character, Jungle, Bridge and Tiger, throughout this scalability report.

SCALABILITY TEST RESULT SUMMARY

The test results show that XGen modeling workloads, or any workload that requires OpenCL, the maximum possible number of users per server is four. Testing shows that with 4 concurrent M60-8Q profiles running simultaneously, all users see less than a 10% degradation in performance, well below discomfort level. We conclude that 4 concurrent users per server is the maximum. The Character modeling/tessellation test can support 16 VMs/Users per server, which is the maximum number of VM’s for the M60-2Q profile. Both the Tiger and Character workflow tests achieve the maximum number of VMs per GPU board, but the tests themselves are not GPU intensive. The frame buffer allocated to the VM was not entirely used by the application during test execution.

The Jungle animation and Bifröst simulation Bridge tests exceeded acceptable usability levels when reaching 10 VMs, and the Bridge test was found to perform better when more vCPU was added to the VM. Both tests put a larger demand on the CPU than GPU. In the case of the Bridge test, reconfiguring vCPU to 10 increased the performance
of the VM. However, as such, the host CPU would become too overallocated if the host had **more than 10 VM's running**. On this class of server, the maximums VM's which are concurrently executing for Bifröst simulation or the Character animation work modelled within the test **reside between 6 and 8 users per server**.

### ABOUT THE APPLICATION: MAYA 2017

Maya provides a comprehensive suite of tools for your 3D content creation work ranging from modeling, animation, and dynamics through to painting and rendering to name but a few. With Maya, you can create and edit 3D models in a variety of modeling formats and animate your models using Maya’s suite of animation tools. Maya also provides a range of tools to allow you to render your animated 3D scenes to achieve photo realistic imagery and animated visual effects.

GPU Override is feature which is enabled by default in Maya 2017, this feature allows the GPU to perform operations which would otherwise be performed on the CPU. GPU Override requires Viewport 2.0 to be active, and the Evaluation Manager must be set to Serial or Parallel modes (Parallel is recommended and set by default). When the Maya preference “GPU Override” is *enabled* in Maya 2017, then standard deforming animation (i.e. Character animation) will offloaded to the GPU rather than CPU. If a user disables “GPU Override”, then these deforming animations would be processed on the CPU.

It is important to note, not all effects will be performed by the GPU even when GPU override is enabled. The following is a list of effects which make more use of the GPU:

- Scenes which use Viewport 2.0 effects such as lighting, shadows, SSAO, motion blur, depth peeling transparency.
- Playback and tumbling with very dense geometry.
- Generally, operations that use OpenCL (i.e. XGen).
- OpenSubdiv smoothing.

The following is list of effect which are CPU limited and do not take advantage of the GPU:

- Playback or tumbling in scenes with very large numbers of objects.
- Rigging in general (i.e. editing Character rigs).
- Modeling on dense meshes.
- Particle simulation and caching.
ABOUT NVIDIA GRID

NVIDIA redefined visual computing by giving designers, engineers, scientists, and graphics artists the power to take on the biggest visualization challenges with immersive, interactive, photorealistic environments. NVIDIA Tesla GPU Accelerators bring this same power to the virtual desktop. Leveraging NVIDIA Tesla™ GPUs, NVIDIA GRID™ delivers virtual workstations from the data center or cloud. Architects, engineers, and designers are now liberated from their desk and can access their applications and data anywhere. The NVIDIA Tesla M60 GPU accelerator works with NVIDIA GRID software to provide the industry’s highest user performance for virtualized workstations, desktops, and applications. This solution allows enterprises to virtualize any application—including professional graphics applications—and deliver them out to any device, anywhere.

Table 2 Tesla M60 Features and Specs

<table>
<thead>
<tr>
<th>Virtualization Use Case</th>
<th>Performance-Optimized Graphics Virtualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU Product</td>
<td>NVIDIA Tesla M60 - designed for the datacenter</td>
</tr>
<tr>
<td>GPU Architecture</td>
<td>NVIDIA Maxwell™</td>
</tr>
<tr>
<td>GPUs per Board</td>
<td>2</td>
</tr>
<tr>
<td>Max User per Board</td>
<td>32(16 per GPU)</td>
</tr>
<tr>
<td>NVIDIA CUDA Cores</td>
<td>4096 NVIDIA CUDA Cores (2048 per GPU)</td>
</tr>
<tr>
<td>GPU Memory</td>
<td>16 GB of GDDR5 Memory (8 per GPU)</td>
</tr>
<tr>
<td>H.264 1080p30 Streams</td>
<td>36</td>
</tr>
<tr>
<td>Max Power Consumption</td>
<td>300 W</td>
</tr>
<tr>
<td>Thermal Solution</td>
<td>Active/Passive</td>
</tr>
<tr>
<td>Form Factor</td>
<td>PCIe 3.0 Dual Slot</td>
</tr>
</tbody>
</table>
TESTING AND METHODOLOGY DESCRIPTION

This section describes the tests performed and the method of testing used to determine sizing and server loads.

Test Environment

The following table describes the server host:

Table 1 Reference host specifications

<table>
<thead>
<tr>
<th>Reference Host</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Supermicro SYS-2028GR-TRT</td>
</tr>
<tr>
<td>CPU</td>
<td>Intel Xeon CPU e5-2698 v3 @2.30Ghz</td>
</tr>
<tr>
<td>Logical Processors</td>
<td>64</td>
</tr>
<tr>
<td>Memory</td>
<td>256 GB</td>
</tr>
<tr>
<td>Storage</td>
<td>26 TB</td>
</tr>
<tr>
<td>GPU’s</td>
<td>2X Tesla M60</td>
</tr>
<tr>
<td>Hypervisor</td>
<td>VMware ESXi, 6.0.0, 3380124</td>
</tr>
<tr>
<td>VDI Management Software</td>
<td>VMware Horizon 7.01</td>
</tr>
<tr>
<td>NVIDIA GRID Software</td>
<td>Virtual Workstation</td>
</tr>
<tr>
<td>NVIDIA GRID Software</td>
<td>369.17(host) 367.43(guest)</td>
</tr>
</tbody>
</table>

Server BIOS Settings

<table>
<thead>
<tr>
<th>Setting</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turbo Boost Technology</td>
<td>Enabled</td>
</tr>
<tr>
<td>Power Settings</td>
<td>Maximum Performance</td>
</tr>
<tr>
<td>Hyperthreading</td>
<td>Enabled</td>
</tr>
</tbody>
</table>

The VM operating system was Windows 7 SP1 with VMware Tools was installed. All testing was run on a single, full HD (1920x1080) monitor.
Test Descriptions

Modelling/Tessellation: Character

The following is a screenshot of the Character test during test execution:

Figure 1 Character test screen shot

Display settings, such as number of vertices, edges, faces, tris and UV’s, determine the complexity of the model. The following screenshot describes the display settings used for the test:

Figure 2 Character test display settings
**Animation: Jungle**

The Jungle test provides basic animation information for a specific object in a scene. Below are two screenshots taken during Jungle test execution:

![Figure 3 Jungle test screenshot](image1.png)

*Figure 3 Jungle test screenshot*

![Figure 4 Jungle test screenshot](image2.png)

*Figure 4 Jungle test screenshot*
Color shading is turned on and off during the tests. The following screenshot describes the complexity of the scene:

![Jungle test display settings](image)

**Figure 5 Jungle test display settings**

**Bifröst Simulation: Bridge**

Autodesk describes Bifröst as a procedural framework that can create simulated liquid and aerodynamic effects using a FLIP (fluid implicit particle) solver. You can generate liquid from emitters and have it fall under gravity, as well as interact with colliders to direct the flow and create splashes, and use fields to create jets and other effects.

![Bridge test screenshot](image)

**Figure 6 Bridge test screenshot**
The following screenshot describes the model:

![Vertices, Edges, Faces, Tris, UVs](image)

**Figure 7 Bridge test display settings**

**XGen Hair: Tiger**

XGen is described by Autodesk as users can procedurally create and style hair, fur and feathers for characters. **With XGen users can** create realistic, self-shadowing fur and short hair on multi-surface NURBS, polygonal and subdivision surface models. The Tiger tests is the most GPU intensive test and requires OpenCL. The following is a screenshot of the Tiger test during test execution:

![Tiger test screenshot](image)

**Figure 8 Tiger test screenshot**

The following screenshot shows the heads up display settings of the scene:

![Vertices, Edges, Faces, Tris, UVs](image)

**Figure 9 Tiger test display settings**

**Importance of vCPU vs. vGPU**

Configuring the virtualized environment is essential for providing a consistently performant user experience. Improper configuration can lead to errors within the
application, sluggish performance or worse yet, application crashes. Autodesk has provided some error detection within the application to assist users in understanding issues which that may use to address issues which can occur due to insufficient GPU resources.

For Maya to take advantage of the GPU and multi-threading, the test scripts automatically enabled Parallel Evaluation and GPU Override.

Figure 10 Maya test script parameters

The following is a screenshot of an error message which occurs if the application does not have the proper amount of graphics memory (frame buffer) available as a resource:

Figure 11 Error message due to low memory config

Additionally, certain workflows such as those captured in the XGen Hair (Tiger) test requires OpenCL. When virtualizing the GPU, and using the GRID M60 card, the only vGPU profile which enables OpenCL is the M60-8Q Profile. If the M60-8Q profile is not assigned for this workload, Maya will crash upon opening. The following application event 1000 error was captured in the Windows Event log.

Figure 12 Event 1000 Application Error

To avoid this application crash, users should consult the Maya output window which is displayed in front of the application. This output window will indicate this issues exist with graphic card drivers and/or Open CL requirements.
Other configuration items, such as vCPU, memory and network are equally essential to provide a great Maya user experience. These configuration items will be highlighted in the upcoming test approach sections.

**Maya 3D Viewport Settings**

Maya has several 3D viewport settings that the application can run. Each of these viewport settings may perform differently, therefore all tests ran using all 3 viewport settings. These settings were applied using environment variables and are as follows:

- DirectX 11
- OpenGL
- CoreProfile

The default viewport setting is CoreProfile.

**Test Metrics**

**VM Metrics**

- Maya Application Logs
  - Total test runtime
- Windows Performance Counters
  - CPU
  - Memory
  - Disk
  - Network
  - NVIDIA GPU

The Windows NVIDIA GPU performance counters were added at NVIDIA GRID 367.17, 367.14 version release. These counters measured frame buffer (FB) used and GPU % Utilization within the VM.

**Host Metrics**

The following host metrics were captured and analyzed:

- Esxtop (Utilization by Core)
- NVIDIA-smi (GPU Usage)

**Test Limitations**

Automated scalability testing can be considered more aggressive by nature than a typical user workflow. In most cases, it is not typical that 10 users will be executing rendering requests simultaneously or even to the degree of which were replicated in multiple test iterations. As such, the test results from our automated scalability testing can be considered worst case scenario and in most cases, a host should be able to support more than the concluded amount of VMs. The degree to which higher
scalability would be achieved is dependent on the typical day to day activities of a user, such as the number of meetings and the length of lunch or breaks, multi-tasking, etc.

This type of peak performance testing leads to conservative estimates of scalability, by design. It is therefore likely that results in typical or even aggressively loaded datacenters will have higher scalability. These results indicate likely minimums for rarefied conditions, to serve as “safe” guidelines.

Methodology Description

The scalability testing was conducted in two phases.

- Single benchmark phase
- Scalability phase

Single Benchmark Phase

In this phase, virtual machines were created with a standard configuration. The test suite of four tests were run on the individual VMs to determine the amount of critical resources each test workflow typically used.

To ensure that each workflow had adequate resources (vCPU, GPU and RAM) to perform the workflows optimally, the single benchmark VMs were intentionally overprovisioned with CPU and Memory; 16 vCPU, 16GB of RAM. Each benchmark test also used the M60-8Q vGPU profile which provided access to the entire GPU. The utilization rates of these resources as well as test run times were then analyzed to determine the optimal VM configuration for each workflow. In some test cases, tests were re-executed at scale after re-configuring VMs.

vGPU Calculation

The following table indicates the GPU memory (frame buffer) utilization rates of the M60-8Q profile. This value was calculated by examining how the 8GB frame buffer was being utilized within the VM during test execution. These values, assisted in identifying the appropriate vGPU profile needed to support each workflow as follows:
Table 2 Typical test resource utilization

<table>
<thead>
<tr>
<th>Workflow</th>
<th>Test Name</th>
<th>% GPU Memory</th>
<th>GPU Memory</th>
<th>vGPU Profile</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bifröst Simulation</td>
<td>Bridge</td>
<td>11%</td>
<td>0.88 GB</td>
<td>M60-1Q</td>
<td></td>
</tr>
<tr>
<td>Modeling Texturing</td>
<td>Character</td>
<td>24%</td>
<td>1.92 GB</td>
<td>M60-2Q</td>
<td></td>
</tr>
<tr>
<td>XGen Hair Tiger</td>
<td>Tiger</td>
<td>34%</td>
<td>2.72 GB</td>
<td>M60-4Q, M60-8Q</td>
<td>OpenCL requirement; M60-8Q See below</td>
</tr>
<tr>
<td>Character Animation</td>
<td>Jungle</td>
<td>19%</td>
<td>1.52 GB</td>
<td>M60-2Q</td>
<td></td>
</tr>
</tbody>
</table>

Note: NVIDIA GRID with Tesla M60 supports both high-end graphics and CUDA/OpenCL accelerated computing.

Since the XGen functionally within the Maya application requires OpenCL, the Tiger test required the M60-8Q profile. Currently OpenCL capabilities are only enabled using the M60-8Q profile.

vCPU Calculation

It is important to note that a VM with more vCPU does not necessarily out perform a VM with less vCPU when using multi-threaded applications. In most of the tests, the 8-vCPU configuration was adequate, however since Maya is multi-threaded, test runtimes for the Bridge and Character tests decreased when the vCPU was altered; the Character VM increased performance when the vCPU count was lowered to 4. Theory of operation for this phenomenon holds that the Maya application generates threads for every vCPU it finds. This can lead to thread contention, which decreases performance. The following tables describes the required vCPU for each test:
vCPU configuration becomes increasing important during the next phase of testing; scalability testing. Allocating too much vCPU to a VM can cause a performance overhead on the host. Allocating too little vCPU can cause sluggish performance no matter how much vGPU the VM has. Aggressive vCPU allocation management is warranted.

**Single Benchmark Results**

The following tables summarizes how the vCPU and vGPU were reconfigured after analyzing the single benchmarks to support each type of workflow:

**Table 3 vCPU configurations for testing**

<table>
<thead>
<tr>
<th>Workflow</th>
<th>Test Name</th>
<th>vCPU Reconfiguration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bifröst Simulation</td>
<td>Bridge</td>
<td>10</td>
</tr>
<tr>
<td>Modeling/Tessellation</td>
<td>Character</td>
<td>4</td>
</tr>
<tr>
<td>XGen Hair</td>
<td>Tiger</td>
<td>8</td>
</tr>
<tr>
<td>Character Animation</td>
<td>Jungle</td>
<td>8</td>
</tr>
</tbody>
</table>

**Table 4 vGPU and vCPU configurations for each test**

<table>
<thead>
<tr>
<th>Workflow</th>
<th>Test Name</th>
<th>vCPU</th>
<th>vGPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bifröst Simulation</td>
<td>Bridge*</td>
<td>10</td>
<td>M60-1Q</td>
</tr>
<tr>
<td>Modeling/Tessellation</td>
<td>Character</td>
<td>4</td>
<td>M60-2Q</td>
</tr>
<tr>
<td>XGen Hair</td>
<td>Tiger</td>
<td>8</td>
<td>M60-8Q</td>
</tr>
<tr>
<td>Character Animation</td>
<td>Jungle</td>
<td>8</td>
<td>M60-2Q</td>
</tr>
</tbody>
</table>

Note: * Bifröst Simulation workflow within the Bridge test reported 11% decrease in test runtime when increasing vCPU from 8 to 10.

**Scalability Testing Phase**

The second phase of testing consisted of scalability test runs for each individual artist workflow. A series of scalability test runs were executed and were incremented; 4, 8, 10, 12, 16 VMs. Between each test run, results and metrics were analyzed to understand how the application was performing as well as how the shared host resources were being utilized. In some test cases, VMs were reconfigured to improve the VM test run time. All configuration used are highlighted in the forthcoming test results sections.
Over 500 tests were run during scalability tests. Test were initially executed using the optimal VM configuration determined in the single benchmark phase.

The vGPU profile assigned to each workflow dictates the amount of density which is achievable in regards to the scalability of concurrently running VMs on the host. For example, XGen Hair (Tiger) workflow, required the M60-8Q profile due to its OpenCL requirement, therefore the maximum number of VMs is 2 per board. Our test server has two NVIDIA Tesla M60 cards, each card has 2 GPU’s, totally 4 GPU’s. Therefore, the maximum amount of currently running VMs using the M60-8Q profile is 4. The following tables describes the maximum number of possible currently running VMs for each of the test workflows and their vGPU profile requirement:

Table 5 Maximum possible VMs for test

<table>
<thead>
<tr>
<th>Workflow</th>
<th>Test Name</th>
<th>vGPU</th>
<th>Max # of VMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>XGen Hair</td>
<td>Tiger</td>
<td>M60-8Q</td>
<td>4</td>
</tr>
<tr>
<td>Modeling/Tessellation</td>
<td>Character</td>
<td>M60-2Q</td>
<td>16</td>
</tr>
<tr>
<td>Character Animation</td>
<td>Jungle</td>
<td>M60-2Q</td>
<td>16</td>
</tr>
<tr>
<td>Bifröst simulation</td>
<td>Bridge</td>
<td>M60-1Q</td>
<td>32</td>
</tr>
</tbody>
</table>

During scalability testing, three of the four workflows (Character, Jungle, Bridge) scaled up to 16 concurrent VMs and the test results were analyzed. The Tiger tests scaled up to 4 VMs.

**Scalability Test Metrics**

The Maya application logs were the first level of analysis. Total test runtimes were post-processed for additional calculations:

- Standard Deviation across all VMs
  - This value can be used as indicator to instability. If a higher value is reported, test run times are slower or faster from one VM to another.
- Average runtime for each scalability test
  - This value is the reported value for the entire scalability test run. For example, a 4VM scalability test, reports the average test runtime of all 4 VM test runtimes.
- Percentage change in test runtime was also calculated.
  - This value was considered the final threshold value for determining the amount of VMs which the host could adequately support without severely degraded user experience.
  - Test thresholds less than 30% were considered to deliver adequate user experience. This threshold value is slightly higher than a typical physical or single user threshold, taking in consideration of the aggressive nature of the stress test. This level therefore became the threshold for our testing.
Once these outputted VM metrics were analyzed, the host metrics were analyzed to determine the test bottleneck to performance.

Scalability Test Results

The following tables describes the test results from our testing and the amount concurrent VMs adequately supported on the host:

Table 6 Concurrent users per host

<table>
<thead>
<tr>
<th>Test Name</th>
<th>Test Area</th>
<th># of VM’s</th>
<th>vCPU</th>
<th>vGPU</th>
<th>% Change Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tiger</td>
<td>XGen-Hair</td>
<td>4</td>
<td>8</td>
<td>M60-8Q</td>
<td>-7.87%</td>
</tr>
<tr>
<td>Character</td>
<td>Modeling/ Tessellation</td>
<td>16</td>
<td>4</td>
<td>M60-2Q</td>
<td>-22.58%</td>
</tr>
<tr>
<td>Jungle</td>
<td>Character Animation</td>
<td>10</td>
<td>8</td>
<td>M60-2Q</td>
<td>-20.79%</td>
</tr>
<tr>
<td>Bridge</td>
<td>Bifröst Simulation</td>
<td>6</td>
<td>10</td>
<td>M60-1Q</td>
<td>-28.21%</td>
</tr>
</tbody>
</table>

NOTE: Overall, all the four artist workflows provided by Autodesk showed that the bottleneck on the host was CPU rather than GPU.

Although the Tiger and Character workflow tests could achieve the maximum number of VMs per GPU board, the tests themselves are not GPU intensive. Therefore, the frame buffer allocated to the VM was not entirely used by the application during test execution.

As in the Tiger and Character tests, the Jungle and Bridge tests also required more CPU than GPU. As such, the test threshold value was quickly surpassed when the amount of concurrent VMs increased to 10. The following table shows the increase in test runtime as the amount of concurrent VMs increased:
### Table 7 Bridge test threshold results

<table>
<thead>
<tr>
<th>Bridge Avg. Test Time- 1Q vGPU Profile - 10vCPU</th>
<th>% Change Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>1VM</td>
<td>6VM</td>
</tr>
<tr>
<td>CoreProfile</td>
<td>02:35.4</td>
</tr>
<tr>
<td>OpenGL</td>
<td>02:35.2</td>
</tr>
<tr>
<td>DX11</td>
<td>02:35.6</td>
</tr>
</tbody>
</table>

As discovered in the single VM test, the Bridge test runtimes were most optional using 10vCPU. However, if the amount of concurrent VMs were increased to 32 to reach to GPU maximum capacity of the GRID M60-1Q profile, the host CPU would be over allocated (320 vCPU) since the host has 64 logical cores with hyper threading enabled. Powering on more VMs meant using more vCPU than the host had as a shared resource.

### CONCLUSION

The test results show that XGen modeling workloads, or any workload that requires OpenCL, the maximum possible number of users per server is four. Testing shows that with 4 concurrent M60-8Q profiles running simultaneously, all users see less than a 10% degradation in performance, well below discomfort level. We **conclude that 4 concurrent users per server is the maximum**. The Character modeling test can **support 16 VMs/Users per server**, which is the maximum number of VM’s for the M60-2Q profile. Both the Tiger and Character workflow tests achieve the maximum number of VMs per GPU board, but the tests themselves are not GPU intensive. The frame buffer allocated to the VM was not entirely used by the application during test execution.

The Jungle and Bridge tests exceeded acceptable usability levels when reaching 10 VMs, and the Bridge test was found to perform better when more vCPU was added to the VM. Both tests put a larger demand on the CPU than GPU. In the case of the Bridge test, reconfiguring vCPU to 10 increased the performance of the VM. However, as such, the host CPU would become too overallocated if the host had more than 10 VM’s running. On this class of server, the maximums VM’s which are concurrently executing for Bifröst simulation or the Character animation work modelled within the test **reside between 6 and 8 users per server**.
ADDITIONAL TESTING: MAYA 2017 SPEC

Maya 2017 SPEC contains eight artist workflows. The 4 workflows referenced in this whitepaper are also included with SPEC. The following is a list of the tests included in SPEC:

- Bifröst Bridge
- Jungle Animation
- Seven Space Animation
- Space Crash
- Seven Character Animation
- XGen Hair Tiger
- Character Model Texture
- Toy store

This SPEC test provides individual elapsed time test results for the different modes tested such as animation, Bifröst, smooth shade (SSAO, edges, texture, all lights), and wire. These elapsed times were compared to a physical NVIDIA workstation graphic card and the test results were within range. No crashes or errors were reported.

The Tiger workflow was modified for SPEC; the same Tiger model is loaded but exercises the GPU more heavily by rotating the model on the screen. This SPEC Tiger workflow intends to model real Maya users who build models as they are animating on the screen and requires both powerful GPU and CPU. Since the Tiger XGen capabilities within Maya requires OpenCL, the SPEC benchmark was ran using the same GRID M60-8Q vGPU (8GB) profile as used in our original test.
The following graphs illustrated GPU and CPU Utilization rates on the VM during a single benchmark test. This graph captures the entire SPEC benchmark (total of 8 tests).

Figure 13 GPU and CPU utilization rates
Notice

The information provided in this specification is believed to be accurate and reliable as of the date provided. However, NVIDIA Corporation (“NVIDIA”) does not give any representations or warranties, expressed or implied, as to the accuracy or completeness of such information. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This publication supersedes and replaces all other specifications for the product that may have been previously supplied.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and other changes to this specification, at any time and/or to discontinue any product or service without notice. Customer should obtain the latest relevant specification before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer. NVIDIA hereby expressly objects to applying any customer general terms and conditions with regard to the purchase of the NVIDIA product referenced in this specification.

NVIDIA products are not designed, authorized or warranted to be suitable for use in medical, military, aircraft, space or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer’s own risk.

NVIDIA makes no representation or warranty that products based on these specifications will be suitable for any specified use without further testing or modification. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer’s sole responsibility to ensure the product is suitable and fit for the application planned by customer and to do the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer’s product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this specification. NVIDIA does not accept any liability related to any default, damage, costs or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this specification, or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this specification. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA. Reproduction of information in this specification is permissible only if reproduction is approved by NVIDIA in writing, is reproduced without alteration, and is accompanied by all associated conditions, limitations, and notices.

ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, “MATERIALS”) ARE BEING PROVIDED “AS IS.” NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA’s aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the NVIDIA terms and conditions of sale for the product.

HDMI

HDMI, the HDMI logo, and High-Definition Multimedia Interface are trademarks or registered trademarks of HDMI Licensing LLC.

OpenCL

OpenCL is a trademark of Apple Inc. used under license to the Khronos Group Inc.

Trademarks

NVIDIA and the NVIDIA logo are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright

© 2016 NVIDIA Corporation. All rights reserved.

www.nvidia.com