NVIDIA GRID vPC Sizing Guide

Application Sizing Guide

SP-10103-001_v01 | August 2020

Document History

SP-10103-001_v01

Version	Date	Authors	Description of Change
01	August 13, 2020	AS, EA, SM	Initial Release

Table of Contents

List of Figures

List of Tables

Chapter 1. Executive Summary

This specification provides insights into how to leverage NVIDIA GRID® virtual PC (vPC) for digital knowledge workers. It provides recommendations based on NVIDIA's nVector knowledge worker benchmarking and cover common questions such as the following:

- Which NVIDIA® GPU should I use for my business needs?
- How do I select the right NVIDIA virtual GPU (vGPU) profile(s) for the types of users I will have?
- What are the advantages of running NVIDIA GRID vPC versus traditional CPU only virtual desktop infrastructure (VDI)?

Digital worker (a.k.a. knowledge worker) workloads will vary per user depending on many factors, including number of applications, the types of applications, file sizes and number of monitors and their resolution. This specification used the NVIDIA nVector as the testing framework for executing a typical knowledge worker workload which simulates application workflow, as well as a tool for capturing real world metrics. Since the number of monitors and their resolution have a direct impact on sizing, our testing to support this specification explored various screen resolutions and number of monitors. Tests were executed on CPU only VM's as well VM's backed by NVIDIA vGPU.

It is recommended that you test your unique workloads to determine the best NVIDIA virtual GPU solution to meet your needs. The most successful customer deployments start with a proof of concept (POC) and are "tuned" throughout the lifecycle of the deployment. Beginning with a POC enables customers to understand the expectations and behavior of their users and optimize their deployment for the best user density while maintaining required performance levels. Continued maintenance is important because user behavior can change over the course of a project and as the role of an individual changes in the organization. A user that was once a light graphics user might become a heavy graphics user when they change teams or are assigned a different project. Management and monitoring tools enable administrators and IT staff to ensure their deployment is optimized for each user.

1.1 About NVIDIA nVector Benchmark

NVIDIA's performance engineering team developed a methodology and benchmarking tool which simulates, at scale, a digital knowledge worker workflow. This workflow is a good representation of knowledge workers commonly used software applications:

- Microsoft Word 2016
- Microsoft Excel 2016
- Microsoft PowerPoint 2016
- Google Chrome web browser and video streaming
- PDF document viewing

These applications will perform various functions throughout the test that replicate a task that a real end user would perform. Microsoft Word, Excel, and PowerPoint creates new content, modify existing content, and move content between applications. Task within these applications include scrolling, zooming, menu navigation, and PDF creation. Google Chrome streams live video and visits interactive websites. Microsoft Edge acts as a PDF viewer.

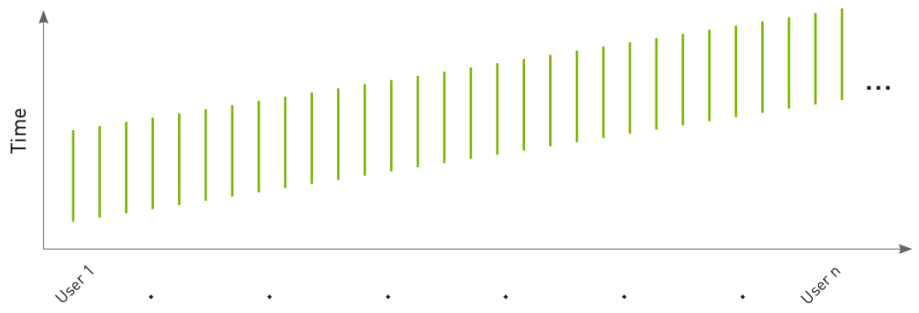
When running the nVector benchmark at scale, nVector randomizes Knowledge Worker (KW) workloads across multiple virtual machines.

The following table shows the workflow of each user and the graph in Figure 1-1 shows cumulative increase in the number of users running workloads through time. Multiple users are tested at a time to simulate scale, with start and end times staggered to be more representative of real VDI environments.

Table 1-1. Simulating Many Users and Behaviors

User 1	User2	User 3	User 4
Google Chrome (Video)	MS Word 2016	Windows Media Player	Google Chrome (Web)
Windows Media Player	Microsoft Edge (PDF)	MS Word 2016	Google Chrome (Video)
MS Word 2016	MS Excel 2016	Microsoft Edge (PDF)	Windows Media Player
Microsoft Edge (PDF)	Google Chrome (Web)	MS Excel 2016	MS Word (2016)
MS Excel 2016	Google Chrome (Video)	Google Chrome (Web)	Microsoft Edge (PDF)

Figure 1-1. Characteristics of NVIDIA's Benchmarking Tool



1.2 What is NVIDIA GRID vPC?

NVIDIA GRID virtual PC software enables the delivery of graphics-rich virtual desktops accelerated by NVIDIA graphics processing units (GPUs). NVIDIA GRID vPC enables sharing the same GPU across multiple virtual machines, delivering a native-PC experience to knowledge workers while improving user density. Because tasks typically done on the CPU are offloaded to the GPU, the user has a much better experience and more users can be supported.

Virtual GPU profiles determine the amount of frame buffer that can be allocated to your virtual machine. The vGPU profiles that are supported on NVIDIA GPU's with NVIDIA GRID software, are the 1B (with 1024 MB of frame buffer) and 2B (with 2048 MB of frame buffer).

Table 1-2. NVIDIA Virtual PC Feature List

Configuration and Deployment	**NVIDIA GRID vPC**
Desktop Virtualization	✓
Remote Desktop Session Host (RDSH) App Hosting	✓
RDSH Desktop Hosting	✓
Windows OS Support	✓
Linux OS Support	✓
GPU Pass-through Support	✓
Bare Metal Support	✓
NVIDIA Graphics Driver	✓
Guaranteed Quality-of-Service Scheduling	✓
Display	**NVIDIA GRID vGPU**
Maximum Hardware Rendered Display	Four HD, Two 4K[1], One 5K[2]
Maximum Resolution	5120 × 2880[2]

Notes:

[1]Support starts with the NVIDIA virtual GPU software March 2018 release (Version 6.0).

[2]5K resolution support starts with the NVIDIA virtual GPU software December 2019 release (10.0).

NVIDIA GRID vPC delivers an engaging user experience for the digital workplace. Employees can be most productive using modern applications and work the way they want, from anywhere. Delivering up to 30% improved density and 3X improved latency over CPU only VDI, IT can cost-effectively scale virtualization to every employee with performance that rivals a physical PC.

1.3 Recommended NVIDIA GPUs for NVIDIA GRID vPC

Density optimized GPUs are typically recommended for knowledge worker virtual desktop infrastructure to run office productivity applications, streaming video, and Windows 10. They are designed to maximize the number of VDI users supported in a server.

Table 1-3. Recommended GPUs for Density and Performance

	NVIDIA T4	NVIDIA M10
Number of cards (Architecture)	1 (NVIDIA Turing™)	1 (NVIDIA Maxwell™)
NVIDIA® CUDA® cores	2560	640 per GPU 2560
Tensor Cores	320	-
NVIDIA® TensorRT™ cores	40	-
Memory size	16 GB GDDR6	8 GB per GPU 32 GB GDDR5
Form factor	PCIe 3.0 single-slot	PCIe 3.0 dual-slot
Power	70 W	225 W
Thermal	Passive	Passive
Optimized for	Density and performance	Density

The NVIDIA T4 GPU is based on the NVIDIA Turing™ architecture. The NVIDIA M10 GPU offers the best user density and performance option for NVIDIA GRID vPC customers. The M10 is a 32 GB dual-slot card which draws up to 225 W of power. Therefore, it requires a supplemental power connector. The T4 is a low profile, 16 GB single-slot card, which draws 70 W maximum and does not require a supplemental power connector. Two NVIDIA T4 GPUs provide 32 GB of frame buffer and support the same user density as a single M10 with 32 GB of frame buffer, but with lower power consumption and more performance. While the M10 provides the best value for knowledge worker deployments, selecting the T4 for this use case brings the unique benefits of the NVIDIA Turing architecture. This enables IT to maximize data center resources by running virtual desktops in addition to virtual workstations, deep learning inferencing, rendering, and other graphics and compute intensive workloads -- all leveraging the same data center infrastructure. This ability to run mixed workloads can increase user productivity, maximize utilization, and reduce costs in the data center. Additional T4 technology enhancements include support for VP9 decoding, which is often used for video playback, and H.265 (HEVC).

The NVIDIA T4 leverages ECC memory and is enabled by default. When enabled, ECC has a 1/15 overhead cost due to the need to use extra VRAM to store the ECC bits themselves. This

will result in a lower frame buffer on the vGPU compared to the physical GPU. It is important to resize your environment when switching from NVIDIA Maxwell GPUs to newer GPUs like Pascal and NVIDIA Turning GPUs. Additional information can be found here.

The maximum number of vGPUs that can be created simultaneously on a physical GPU is defined by the amount of frame buffer per VM, and thus how many VMs can share that physical GPU. For example, an NVIDIA GPU which has 16 GB of GPU memory, can support up to 16 1B profiles (16 GB total with 1GB per VM). You cannot oversubscribe GPU memory and it must be shared equally for each physical GPU.

The NVIDIA T4 and M10 are the recommended GPU's for NVIDIA GRID vPC since these GPUs are optimized for density. The complete list of NVIDIA GPUs that support NVIDIA GRID vPC can be found here.

Chapter 2. Testing Methodology

2.1 Single VM Testing

The first phase of testing explored the impact of higher resolution and multi-monitor scenarios, and the following tests were executed:

Table 2-1. Single VM Testing

Resolution	Monitors
High Definition (HD) 1920 × 1080	2
Quad High Definition (QHD) 2560 × 1440	2
	3
	4
4K (4096 × 2160)	1
	2
5K (5120 × 2880)	1

Note: This table reflects the configurations that were not tested. These configurations are not our recommendations.

Tests were executed on a single VM using 1B and 2B vGPU profiles to determine the most optimal vGPU profile based upon the nVector KW workload. Tests were executed on CPU only as well as vGPU enabled environments.

2.1.1 Test Environment

Phase one testing leveraged two physical servers with one hosting the target NVIDIA GRID vPC VMs and the second hosting the virtual clients. Both server hosts used VMware vSphere ESXi 6.7.0 and NVIDIA GRID 10.1. The target VM acts as a standard GRID vPC VDI that an end user would connect to and the virtual client acts as an example of an endpoint that the end user would use to connect to the target VM.

Table 2-2. Testing Environment

Host Configuration	VM Configuration	Virtual Client
PowerEdge R740 Rack Server	vCPU: 4	vCPU: 4
Intel® Xeon® Gold 6148 @ 2.40 GHz	vRAM: 6144 MB	vRAM: 4096 MB
VMware ESXi, 6.7.0, 15160138	NIC: 1 (vmxnet3)	NIC: 1 (vmxnet3)
Number of CPUs: 40 (2 × 20)	Hard disk: 40 GB	Hard disk: 40 GB
Memory: 512 GB	Virtual Hardware: vmx-13	Virtual Hardware: vmx-13
Storage: Local Flash	VMware Horizon 7.9	VMware Horizon 7.9
Power Setting: High Performance	Blast Extreme (4:4:4)	Blast Extreme (4:4:4)
GPU: 6 × T4	vGPU Driver: GRID 10.1 (Windows Driver 442.06)	vGPU Driver: GRID 10.1 (Windows Driver 442.06)
Scheduling Policy: 0x00 (Best Effort)	Guest OS: Windows 10 Enterprise 1903	Guest OS: Windows 10 Enterprise 1903

2.1.2 Test Metrics – Frame buffer Usage

Frame buffer utilization is based upon many factors including application load, monitor configuration, and screen resolution. Since our test focuses on the impact of higher resolutions and multi-monitor scenarios, frame buffer utilization is a critical test metric.

2.1.2.1 GPU Profiler

GPU Profiler is commonly used tool which can quickly capture resource utilization while a workload is being executed on a virtual machine. This tool is typically used during a POC to help size the virtual environment to ensure acceptable user performance. GPU Profiler was running on a single VM with various vGPU profiles while the nVector knowledge worker workload was running. The following metrics were captured:

- Frame buffer %
- vCPU %
- RAM %
- Video Encode
- Video Decode

Figure 2-1. GPU Profiler

A good rule of thumb to follow is that frame buffer utilization should not exceed 90% for a short time or an average over 70% on the 1 GB (1B) profile. If high utilization is noted, then the vPC VM should be assigned a 2 GB (2B) profile. These results are reflective of the work profile mentioned in Section 1.1. Due to users using application in different ways we recommend performing your own POC with your workload.

2.2 Scalability Testing

Typical VDI deployments have two conflicting goals: Achieving the best possible user experience and maximizing user density on server hardware. Problems can arise as density is scaled up, because after a certain point it can negatively impact user experience. Phase two testing used nVector to execute tests at scale on 64 VM's and leveraged 2 – High Definition (1920 × 1080) monitors. Capacity planning for the server is often dependent upon server resource utilization metrics as well as user experience, this phase of testing examined both and the following sections summarize their importance and how to analyze these metrics.

2.2.1 Server Utilization Metrics

Observing overall server utilization will allow you to assess the trade-offs between end user experience and resource utilization. To do this, monitoring tools periodically samples CPU core and GPU utilization during a single workload session. To determine the 'steady state' portion of the workload, samples are filtered, leaving out the times when users have all logged on and the workload start ramps up and down. Once steady state has been established, all samples are aggregated to get the total CPU core utilization on the server.

The utilization of the GPU compute engine, the frame buffer, the encoder, and decoder can all be monitored and logged through NVIDIA System Management Interface (nvidia-smi), a command line interface tool. In addition, NVIDIA vGPU metrics are integrated through management packs like VMware vRealize Operations. For our testing purposes, nVector automated the capture of the following server metrics. Since it is highly recommended that you test your unique workloads during a POC, nvidia-smi commands can run on the hypervisor, which will allow you to monitor GPU utilization of the physical GPU. Refer to Chapter 4 for further syntax information.

2.2.2 User Experience Metrics

NVIDIA's nVector benchmarking tool has built-in mechanisms to measure user experience. The next section, "Latency Metrics" will dig deeper into how the end user experience is measured and how results are obtained.

2.2.2.1 Latency Metrics

Latency defines the response or feel of the end user when working with applications in the VDI. Increased latency can provide a poor experience that can include mouse cursor delay, text display issues when typing, and audio/video sync issues. The lower latency the better.

Imagine that you are working on PowerPoint and adding a shape and resizing it. On the first attempt, this process is instantaneous. However, the second attempt is delayed by several seconds or is sluggish. With such inconsistency, the user tends to overshoot or have trouble getting the mouse in the right position. This lack of a consistent experience can be very frustrating. Often, it results in the user experiencing high error rates as they click too fast or too slow, trying to pace themselves with an unpredictable response time. NVIDIA's nVector benchmarking tool measures the variation in end user latency and how frequently it is experienced.

2.2.2.2 Remoted Frames Metrics

Frame rate metrics are captured on the endpoint and provides a great metric on the possible end user experience. The average frame rate is captured and calculated across the simulated workload. A lower frame rate can cause slow response during screen fresh and stuttering during scrolling or zooming. The higher frame rate the better.

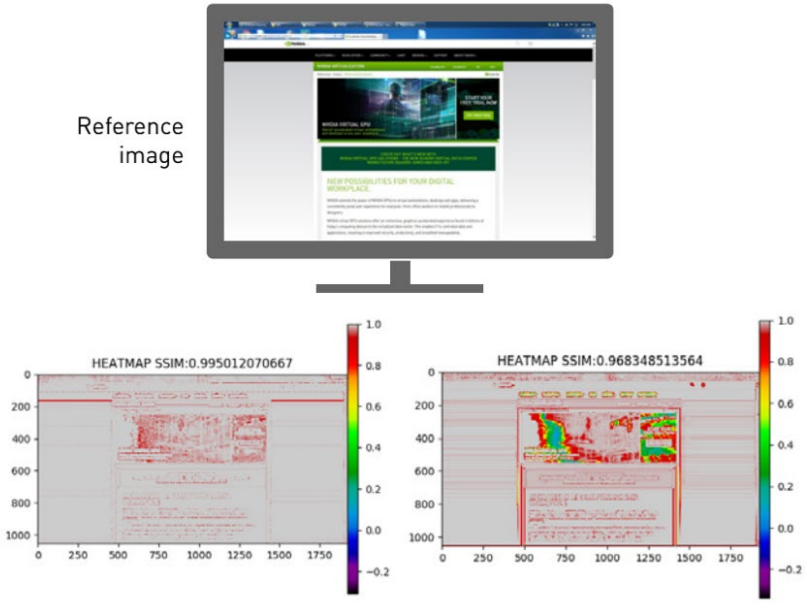
Remoted frames are a common measure of user experience. For the entire duration of the workload, NVIDIA's nVector benchmarking tool collects data on the 'frames per second' provided by the remote protocol vendor. The tool then tallies the data for all VDI sessions to get the total number of frames remoted for all users. Hypervisor vendors likewise measure total remoted frames as an in indicator of quality of user experience. The greater this number, the more fluid the user experience.

2.2.2.3 Image Quality

Image quality is determined by the remoting protocol, configuration of VDI environment, and endpoint capability of the end point. For this sizing guide the protocol used is VMware Blast Extreme with High Color Accuracy (HCA) YUV444. HCA no longer removes any chroma information from images and provides a much better image quality.

NVIDIA's nVector benchmarking tool uses a lightweight agent on the VDI desktop and the client to measure image quality. These agents take multiple screen captures on the VDI desktop and on the thin client to compare later. The structural similarity (SSIM) of the screen capture taken on the client is computed by comparing it to the one taken on the VDI desktop. When the two images are similar, the heatmap will reflect more colors above the spectrum shown on its right with an SSIM value closer to 1.0 (Figure 2-2). As the images become less similar, the heatmap will reflect more colors down the spectrum with a value less than 1.0. More than a hundred pairs of images across an entire set of user sessions is obtained. The average SSIM index of all pairs of images is computed to provide the overall remote session quality for the entire population of all users. The threshold SSIM value is 0.98, scores which are above 0.98 indicate good image quality with 1.0 being perfect.

Figure 2-2. SSIM as Measure of Image Density



Same remoting protocol with two different settings

Chapter 3. Test Findings

3.1 Single VM Multi-Monitor Resolution Test Results

The following table summarizes the results of Phase 1 testing where we explored the impact on frame buffer (FB) for higher resolution and multi-monitor scenarios based upon the nVector KW workload. As monitor resolutions continue to increase, more pixels are being delivered to the screen. As a result, the frame buffer usage in a virtual environment increases. While HD (1920 × 1080) is currently the most common resolution, an increasing number of devices are being released with higher resolution screens.

Table 3-1. nVector Knowledge Worker Workload Test Results

Resolution	Monitors	vGPU Profile
High Definition (1920 × 1080)	2	1B
Quad High Definition (2560 × 1440)	2	2B
	3	2B
	4	2B
4K (4096 × 2160)	1	2B
	2	2B
5K (5120 × 2880)	1	2B

Note: Based upon benchmark testing, test your own workloads to ensure FB sizing is appropriate for your users.

Knowledge worker workloads will vary per user depending on many factors, including number of applications, the types of applications, file sizes and number of monitors and their resolution. Additional monitor and resolution support, including mixed displays, can be found here. It is highly recommended that you test your own workloads during a POC since mileage may vary. Our nVector test results should be used for guidance purposes only.

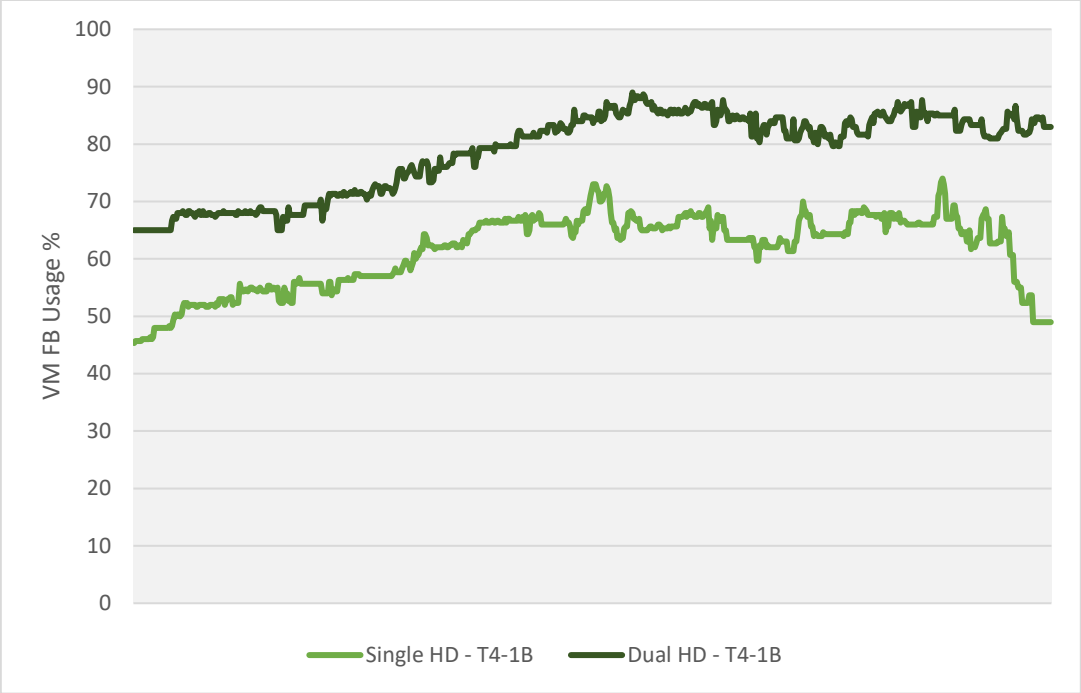
The results of Phase 1 frame buffer analysis are used for sizing purposes and since the maximum number of vGPUs which can then be created (and then assigned to a VM) is defined

by the amount of GPU memory per VM. The following sections describe the frame buffer usage captured on the VM for nVector KW workload.

3.1.1 High Definition (1920 × 1080) Displays

When the number of monitors were increased, more pixels are being delivered to the screen. Our nVector KW workload reported an average of 15% increase of FB usage when monitors were increased from 1 to 2. Our nVector test results illustrate the vPC 1B profile size was able to support 2 high definition monitors when executing the nVector KW workload. The following figure illustrate the frame buffer usage while KW workload executed using dual high definition monitors:

Figure 3-1. Frame Buffer Usage with Dual High Definition Monitors



3.1.2 Quad High Definition (2560 × 1440) Displays

Quad High Definition (2560 × 1440) resolution tests were executed using 2, 3, and 4 monitors. Quad high definition has almost double as many pixels as HD, therefore FB requirements for Quad High Definition monitors are greater than HD. Overall, our nVector test results for KW workload illustrates that 2B profile was sufficient for 2 – Quad High Definition monitors. When monitors were increased from 2 to 3 there was a 15% increase of FB usage. With this in mind, the 2B profile provided the most adequate amount of FB for 3 and 4 Quad High Definition monitors. The following sections describes our test findings.

3.1.2.1 Dual QHD Monitor Test Results

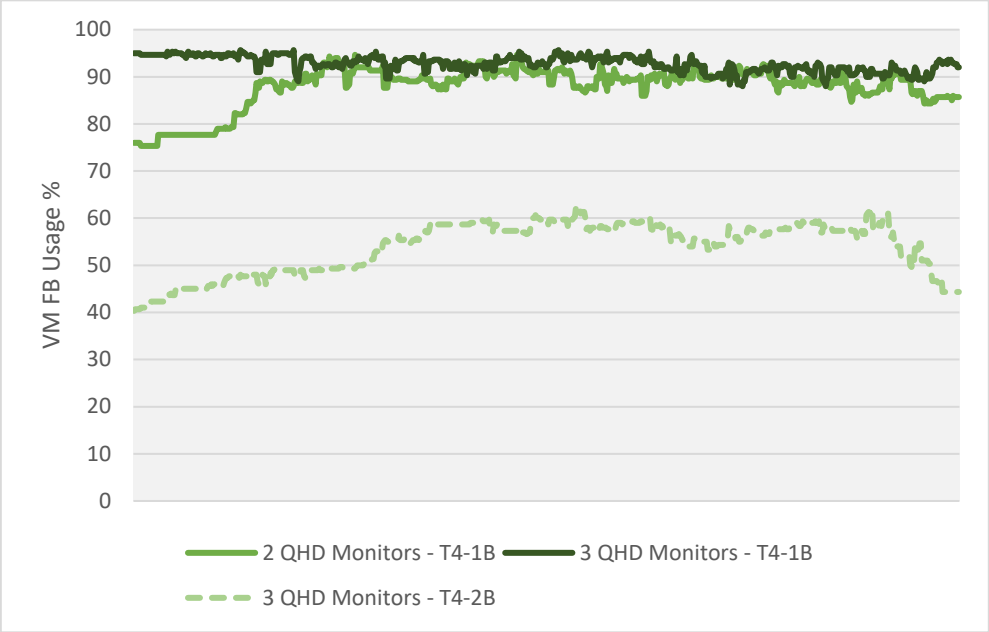
The following figure illustrates the frame buffer usage captured while executing the nVector KW workload on dual quad high definition monitors.

Figure 3-2. Frame Buffer Usage with Dual Quad High Definition Monitors

3.1.2.2 Triple QHD Monitor Test Results

The following figure illustrates impact on frame buffer when monitors are increased from 2 to 3 monitors in separate nVector KW workload tests. The FB usage of the 1B profile increased past 90% frequently throughout the test, indicating that user experience may be impacted. With this in mind, the 2B profile would be the correct vGPU profile for our nVector KW workflow.

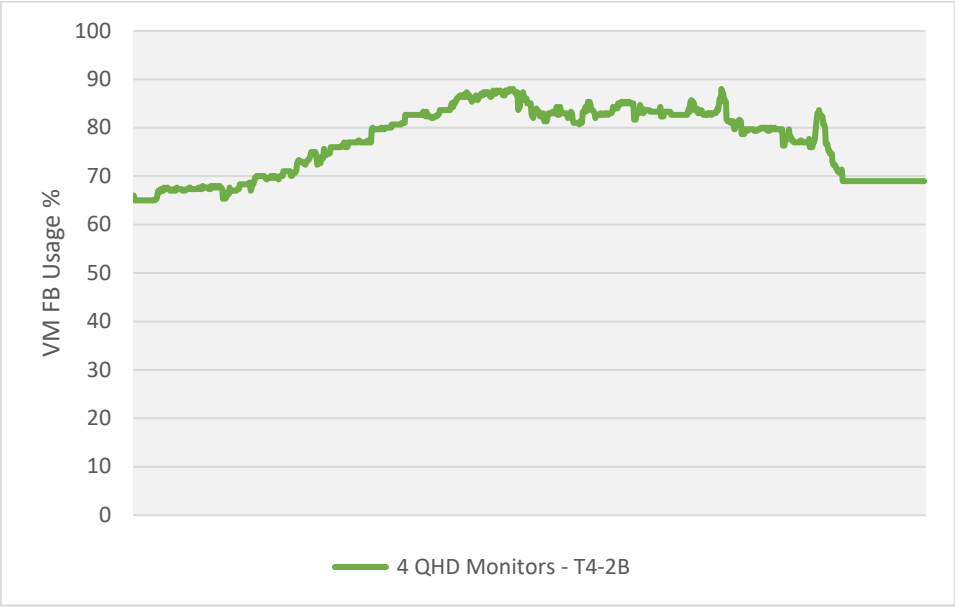
Figure 3-3. Frame Buffer Usage with Triple Quad High Definition Monitors



3.1.2.3 Quad QHD Monitor Test Results

Using the knowledge gained from our testing of 3 monitors, the 4-monitor testing used the 2B profile. The following figure illustrates FB usage when we increased the vGPU profile to 2B and ran 4 monitors:

Figure 3-4. 2B vGPU Profile with 4 Monitors Frame Buffer Usage



3.1.3 4K (4096 × 2160) Displays

Tests were executed using a single 4K monitor as well as dual 4K monitor. The test results using nVector KW workload illustrated that a 1B vGPU profile for a single 4K monitor was sufficient. However, it certainly utilized the FB and in certain cases a 2B profile may be justified. Based upon this information, when the number of monitors were increased to 2 - 4K monitors, the 2B profile was best suited for the nVector KW workload.

3.1.3.1 Single 4K Monitor Test Results

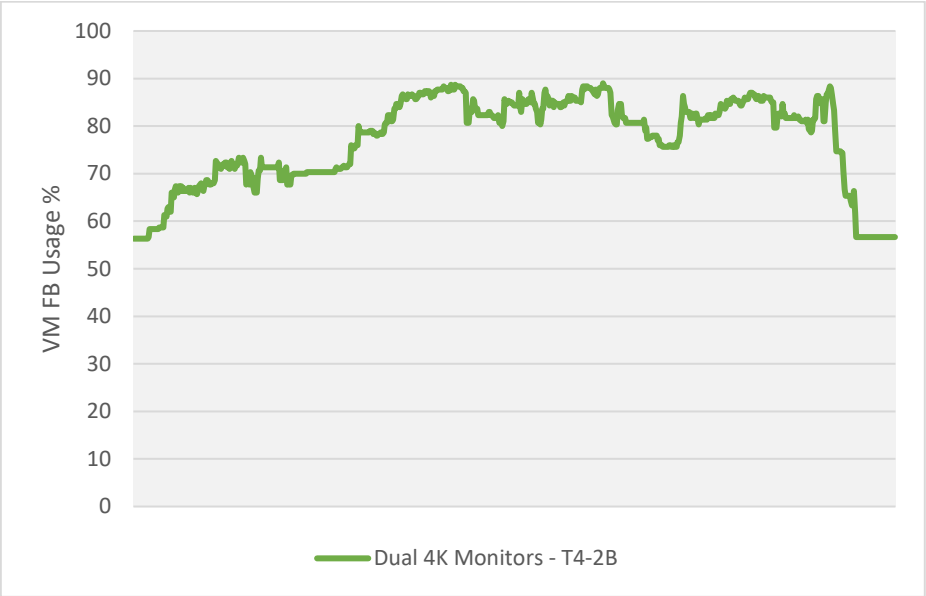
The following figure illustrates the FB usage of the nVector KW workload using the 2B profile.

Figure 3-5. nVector KW Workload using 2B Profile Frame Buffer Usage

3.1.3.2 Dual 4K Monitor Test Results

Using the knowledge gained from our single 4K monitor test, the dual 4K monitor testing used the 2B profile. The following figure illustrates FB usage when we increased the vGPU profile to 2B and used 2 monitors while executed the nVector KW workload.

Figure 3-6. Increased vGPU Profile to 2B with two Monitors Frame Buffer Usage



3.1.4 5K (5120 × 2880) Display

NVIDIA GRID vPC supports only a single 5K monitor and for our nVector KW workload, the 2B vGPU profile was chosen based upon the known FB requirements of 5K resolution. 5K displays have a resolution of about 7x's the number of pixels than high definition displays (1920 × 1080). The following figure illustrates FB Usage for the nVector KW workload.

Figure 3-7. Frame Buffer Usage for nVector KW Workload

3.2 Phase 2: Multi-Monitor Resolution Scalability Test Results

Running a single VM on a large environment does not allow you to capture the usage of a production environment. Since full HD is currently the most common resolution, the results of our scalability testing within this specification focuses on dual high definition (1920 × 1080) monitors. Phase 2 testing used the results of Phase 1 testing, where it was concluded that for the nVector KW workload, the 1B profile was sufficient for dual high definition monitors. It is highly recommended that you test your own workloads during a POC since mileage may vary. Our nVector test results should be used for guidance purposes only.

For the purpose of our testing, the scale was configured for 64 VMs on the ESXi host. For further insight regarding the increased VDI density when comparing CPU-only to NVIDIA vPC, we also ran 32 VM's scalability tests.

The following table summarizes the multi-monitor high resolution test environment as well as how many NVIDIA GPU's were used for each scalability test.

Table 3-2. Multi-Monitor High Resolution Test Environment

# of VMs at Scale	# of NVIDIA GRID Cards	vGPU Profile	Monitor Resolution	# of Monitors
64	2	M10-1B	1920 × 1080	2
32	1	M10-1B	1920 × 1080	2

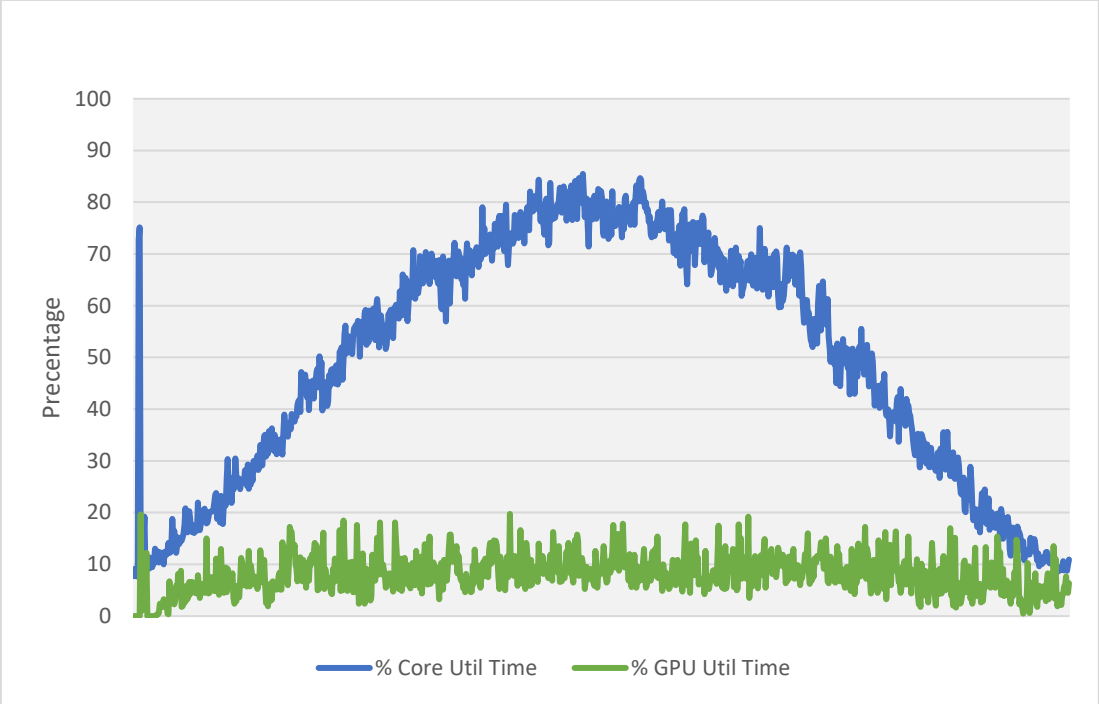
During this process, the benchmark was used to execute various nVector KW workflows across multiple VMs with start and end times staggered across the environment. Tests were executed on CPU only VMs as well as vGPU VM's. The following section illustrates CPU only versus vGPU Server Metrics as well as user experience test results for the nVector KW workflow.

3.2.1 Server Utilization Metrics

Choosing the correct CPU for virtualization and proper configuration can have a direct effect on scalability even when a virtual GPU is present. Processor resources are often hyper-threaded and overprovisioned to a certain degree. In terms of CPU specs, you should evaluate the number of cores and clock speed. For our testing, we choose to test on two different CPU configurations and by doing so, our test results illustrate that scalability for vPC is often dependent what CPU configuration you choose for your deployment. The following section describe our test findings when using 64 VMs:

The following figure illustrates CPU Core utilization using Intel Xeon Gold 6154 CPU at 3.0 GHz for 64 VMs. This server configuration has 36 cores with hyperthreading enabled.

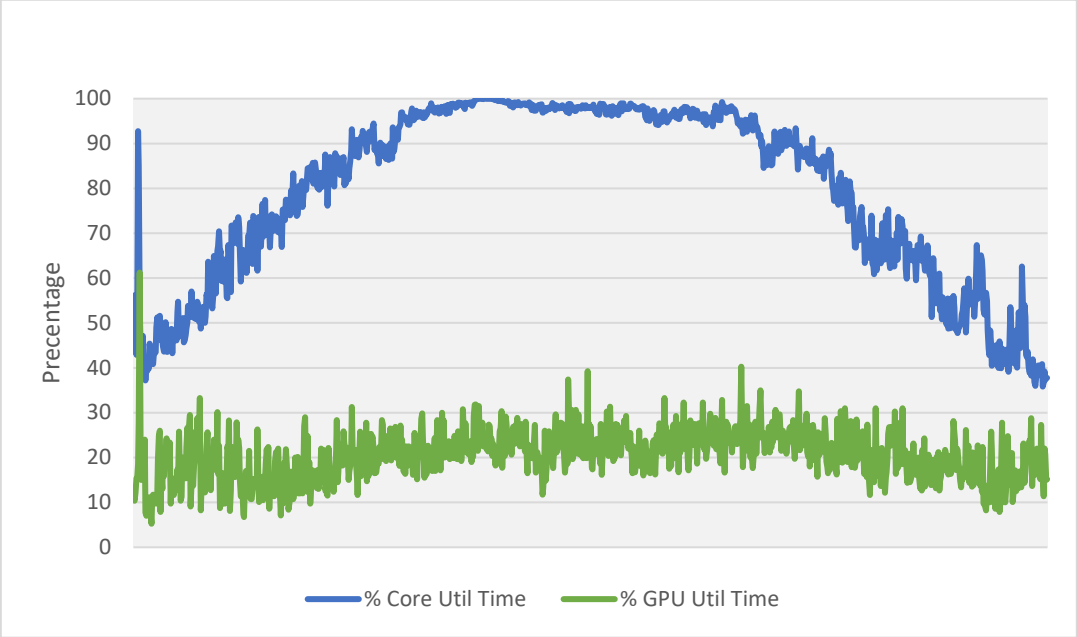
Figure 3-8. Intel Xeon Gold 6154 Utilization



It is also important to note, GPU utilization rates illustrated in Figure 3-8 indicates there is not a GPU bottleneck. Meaning, the server has plenty of head room within the GPU compute engine. GPU Util time is being reported by averaging utilization across the two M10 GPUs in the server.

The following figure illustrates CPU Core utilization using Intel Xeon Gold 6140 CPU at 2.3 GHz for 64 VM's. This server configuration has 36 cores with hyperthreading enabled. During the peak test execution, CPU Core Util Time reached 100% therefore the server is doing more than what it has capacity for. Since this server has the same GPU as the 6154 CPU Server, as expected, GPU utilization shows there is not a GPU bottleneck.

Figure 3-9. Intel Xeon Gold 6140 Utilization

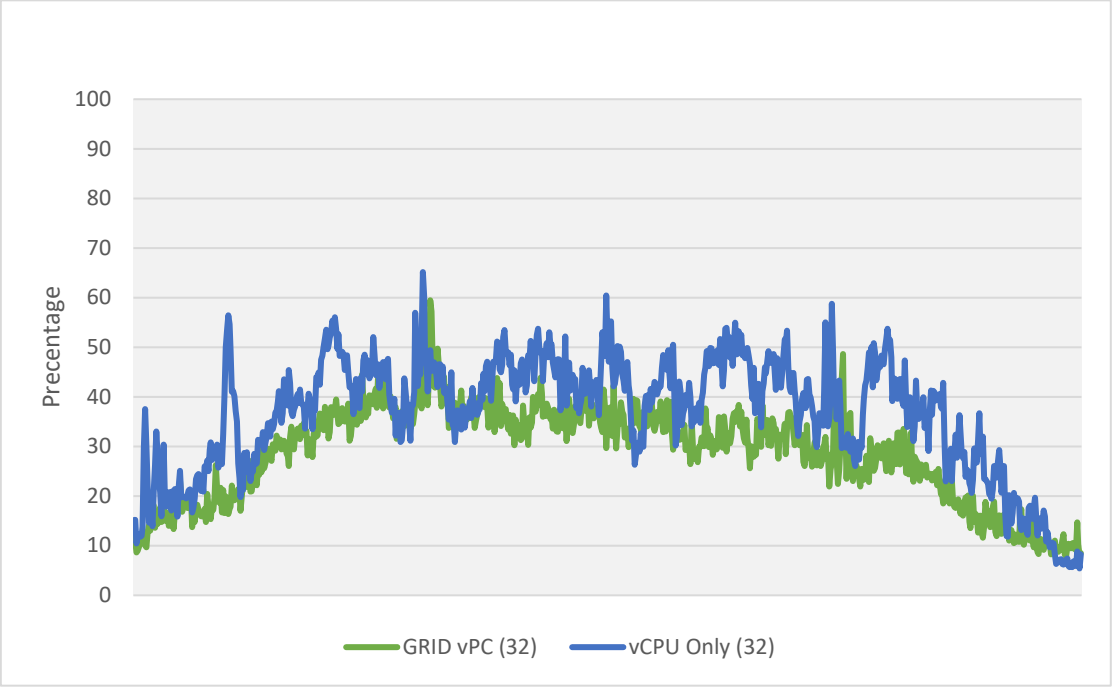


3.2.1.1 Reduced CPU and increased VDI density

When introducing a NVIDIA vGPU into the environment, virtual systems will no longer rely solely on CPU for graphics processing. Systems without a GPU have higher overall CPU usage due to the additional GPU requirements in Windows 10 and other applications. With this in mind, the 64 VM vPC test results illustrated that if a bottleneck was to occur in the system, it would be CPU not GPU. Therefore, we ran a 32 VM test to illustrate increased VDI density when comparing vPC to a CPU-only VDI environment.

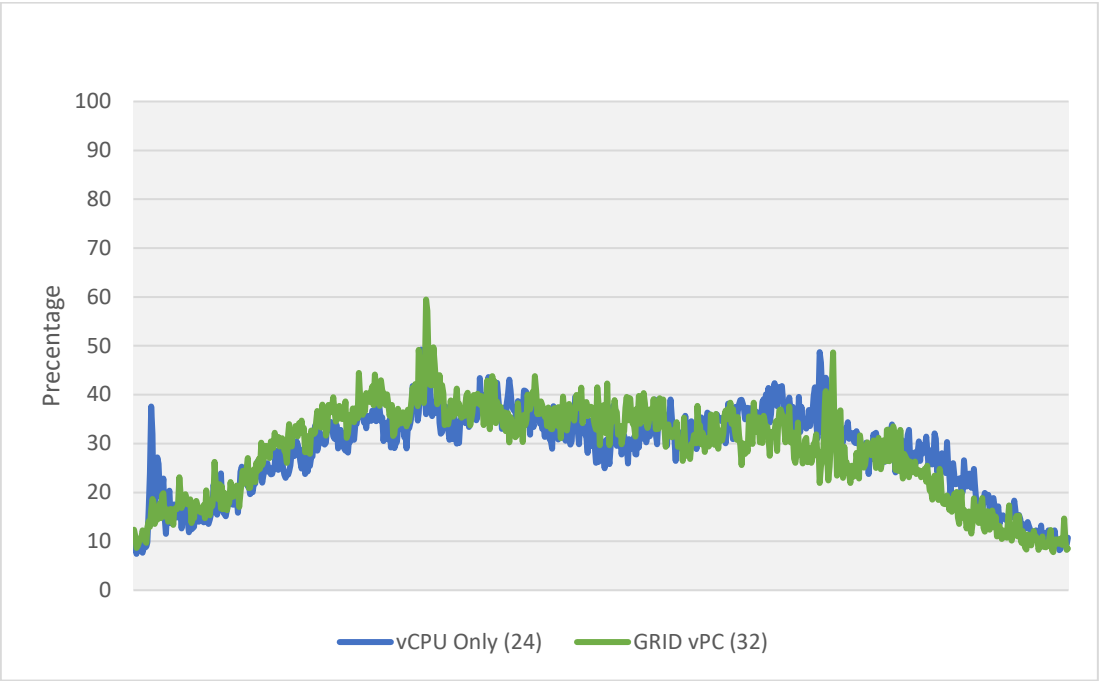
The chart in Figure 3-10 compares CPU Core Utilization while executing the nVector KW workload for CPU-only VM's and NVIDIA GRID vPC. There is a consistent gap in server CPU utilization when comparing the two, which averages 15%. This results in a net decrease in CPU usage while applications and users are active.

Figure 3-10. CPU Core Utilization on ESXi Host



To further illustrate the importance of NVIDIA GRID vPC and value of reducing CPU usage, we ran a 24 VM CPU-only test using the nVector KW workload. The following figure illustrates the CPU Core Utilization rates for both CPU-only and NVIDIA GRID vPC VMs.

Figure 3-11. CPU Core Utilization on ESXi Host



When the number of CPU-only VMs were lowered to 24, the CPU utilization rate dropped to the level of 32 GPU-enabled VMs .This means you can support up to 33% more users on GPU-enabled VDI environment with a better user experience.

3.2.2 nVector User Experience Metrics

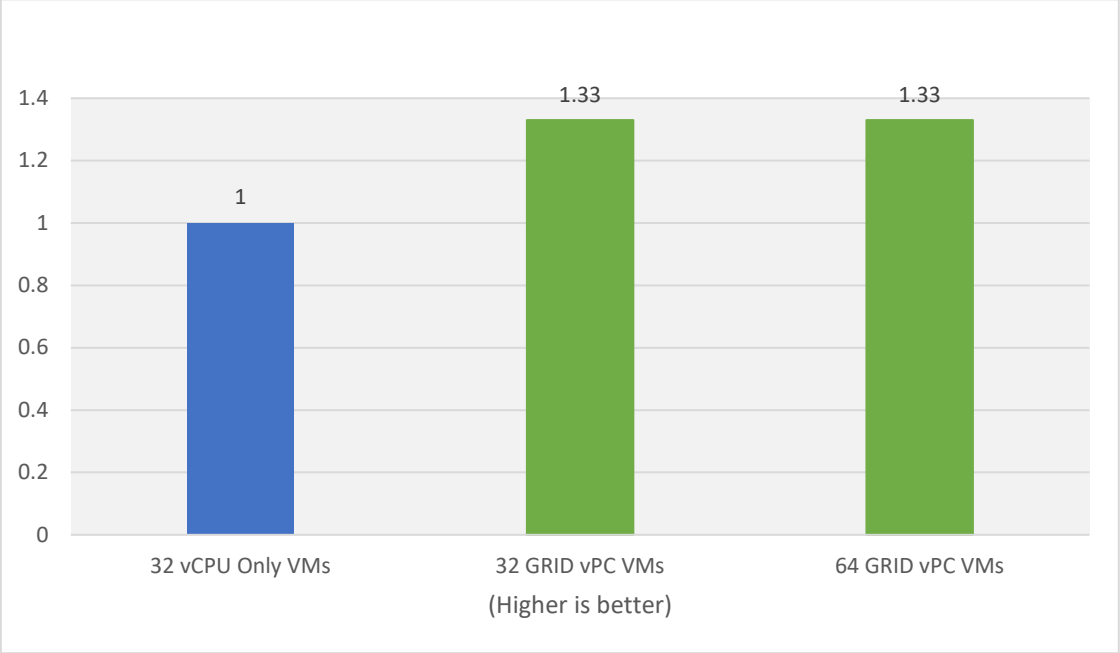
To further assess the trade-offs between end user experience and resource utilization, we used NVIDIA nVector built-in mechanisms to measure user experience. The following sections describe our findings for the nVector KW workload.

3.2.2.1 Frame Rate

The nVector benchmark tool captures frame rate which provides a great metric in determining the end user experience. Providing a consist and high frame rate can lead to a smoother experience for the user, while an inconsistent frame rate will create a less than acceptable experience.

The following figure illustrates the frame rate differences for dual high definition 1920 × 1080 monitors while running the nVector KW workload. The average frame rates were 1.3×'s more using NVIDIA GRID vPC than CPU only.

Figure 3-12. Frame Rate

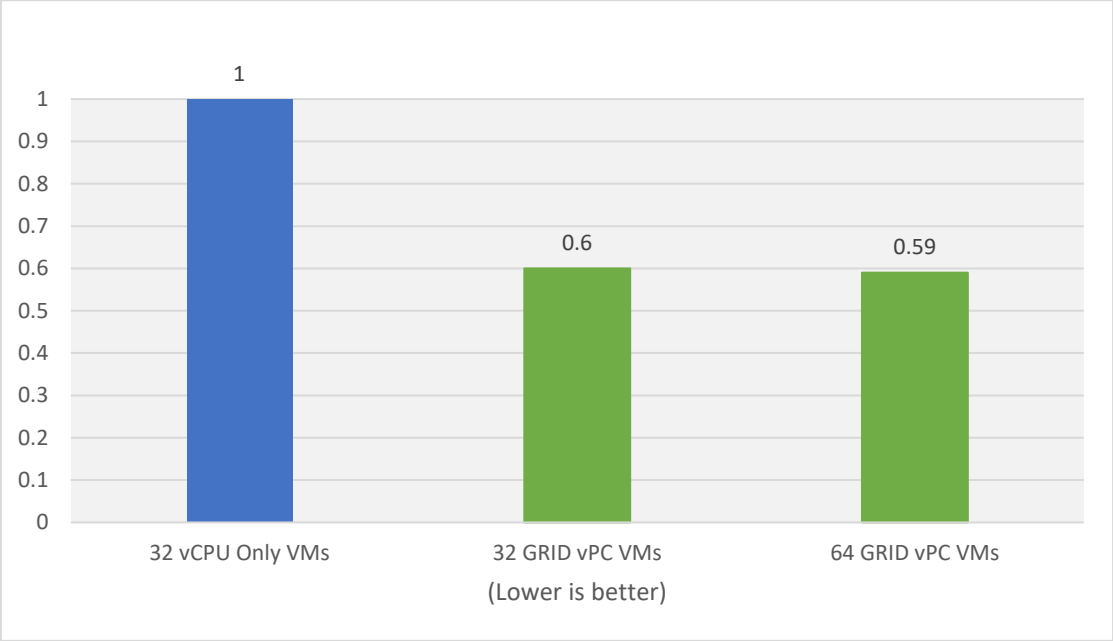


3.2.2.2 Latency Metrics

Another important metric captured by the nVector benchmark tool is latency or in this case end user latency. Latency can affect thing such as mouse speed, characters showing up on the screen behind what is typed, and poor video playback.

The following figure illustrates end user latency for dual high definition 1920 × 1080 monitors while running the nVector KW workload. User latency was 50% less using NVIDIA GRID vPC when compared to CPU-only VMs.

Figure 3-13. End User Latency

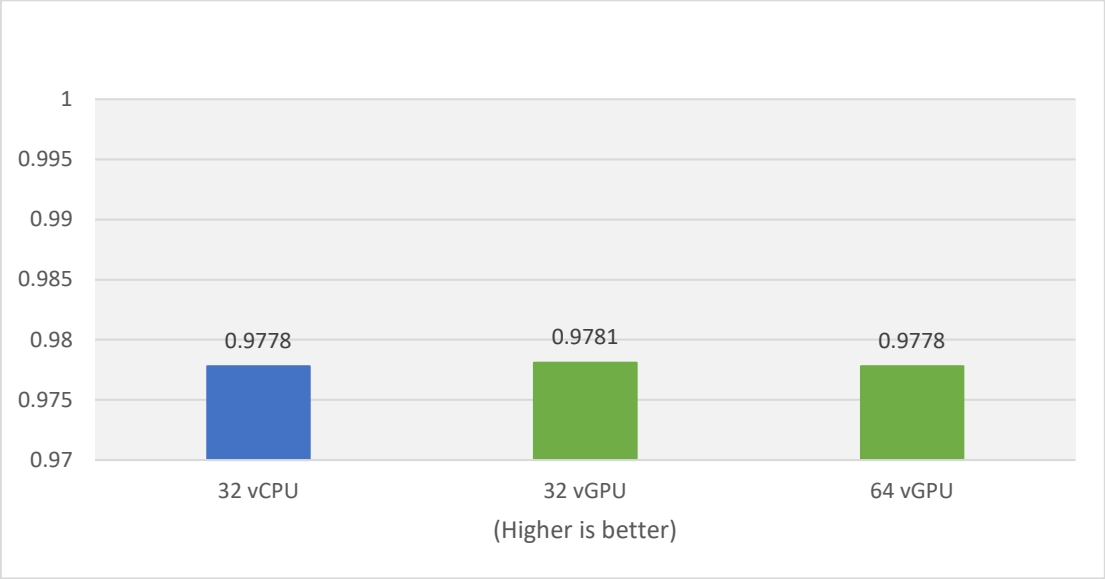


3.2.2.3 Image Quality

The nVector benchmark tool calculates image quality and it is determined by the remoting protocol, the configuration, and policies set in the VDI environment (refer to Appendix A regarding the configuration used within our testing). Poor image quality, under .90, can cause issues with text display, line sharpness, and other graphical issues.

Our nVector testing illustrates that GPU-accelerated VMs deliver uncompromised image quality as SSIM of the screen capture using dual high definition 1920 × 1080 monitors. Both CPU only and NVIDIA GRID vPC reported higher than the 0.90 threshold.

Figure 3-14. Image Quality



Chapter 4. Deployment Best Practices

4.1 Run a Proof of Concept

The most successful deployments are those that balance user density (scalability) with quality user experience. This is achieved when NVIDIA GRID vPC virtual machines are used in production while objective measurements and subjective feedback from end users is gathered.

Table 4-1. Proof of Concept

Objective Measurements	Subjective Feedback
Loading time of application	Overall user experience
Loading time of dataset	Application performance
Utilization (CPU, GPU, network)	Zooming and panning experience

4.2 Leverage Management and Monitoring Tools

NVIDIA GRID vPC on NVIDIA GPUs provides extensive monitoring features enabling IT to better understand usage of the various engines of an NVIDIA GPU. The utilization of the compute engine, the frame buffer, the encoder, and decoder can all be monitored and logged through a command line interface tool nvidia-smi, accessed on the hypervisor or within the virtual machine. In addition, NVIDIA vGPU metrics are integrated with Windows Performance Monitor (PerfMon) and through management packs like VMware vRealize Operations.

To identify bottlenecks of individual end users or of the physical GPU serving multiple end users, execute the following nvidia-smi commands on the hypervisor.

Virtual Machine Frame Buffer Utilization:

```
nvidia-smi vgpu -q -l 5 | grep -e "VM ID" -e "VM Name" -e "Total" -e "Used"
-e "Free"
```

Virtual Machine GPU, Encoder and Decoder Utilization:

```
nvidia-smi vgpu -q -l 5 | grep -e "VM ID" -e "VM Name" -e "Utilization" -e
"Gpu" -e "Encoder" -e "Decoder"
```

Physical GPU, Encoder and Decoder Utilization:

```
nvidia-smi -q -d UTILIZATION -l 5 | grep -v -e "Duration" -e "Number" -e
"Max" -e "Min" -e "Avg" -e "Memory" -e "ENC" -e "DEC" -e "Samples"
```

4.3 Understand Your Users

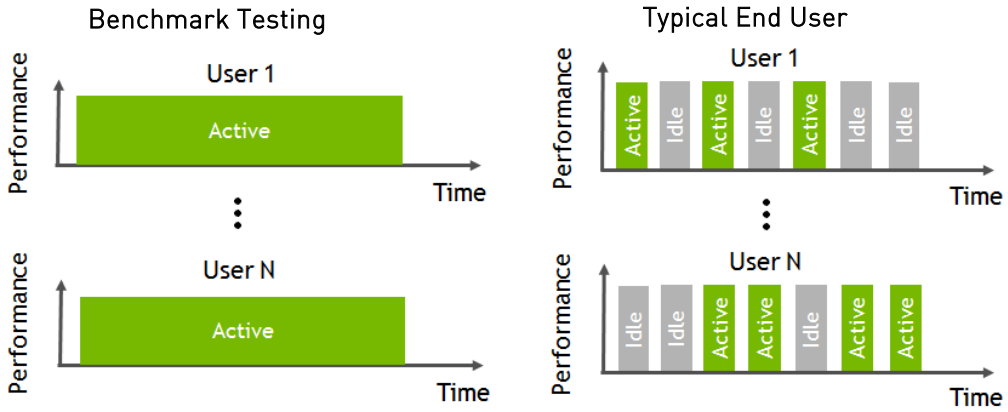
Another benefit of performing a POC prior to deployment is that it enables more accurate categorization of user behavior and GPU requirements for each virtual workstation. Customers often segment their end users into user types for each application and bundle similar user types on a host. Light users can be supported on a smaller GPU and smaller profile size while heavy users require more GPU resources, a large profile size and, may be best supported on an upgraded vGPU license like NVIDIA® Quadro® Virtual Data Center Workstation (Quadro vDWS).

4.4 Use Benchmark Testing

Benchmarks like nVector can be used to help size a deployment but they have some limitations. The nVector benchmarks simulate peak workloads, when there is the highest demand for GPU resources across all virtual machines. The benchmark does not account for the times when the system is not fully utilized, for which hypervisors, and the best effort scheduling policy can leverage to achieve higher user densities with consistent performance.

The graphic in Figure 4-1 demonstrates how workflows processed by end users are typically interactive, which means there are multiple short idle breaks when users require less performance and resources from the hypervisor and NVIDIA vGPU. The degree to which higher scalability is achieved is dependent on the typical day to day activities of your users, such as the number of meetings and the length of lunch or breaks, multi-tasking, etc. It is recommended to test and validate to meet the needs of your users.

Figure 4-1. Benchmark Testing



NVIDIA used the nVector benchmarking engine to conduct vGPU testing at scale. This benchmarking engine automates the testing process from provisioning virtual machines, establishing remote connections, executing KW workflow, and analyzing the results across all virtual machines. Test results shown in this specification are based on the nVector KW benchmarks which was running in parallel on all virtual machines with metrics averaged.

4.5 Understanding the GPU Scheduler

NVIDIA GRID vPC provides three GPU scheduling options to accommodate a variety of QoS requirements of customers. Additional information regarding GPU scheduling can be found here.

- Fixed share scheduling always guarantees the same dedicated quality of service.
- Best effort scheduling provides consistent performance at a higher scale and therefore reduces the TCO per user.
- Equal share scheduling provides equal GPU resources to each running VM. As vGPUs are added or removed, the share of GPU processing cycles allocated changes, accordingly, resulting in performance to increase when utilization is low, and decrease when utilization is high.

Chapter 5. Summary

When sizing an NVIDIA GRID vPC deployment for Knowledge Workers, NVIDIA recommends conducting a POC and fully analyzing resource utilization using objective measurements and subjective feedback. The best effort scheduler option is recommended for enterprise deployments, and user density will be dependent on the hardware configuration and user types.

To see how you can virtualize Digital Knowledge Worker workloads using NVIDIA GRID vPC software, try it for free. Learn more about NVIDIA GRID vPC here.

Appendix A. Frame Buffer Utilization Master List

Workload	NVIDIA GRID Card	Resolution	Monitors	Peak FB Usage	Average FB Usage	Idle FB Usage
vPC M10-1B Profile – 1 GB of Frame Buffer						
Knowledge	M10	1920x1080	1	68% (696 MB)	50.62% (518 MB)	28% (287 MB)
Knowledge	M10	1920x1080	2	69.67% (714 MB)	52.86% (541 MB)	40% (410 MB)
vPC T4-1B Profile – 1 GB of Frame Buffer						
Knowledge	T4	1920x1080	1	74% (758 MB)	61.37% (628 MB)	47% (481 MB)
Knowledge	T4	1920x1080	2	90% (922 MB)	76.07% (779 MB)	65% (666 MB)
Knowledge	T4	2560x1440 (QHD)	2	97.3% (996 MB)	84.06% (861 MB)	77% (788 MB)
vPC T4-2B Profile – 2 GB of Frame Buffer						
Knowledge	T4	2560x1440 (QHD)	2	68% (1,393 MB)	53.93% (1,104 MB)	42% (860 MB)
Knowledge	T4	2560x1440 (QHD)	3	88% (1,802 MB)	77.33% (1,583 MB)	53% (1,085 MB)
Knowledge	T4	2560x1440 (QHD)	4	97% (1,987 MB)	75.95% (1,556 MB)	65% (1,331 MB)
Knowledge	T4	4096x2160 (4K)	1	82.33% (1,686 MB)	59.96% (1,228 MB)	37% (756 MB)
Knowledge	T4	4096x2160 (4K)	2	95% (1,946 MB)	77.08% (1,579 MB)	58% (1,188 MB)
Knowledge	T4	5120x2880 (5K)	1	96% (1,966 MB)	71.52% (1,465 MB)	39% (799 MB)

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation ("NVIDIA") makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. 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 document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information 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 ("Terms of Sale"). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

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 this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform 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 document. NVIDIA accepts no 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 document 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 document. 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 document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND 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. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. 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 Terms of Sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, CUDA, NVIDIA GRID, NVIDIA Maxwell, NVIDIA Turing, Quadro, and TensorRT 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

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
http://www.nvidia.com