



NVIDIA Virtual Applications on Citrix Virtual Apps with VMware ESXi Sizing Guide

Application Sizing Guide

Document History

nv-quadro-vapps-citrix-sizing-guide-v1.3-011421.docx

| Version | Date | Authors | Description of Change |
|---------|--------------|-------------|-----------------------|
| 01 | Sep 8, 2020 | JJC, AS, EA | Initial Document |
| 02 | Jan 08, 2021 | CW | Version 1.2 |
| 03 | Jan 14, 2021 | AFS | Marking Update |

Table of Contents

| | |
|---|-----------|
| Chapter 1. Executive Summary..... | 5 |
| 1.1 What is NVIDIA vApps?..... | 5 |
| 1.2 Why NVIDIA vGPU? | 6 |
| 1.3 NVIDIA vGPU Architecture..... | 6 |
| 1.4 Recommended NVIDIA GPU's for NVIDIA vApps..... | 8 |
| Chapter 2. Sizing Methodology..... | 9 |
| 2.1 vGPU Profiles | 9 |
| 2.2 Virtual Machines per Physical Host | 9 |
| 2.3 vCPU Oversubscription | 10 |
| 2.4 Sizing vCPU | 10 |
| Chapter 3. Tools..... | 12 |
| 3.1 GPU Profiler | 12 |
| 3.2 NVIDIA System Management Interface (nvidia-smi)..... | 13 |
| 3.3 VMWare ESXtop | 14 |
| 3.4 VMWare vROPS | 14 |
| Chapter 4. Performance Metrics | 15 |
| 4.1 VDA Virtual Machine Metrics | 15 |
| 4.1.1 Framebuffer Usage | 15 |
| 4.1.2 vCPU Usage | 15 |
| 4.1.3 Video Encode/Decode | 16 |
| 4.2 Physical Host Metrics..... | 16 |
| 4.2.1 CPU Core Utilization..... | 16 |
| 4.2.2 GPU Utilization..... | 16 |
| Chapter 5. Test Findings | 17 |
| 5.1 Phase 1: Single VDA Testing | 17 |
| 5.2 Phase 2: Full Scale Testing..... | 19 |
| Chapter 6. Deployment Best Practices | 22 |
| 6.1 Understand Your Environment..... | 22 |
| 6.2 Run a Proof of Concept..... | 22 |
| 6.3 Leverage Management and Monitoring Tools | 23 |
| 6.4 Understand Your Users and Applications..... | 23 |
| 6.5 Use Benchmark Testing | 23 |
| 6.6 Understanding the GPU Scheduler..... | 23 |
| Chapter 7. Summary..... | 24 |
| 7.1 Process for Success..... | 24 |
| 7.2 Accelerate your Citrix Virtual Apps with NVIDIA vApps | 25 |

Appendix A. About NVIDIA nVector Benchmark 26

Appendix B. Microsoft Framebuffer Usage Patch..... 28

Appendix C. Lab Environment..... 29

Chapter 1. Executive Summary

This document provides insights into how to leverage NVIDIA Virtual Applications (vApps) for digital knowledge workers using Citrix Virtual Applications on VMware ESXi. It provides recommendations based on NVIDIA's nVector knowledge worker benchmarking and covers common questions such as:

- ▶ Which NVIDIA® GPU should I use for my business needs?
- ▶ How do I select the right NVIDIA virtual GPU (vGPU) profile(s) for Citrix VDAs for the types of users I have?
- ▶ How do I appropriately size my Citrix Virtual Apps environment with NVIDIA Virtual Applications?

Workloads will vary per user depending on many factors, including number of applications, the types of applications, and file sizes. The workload used for this document is from an internally developed NVIDIA benchmarking tool called nVector (refer to [7.2 Appendix A](#) for additional details). It is strongly 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 monitoring is important because user behavior can change over the course of a project and as the role of an individual changes in the organization. Users who were once light graphics users could become heavy graphics users when they change teams or are assigned a different project. Applications also have ever-increasing graphical requirements too. Management and monitoring tools enable administrators and IT staff to ensure their deployment is optimized. Through this document, you will gain an understanding of these tools, as well as the key resource usage metrics to monitor during your POC and product lifecycle.

1.1 What is NVIDIA vApps?

NVIDIA Virtual Applications (vApp) software is a GPU accelerated solution for organizations deploying Citrix Virtual Apps and Desktops, RDSH or other app streaming or multi-session-based solutions. Designed to deliver PC Windows applications at full performance, NVIDIA vApps allow users to access any Windows application at full performance on any device, anywhere. Windows Server hosted RDSH desktops are also supported by NVIDIA vApps.

NVIDIA vApps deliver an engaging user experience for the digital workplace. Employees can be more productive using modern applications and work the way they want, from anywhere. IT can cost-effectively scale virtualization to every employee with performance that rivals a physical PC. With GPU sharing, multiple VMs serving as Citrix VDA's can be powered by a single datacenter GPU,

maximizing utilization and affordability. With support for all major hypervisor virtualization platforms, including VMWare vSphere, datacenter admins can use the same management tools for their GPU-accelerated servers as they do for the rest of their data center.

Please refer to the [NVIDIA vGPU Licensing Guide](#) for additional information regarding feature entitlements which are included with the NVIDIA vApps software license.

1.2 Why NVIDIA vGPU?

NVIDIA vApps software is based upon NVIDIA virtual GPU (vGPU) technology and includes the NVIDIA graphics driver that is required by graphic intensive applications. NVIDIA vGPU enables multiple virtual machines (VMs) to have simultaneous, direct access to a single physical GPU. vGPU uses the same NVIDIA drivers that are deployed on non-virtualized operating systems. By doing so, NVIDIA vGPU provides VMs with high performance graphics and application compatibility, as well as cost-effectiveness and scalability, since multiple VMs can be customized to specific tasks that may demand more or less GPU compute or memory.

With NVIDIA vApps, you can gain access to the most powerful GPUs in a virtualized environment and gain vGPU software features such as:

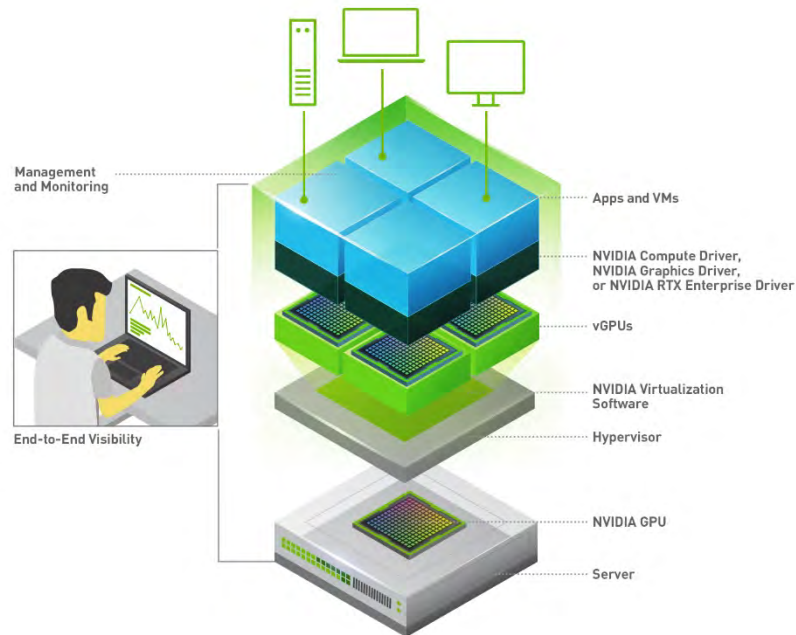
- ▶ Management and monitoring – streamline data center manageability by leveraging hypervisor-based tools.
- ▶ Live Migration – Live migrate GPU-accelerated VMs without disruption, easing maintenance and upgrades.
- ▶ Security – Extend the benefits of server virtualization to GPU workloads.
- ▶ Multi-Tenant – Isolate workloads and securely support multiple users.

Factors that should be considered during POC include things like which NVIDIA vGPU certified [OEM server](#) you've selected, which NVIDIA GPUs are supported in that platform, as well as any power and cooling constraints which you have may in your data center.


1.3 NVIDIA vGPU Architecture

The high-level architecture of an NVIDIA virtual GPU enabled virtual application environment is illustrated below in Figure 1.1. Here, GPUs are showing in the server, and the NVIDIA vGPU manager software (vib) is installed on the host server. This software enables multiple VMs to share a single GPU or if there are multiple GPUs in the server, they can be aggregated so that a single VM can access multiple GPUs. This GPU enabled environment provides an engaging user experience because graphics can be offloaded to the GPU and is no longer done by the CPU. Physical NVIDIA GPUs can support multiple virtual GPUs (vGPUs) and can be assigned directly to guest VMs under the control of NVIDIA's Virtual GPU Manager running in a hypervisor. Guest VMs use the NVIDIA vGPUs in the same manner as a physical GPU that has been passed through by the hypervisor. For NVIDIA vGPU deployments, the NVIDIA vGPU software automatically selects the correct type of license based on the vGPU type assigned.

Figure 1.1 NVIDIA vGPU Platform Solution Architecture



NVIDIA vGPUs are comparable to conventional GPUs in that they have a fixed amount of GPU Memory and one or more virtual display outputs or heads. Multiple heads support multiple displays. Managed by the NVIDIA vGPU Manager installed in the hypervisor, the vGPU Memory is allocated out of the physical GPU frame buffer at the time the vGPU is created. The vGPU retains exclusive use of that GPU Memory until it is destroyed.

 **Note:** These are virtual heads, meaning on GPUs there is no physical connection point for external physical displays.

All vGPUs resident on a physical GPU share access to the GPU's engines, including the graphics (3D) and video decode and encode engines. VM's guest OS leverages direct access to the GPU for performance and critical fast paths. Non-critical performance management operations use a para-virtualized interface to the NVIDIA Virtual GPU Manager.

1.4 Recommended NVIDIA GPU's for NVIDIA vApps

Density optimized GPUs are typically recommended for knowledge worker virtual applications such as office productivity applications, streaming video, and Windows 10. They are designed to maximize the number of concurrent sessions supported in a server.

| | NVIDIA T4 | NVIDIA M10 |
|------------------------|-------------------------|-----------------------------|
| # Cards [Architecture] | 1 (NVIDIA Turing™) | 1 (NVIDIA Maxwell™) |
| Memory Size | 16GB GDDR6 | 8 GB per GPU 32 GB GDDR5 |
| Form Factor | PCIe 3.0 Single Slot | PCIe 3.0 dual-slot |
| Power | 70W | 225 W |
| Thermal | Passive | Passive |
| Optimized for | Density and Performance | Density |

The NVIDIA® M10 is based upon Maxwell GPU architecture whereas the NVIDIA® T4 GPU is based on the newer generation NVIDIA Turing™ architecture. This document focuses on the T4 which offers the additional generational technology enhancements that includes support for VP9 decoding, which is often used for video playback, and H.265 (HEVC). The NVIDIA® M10 is still considered an acceptable GPU for vApps deployments.

The T4 is a low profile, 16 GB single-slot card, which draws 70 W maximum and does not require a supplemental power connector. 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.

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, therefore the amount of frame buffer that is useable by vGPU is reduced. It is important to resize your environment when switching from Maxwell GPUs to newer GPUs like Pascal and Turing GPUs. Additional information can be found [here](#).

Chapter 2. Sizing Methodology

It is highly recommended that a proof of concept is performed prior to a full deployment in order to gain a better understanding of how your users work and how much GPU resources they really need. This includes analyzing the utilization of all resources, both physical and virtual, as well as gathering subjective feedback to optimize the configuration to meet the performance requirements of your users and for best scale. Benchmarks like those used within the guide can be used to help size a deployment, but they have some limitations. The following sections cover topics and the methodology that should be considered for sizing.

2.1 vGPU Profiles

NVIDIA vGPU software allows you take an NVIDIA GPU and partition or fractionalize the datacenter GPU. These virtual GPU resources are then assigned to VMs in the VMWare vSphere management console using vGPU profiles. Virtual GPU profiles determine the amount of GPU framebuffer that can be allocated to your Citrix Virtual Deliver Agent (VDA) virtual machines. Determining the correct vGPU profile will improve your total cost of ownership, scalability, stability, and performance of your Citrix Virtual Apps environment.

It is important to consider which vGPU profile will be used within a deployment since this will ultimately determine how many vGPU backed VDA VMs will be deployed. When choosing a vGPU profile, it is important to balance the benefits of multiple VDAs, which provides greater redundancy by using a smaller vGPU profile, against the benefits of reduced licensing costs by lowering the number of VDAs and using a larger vGPU profile. With this in mind, an 8A or higher vGPU profile is recommended for Citrix Virtual Apps to reduce licensing cost, while maintaining redundancy at the VM layer. 1A, 2A, & 4A vGPU profiles are only recommended for use cases that involve a single user per VDA with a server operating system.

2.2 Virtual Machines per Physical Host

The number VMs serving as Citrix VDAs is dependent on the type of NVIDIA GPU being used, the number of physical GPUs in your host, and the vGPU profile assigned to the Citrix VDAs. Because GPU memory cannot be oversubscribed, the amount of total GPU memory that is available for all VDA VMs on a physical host is constant and how much is available is dependent on the type of NVIDIA GPU in the host server. For the testing scenario, 4 NVIDIA T4s are used per server. The T4 has 16 GB of GPU memory, which provides a total of 64 GB of GPU memory per server.

Overall, to fully utilize all GPUs in a host server, the total number of VDA VMs per host should equal the total amount of GPU memory in the host divided by the amount of framebuffer per VM (assigned to the VM using a vGPU profile). As discussed previously, this document will focus on the 8A profile for the environment to balance VDA VM redundancy and licensing cost reduction. The following formula was used:

$$\text{\# of VDA VMs} = \frac{\text{\# of T4's in host} \times \text{GPU Memory on T4 card}}{\text{Framebuffer of vGPU Profile}}$$

$$\text{\# of VDA VMs} = \frac{4 \times 16}{8}$$

$$\text{\# of VDA VMs} = 8$$

2.3 vCPU Oversubscription

Most modern server-based CPUs and hypervisor CPU schedulers have feature sets (e.g. Intel's Hyperthreading or AMD's Simultaneous Multithreading) that allows for oversubscribing CPU resources. This means that the total number of virtualized CPUs (vCPU) can be greater than the total number of physical CPU cores in a server. The oversubscribing ratio can have a dramatic impact on the performance and scalability of your NVIDIA vApps deployment via Citrix Virtual Apps. Citrix has found that a 2:1 oversubscription ratio is optimal for most Citrix Virtual Apps workloads.

One of the primary factors that impacts performance and scalability of an NVIDIA vApps deployment via Citrix Virtual Apps is the workload itself. For most task and knowledge worker workloads, NVIDIA recommends utilizing a 2:1 CPU oversubscription ratio as a starting point. Actual oversubscription ratios may vary depending on your application and VDA VM mix.

For additional information on CPU over-subscription in a Citrix Virtual App environment, refer to the [Citrix Virtual Apps and Desktops Single-Server Scalability Product Documentation](#). In general, magic multipliers are not recommended for graphic intensive workloads such as the nVector workload (refer to 7.2 Appendix A for additional details) which was used for our sizing purposes.

2.4 Sizing vCPU

Allocating the correct number of vCPUs per VDA VM is crucial for optimizing the performance and scalability of your NVIDIA vApps via Citrix Virtual Apps deployment. Choosing the correct number of vCPUs to assign to VDA VM is dependent on the number of physical cores in your host server, the number of VMs per host, and the CPU over-subscription ratio.

The following formula describes how to calculate the number of vCPU's per VDA VM.

$$\text{vCPU per VDA VM} = \frac{(\text{Cores per Host}) \times (\text{Oversubscription Ratio})}{(\text{VDA VMs per Host})}$$

For this test environment, the Intel Xeon Gold 6240R processor, in a dual-socket Rack Server was used which has 24 cores, therefore there are **48 total cores** in the dual-socket rack server. As discussed, in Virtual Machines per Physical Host, the T4-8A vGPU profile was assigned to VDA VMs, and there were **8 VDA VMs per host**. Using an **oversubscription ratio of 2:1** the number of vCPUs per VDA VM is 12 vCPU. The formula is as follows:

$$\text{vCPU per VDA VM} = \frac{48 \times 2}{8}$$

$$\text{vCPU per VDA VM} = 12 \text{ vCPU}$$

Choosing a **12 vCPU** configuration for the VDA VMs ensures that the application stack will take advantage of the advanced feature sets offered by modern server-based CPUs and hypervisor CPU schedulers which will improve scalability and performance.

Some environments may see increased performance when using higher oversubscription ratios. In the test environment, 12 vCPU are utilized as a starting point to ensure all CPU and hypervisor feature sets are used. Increasing the vCPU count, thereby increasing the oversubscription ratio, may be beneficial to your environment and should be examined during your POC/trial period.

Chapter 3. Tools

There are several NVIDIA specific and third-party industry tools that can help validate your POC while optimizing for the best user density and performance. The tools covered in this section are:

- ▶ GPU Profiler
- ▶ NVIDIA-SMI
- ▶ ESXtop
- ▶ vROPS

These tools allow you to analyze the utilization of all resources, both physical and virtual, to optimize the configuration to meet the performance requirements of your users and for best scale. These tools are useful during your POC to ensure your test environment will not only accurately represent production, but also in a live production environment. It is important to continually use these tools to help ensure system health, stability, and scalability, as your deployment needs will likely change over time.

3.1 GPU Profiler

GPU Profiler (available on GitHub) is a 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 in order to ensure acceptable user performance. GPU Profiler can be run on a single VM with various vGPU profiles. The following metrics can be captured:

- ▶ Framebuffer %
- ▶ GPU Utilization
- ▶ vCPU %
- ▶ RAM %
- ▶ Video Encode
- ▶ Video Decode
- ▶ Session Count



3.2 NVIDIA System Management Interface (nvidia-smi)

The built-in NVIDIA vGPU Manager provides extensive monitoring features to enable IT to better understand usage of the various engines of an NVIDIA vGPU. 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.

To identify bottlenecks of the physical GPU, which is serving Citrix VDA VM's, execute the following `nvidia-smi` commands on the hypervisor in a Shell session using SSH.

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"
```

Additional information regarding `nvidia-smi` is located [here](#). It is important to note, option `-f FILE, --filename=FILE`, which can redirect query output to a file (for example, `.csv`).

3.3 VMWare ESXtop

ESXtop is a VMware tool for capturing host-level performance metrics in real time. It can display information about physical host state information for each processor, the host's memory utilization, as well as the disk and network usage. VM level metrics are also captured.

Collecting ESXtop and piping it directly into a zip file is usually the preferred capture method to reduce disk space usage. Below is an example command to capture a one-hour data sample.

```
esxtop -b -a -d 15 -n 240 | gzip -9c > esxtopoutput.csv.gz
```

“-b” stands for batch mode, “-a” will capture all metrics, “-d 15” is a delay of 15 seconds and “-n 240” is 240 iterations resulting in a capture window of 3600 seconds or one hour.

Additional information on VMware's ESXtop can be found [here](#).

3.4 VMWare vROPS

NVIDIA Virtual GPU Management Pack for VMware vRealize Operations enables you to use a VMware vRealize Operations cluster to monitor the performance of NVIDIA physical GPUs and virtual GPUs.

VMware vRealize Operations provides integrated performance, capacity, and configuration management capabilities for VMware vSphere, physical and hybrid cloud environments. It provides a management platform that can be extended by adding third-party management packs. For additional information, see the [VMware vRealize Operations documentation](#).

NVIDIA Virtual GPU Management Pack for VMware vRealize Operations collects metrics and analytics for NVIDIA vGPU software from virtual GPU manager instances. It then sends these metrics to the metrics collector in a VMware vRealize Operations cluster, where they are displayed in custom NVIDIA dashboards.

Additional information on NVIDIA's Virtual GPU Management Pack for VMware vRealize Operations can be found [here](#).

Chapter 4. Performance Metrics

The tools described in [Chapter 3](#) allow you to capture key performance metrics which are discussed in the upcoming sections. It is important to collect metrics during your POC as well as on a regular basis in a production environment to ensure optimal virtual application delivery.

Within a Citrix Virtual Application environment, there are two tiers of metrics which can be captured: VM level (Citrix VDA VMs) and Server level. Each tier has its own performance metrics, and all must be validated to ensure optimal performance and scalability.

4.1 VDA Virtual Machine Metrics

As mentioned in [Chapter 3](#), the GPU Profiler and VMware vRealize Operations (vROPS) are both great tools for understanding resource usage metrics within VMs. The following sections cover the metrics captured by these tools in greater detail and are useful during a POC for monitoring an existing deployment in order to further understand potential performance bottlenecks.

4.1.1 Framebuffer Usage

In a virtualized environment, framebuffer is the amount of vGPU memory that is exposed to the guest operating system. If high framebuffer utilization is noted, then the vApps VM is more prone to produce suboptimal user experience with potentially degraded performance and crashing. Since users interact and work differently within software applications, it is recommended performing your own POC with your workload to determine framebuffer thresholds within your environment. A good rule of thumb to follow is that a VM's framebuffer usage should not exceed 90% frequently to ensure optimal user experience.

4.1.2 vCPU Usage

When using NVIDIA vApps, vCPU usage can be just as important as the VM's vGPU framebuffer usage. Since all workloads require CPU resources, vCPU usage should not bottleneck and is crucial for optimal performance. Even when a process is programmed to utilize a vGPU for acceleration, vCPU resources will still use it to some level.

4.1.3 Video Encode/Decode

NVIDIA GPUs contain a hardware-based encoder and decoder which provide fully accelerated hardware-based video decoding and encoding for several popular codecs. Beginning with the GPU Kepler generation, complete encoding (which can be computationally complex) is offloaded from the CPU to the GPU using NVENC. Hardware based decoder (referred to as NVDEC) provides faster real-time decoding for video playback applications. When NVIDIA hardware-based encoder and decoder are being used, usage metrics can be captured.

Using Citrix HDX3DPro, the Video Encoder Sessions metric will show how many active user connections exist on the Citrix VDA VM. Video Encoder Usage metric captures the utilization of the encoder on the NVIDIA GPU by the protocol.

4.2 Physical Host Metrics

As mentioned in [Chapter 3](#), the NVIDIA System Management Interface (nvidia-smi) and VMWare ESXtop are both great tools for understanding resource usage metrics for a physical host. The following sections cover the metrics which are useful during a POC or for monitoring an existing deployment to further understand potential performance bottlenecks.

4.2.1 CPU Core Utilization

VMWare's ESXtop utility is used for monitoring physical host state information for each CPU processor. The % Total CPU Core Utilization is a key metric to analyze to ensure optimal VM performance. As mentioned previously, each process within a VM will be executed on a vCPU; therefore, all processes running within a VDA VM will utilize some portion of physical cores on a host for execution. If there are no available host threads for execution, processes in a VM will be bottlenecked and can cause significant performance degradation.

It is also important to note, how a hypervisor scheduler allocates CPU resources to each VM can also result in bottlenecks. Following the guidelines outlined in [Chapter 2](#) will help ensure these bottlenecks do not occur in your environment.

4.2.2 GPU Utilization

NVIDIA System Management Interface (nvidia-smi) is used for monitoring GPU Utilization rates which reports how busy each GPU is over time and it can be used to determine how much Citrix VDA VM's are using the NVIDIA GPUs in the host server.

For most vApps deployment, GPU usage will remain well below 100%. If you find that your deployment is encountering 100% GPU Utilization during your POC, reach out to your NVIDIA representative. If there are no issues with your deployment strategy, you may need to upgrade your deployment from session-based to a traditional VDI which can be licensed with NVIDIA vPC or RTX vWS.

Chapter 5. Test Findings

5.1 Phase 1: Single VDA Testing

The first phase of testing explored the impact of increased user density and varied vCPU count. During this phase, the GPU framebuffer and vCPU usages were closely analyzed on the VDA VM to ensure correct sizing. Tests were executed on a single VDA VM using a T4-8A vGPU profile and user sessions were executing the nVector Knowledge worker workload (refer to 7.2Appendix A for additional details) via the Citrix HDX-3D Pro protocol connecting to a client with a single HD (1920x1080) monitor.

Tests were first executed on a correctly sized VDA VM using 12vCPU with a T4-8A profile and 10 user sessions. The following graph illustrates that GPU framebuffer utilization remained within the expected thresholds and are not bottlenecked. This means the VM had plenty of head room in terms of vGPU memory to provide a rich end-user experience.

Figure 5-1 GPU Framebuffer Utilization

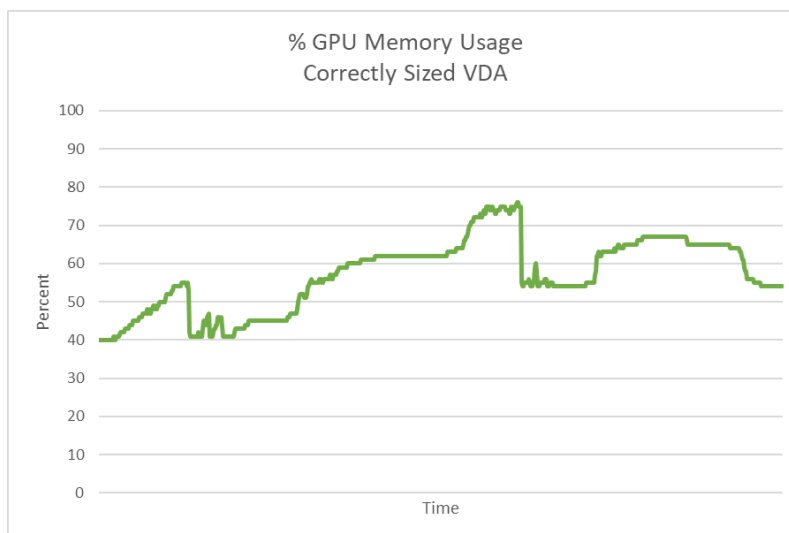
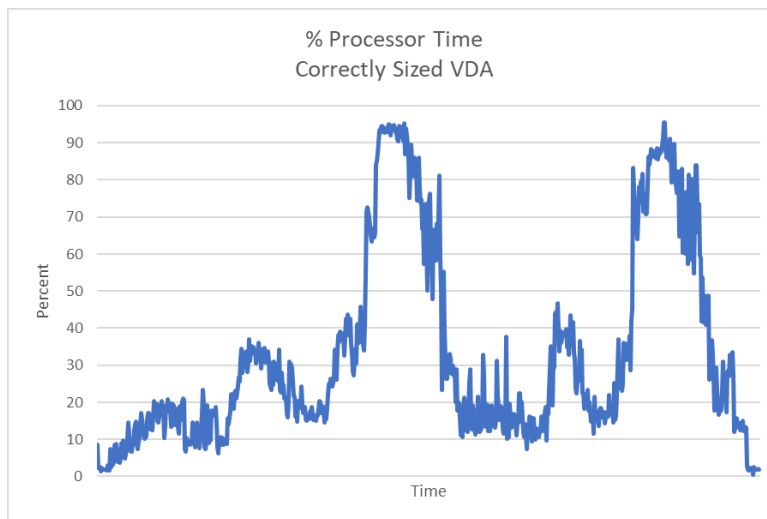


Figure 5-2 illustrates the vCPU metrics on the Citrix VDA VM (% CPU Processor Time). During peak test execution, CPU utilization reached past 90% therefore the vCPU was adequately being used but did not reach the point where CPU resources were flat lining and becoming a bottleneck.

Figure 5-2 vCPU Present Processor Time



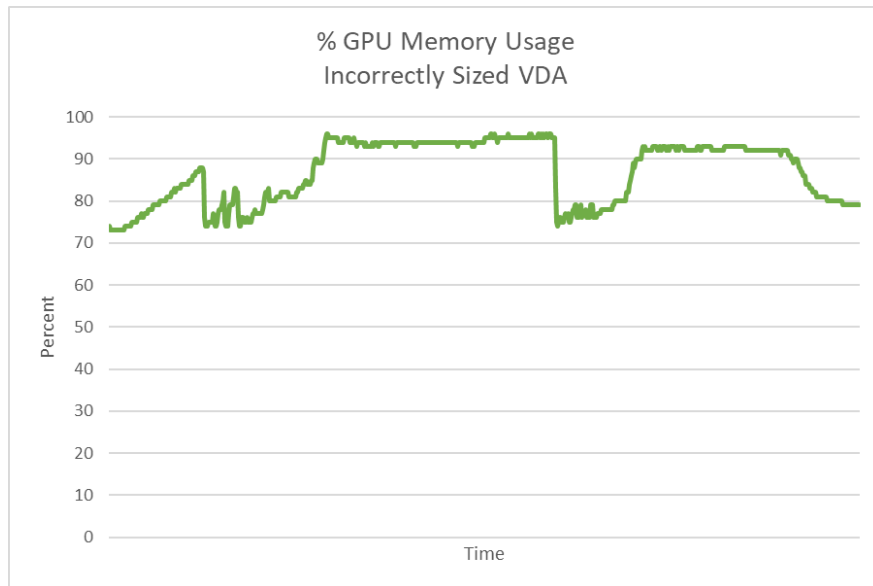
A VDA's % Processor Time continuously remained beyond 95% for an extended period of time. This is an indication that there is a vCPU bottleneck, therefore no threads are available for execution, which can severely degrade performance.

Overall, when comparing vGPU and vCPU usage metrics, the Citrix VDA VM is correctly sized with 12vCPU with a T4-8A profile and 10 user sessions. This resulted in a balanced configuration based upon the nVector workload, but if a bottleneck were to occur, the VM is more prone to run out of vCPU rather than vGPU resources.

Now let us look at a second test scenario where user connection count doubled, and the amount vCPU allocated to the Citrix VDA VM was increased (24 vCPU on single VDA VM with a T4-8A profile, and 20 Citrix HDX-3D Pro connections). In this scenario, end user experience can suffer drastically when the number of concurrent sessions sharing the vGPU resources have exhausted the available framebuffer. Allocating more vCPU resources to the VM will not make up for the lack of vGPU resources. The only way to improve performance is to reduce the user count.

The following graph illustrates the GPU Memory usage when 20 users are connected to the single VDA VM.

Figure 5-3 % GPU Memory Usage with an Incorrectly Sized VDA



In this test scenario, the Citrix VDA VM framebuffer has been exhausted and is exceeding 90% frequently as seen in the above graph, which can result in severely decreased user experience, performance, and crashes. Workloads will vary depending on many factors, including number of applications, the types of applications, and file sizes. It is highly recommended that you test your own workloads during a POC, since mileage may vary. The nVector test results in this document should be used for guidance purposes only.

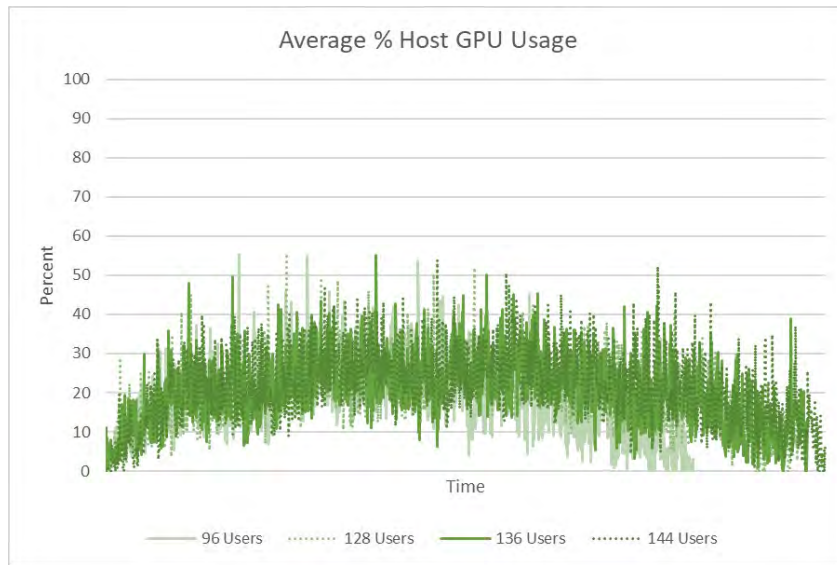
Overall, the results of Phase 1 single VDA testing were used for guidance and sizing purposes but will also be used in Phase 2 scalability testing. Single VDA VM testing does not provide an accurate representation of how deployments will respond at scale since the advanced feature sets offered by modern server-based CPUs, like hyperthreading, is not fully utilized.

5.2 Phase 2: Full Scale Testing

Phase 2 testing focused on increasing user count as well as the number of Citrix VDA VMs. Tests were executed with 8 Citrix VDA VMs that were configured with the optimal vCPU count and vGPU profile used within Phase 1; 12 vCPU and T4-8A profiles. Tests were scaled to 96, 128, 136, and 144 concurrent user sessions. User sessions were scaled in order to determine the maximum number of concurrent user sessions which the server could support based upon available physical server resources. With this in mind, Phase 2 testing primarily focused on analyzing Host resource metrics in order to identify potential bottlenecks.

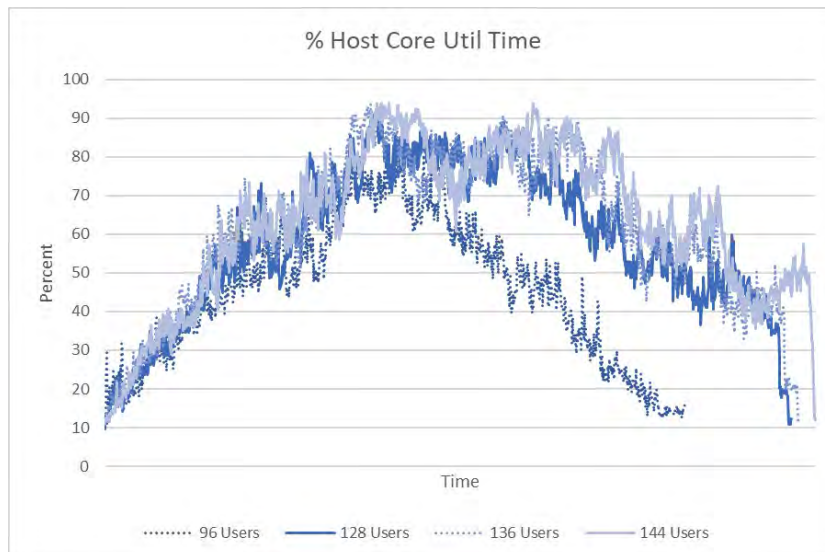
The graph below illustrates host GPU utilization as the number of concurrent user connections scaled from 96, 128, 136, & 144. It is important to note, host GPU utilization rates illustrated in Figure 5-4 indicates there is not a GPU bottleneck. Meaning, the server has plenty of head room with the GPU compute engine. GPU Utilization is being reported by averaging utilization across the four NVIDIA T4 GPUs in the server.

Figure 5-4 GPU Utilization as Concurrent User Connections Are Scaled



The following figure illustrates total CPU Core utilization using Intel® Xeon® Gold 6240R @ 2.40 GHz for 96, 128, 136, 144 concurrent user sessions. Comparing the previously shown host GPU graph to CPU usage below, the utilization metrics illustrate that if a bottleneck were to occur on the host, CPU would be more likely to be the bottleneck.

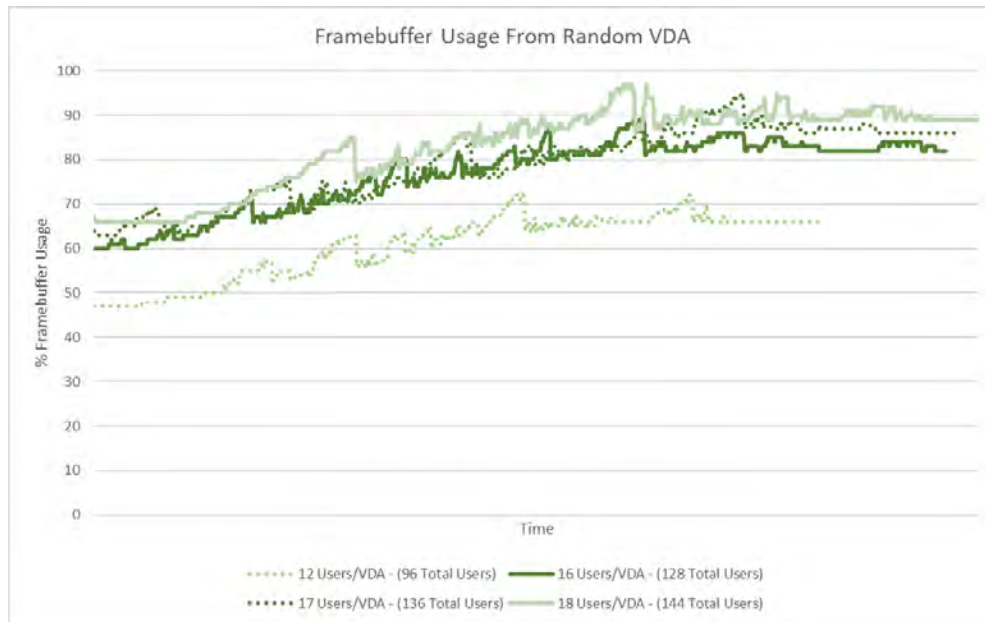
Figure 5-5 CPU Utilization as Concurrent User Connections Are Scaled



During peak test execution, total CPU Core Util Time reached 96% only for a short period, therefore it may appear that the server resources are adequate to support 144 concurrent users as tests scaled. However, nearly all physical CPU cycles have been exhausted and the probability of pegging all CPU threads is extremely high.

At this point, it is important to continue performance investigation by looking further into the different tiers within the architecture, specifically at the VM level GPU framebuffer metrics. The below graph shows the framebuffer usage from one of the eight Citrix VDA VMs as tests were scaled.

Figure 5-6 Framebuffer Usage of Citrix VDA VMs as Tests Are Scaled



As mentioned in section [4.1.1](#), to provide engaging user experience, framebuffer utilization should not exceed 90% frequently. As the number of concurrent user sessions scaled across the 8 Citrix VDA VMs, the number of users per VM increased from 12, 16, 17, 18 concurrent users.

With 18 concurrent users (144 total sessions on the server), the framebuffer usage on the VM exceeded 90% frequently once user count scaled. The 16 concurrent users per Citrix VDA VM test scenario was the most optimal, since the framebuffer usage did not exceed 90% frequently, resulting in best user experience. Therefore, 16 concurrent users per Citrix VDA VM is the maximum amount of concurrent user sessions that would be recommended for our nVector knowledge workload.

This test scenario illustrated that Intel® Xeon® Gold 6240R @ 2.40 GHz, has enough Server resources to support a maximum of 128 concurrent users for the nVector workload. However, it important to keep in mind that choosing the correct CPU for virtualization and proper configuration can have a direct effect on scalability even when a virtual GPU is present. In terms of CPU specs, you should evaluate the number of cores and clock speed. It is highly recommended that you test your own server and workloads during a POC since mileage may vary.

Chapter 6. Deployment Best Practices

6.1 Understand Your Environment

IT infrastructure is highly complex involving multiple server types, with varying CPUs, memory, storage, and networking resources. Deployments often involve a geographically dispersed user base, with multiple datacenters, and a mixture of cloud-based compute and storage resources. Define the scope of your deployment around these variables and run a POC for each of the scoped deployment types.

Other factors include things like which NVIDIA vGPU certified OEM server you've selected, which NVIDIA GPUs are supported in that platform, as well as any power and cooling constraints which you have may in your datacenter. For further information regarding installation and server configuration steps, please refer to the NVIDIA vGPU Citrix Virtual Apps & Desktops on VMware vSphere Deployment Guide.

6.2 Run a Proof of Concept

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

| 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 |

6.3 Leverage Management and Monitoring Tools

As discussed in [Chapter 3](#), there are several NVIDIA specific and third-party industry tools that will help validate that your deployment and to ensure it is providing an acceptable end-user experience and optimal density. Failure to leverage these tools can result in additional unnecessary risk and poor end-user experience.

6.4 Understand Your Users and Applications

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 application. Light applications used by knowledge workers or task workers can be supported on a GPU intended for vApps, while GPU intensive applications may be best supported on an upgraded vGPU license like NVIDIA Virtual PC (vPC) or RTX Virtual Workstation (vWS). Work with your application ISV and NVIDIA representative to help you determine the correct license(s) for your deployment needs.

6.5 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.

6.6 Understanding the GPU Scheduler

vApps provide three GPU scheduling options to accommodate a variety of QoS requirements of customers. For a knowledge worker workload, NVIDIA recommends using the default scheduler of **Best effort scheduling**. Additional information regarding GPU scheduling can be found [here](#).

- ▶ **Fixed share scheduling** guarantees the same dedicated quality of service at all times.
- ▶ **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 7. Summary

The most successful customer deployments start with a proof of concept (POC) and are “tuned” throughout the lifecycle of the deployment. Management and monitoring tools enable administrators and IT staff to ensure their deployment is optimized for each user. Due to applications being used in different ways, it is recommend performing your own POC with your workload. The results shared within this guide are reflective of the work profile captured within the nVector workload. 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.

7.1 Process for Success

Successful NVIDIA vApps deployments follow these steps to deliver a rich accelerated end user experience via Citrix Virtual Applications.

- ▶ Scope your environment for the needs of each application and user type.
- ▶ Implement the NVIDIA recommended sizing methodology for your Citrix VDAs.
- ▶ Run a proof of concept for each deployment type.
- ▶ Utilize benchmark testing to help validate your deployment.
- ▶ Utilize NVIDIA specific and industry wide performance tools for monitoring.
- ▶ Ensure performance and experience metrics are within acceptable thresholds.

NVIDIA vApps are most suitable for mid-range graphic applications. If you find that your deployment is encountering 100% GPU Utilization during your POC, you may need to upgrade your deployment due to the additional graphic requirements from higher-end software applications. As such, NVIDIA vApps session-based solution is not ideal and upgrading to a traditional VDI, licensed with NVIDIA vPC or RTX vWS, may be more suitable.

7.2 Accelerate your Citrix Virtual Apps with NVIDIA vApps

The need for accelerated computing in a Citrix Virtual Apps environment has grown as the demands of end users have increased, and application workloads increasingly rely on accelerated visual computing platforms. Workloads like webinars, training videos, and 3-D web content can all benefit from NVIDIA vApps for Citrix Virtual Applications. NVIDIA vApps can accelerate, enhance, and provide a rich user experience that will bring your Citrix Virtual Apps deployment into the modern age of application virtualization. To see how you can virtualize Digital Knowledge Worker workloads using NVIDIA vApps software, [try it for free](#).

Appendix A. 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. Tasks within these applications include scrolling, zooming, menu navigation, and PDF creation. Google Chrome streams 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 Citrix Virtual Desktop Agents (VDAs).

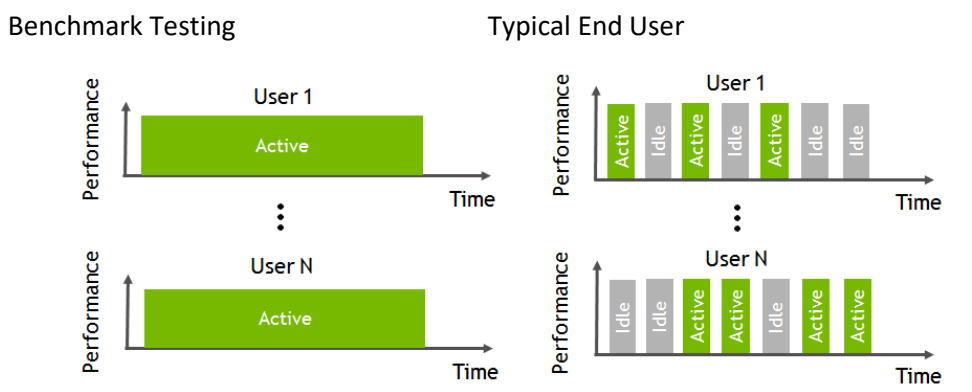
Simulating Many Users, Many Behaviors

| User #1 | User #2 | 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 3. Characteristics of NVIDIA's Benchmarking Tool. The above table shows the workflow of each user. The graph 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.

The graphic below 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.



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 application guide are based the nVector KW benchmarks, which was run in parallel on all virtual machines with metrics averaged.

Appendix B. Microsoft Framebuffer Usage Patch

KB4586830 & KB458639 address an issue with incorrect Canonical Display Driver (CDD) buffer flushing, which degrades performance in Remote Desktop Protocol (RDP) Windows 2000 Display Driver Model (XDDM) scenarios. This issue affects applications that use graphics processing units (GPU) to operate, such as Microsoft Teams, Microsoft Office, and web browsers.

- ▶ Server 2016 – [KB4586830](#)
- ▶ Server 2019 – [KB4586839](#)

Please follow the steps below to enable KB4586830 on Server 2016. It is not enabled by default post installation.

- To Enable the fix - reg add
HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Policies\Microsoft\FeatureManagement\Overrides /v 1826589834 /t REG_DWORD /d 1 /f
- To Disable the fix - reg add
HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Policies\Microsoft\FeatureManagement\Overrides /v 1826589834 /t REG_DWORD /d 0 /f

Appendix C. Lab Environment

Testing leveraged two physical servers with one hosting the target vApps Citrix VDA VMs and the second hosting the virtual clients. Both server hosts used VMware vSphere ESXi 6.7.0 and NVIDIA Virtual GPU Manager 11. The target VM acts as a standard vApps Citrix VDA 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. To capture a wide variety of common knowledge worker applications as mentioned in [Appendix A](#), a published desktop was used with multiple applications installed on the multi-session Server operating system.

In order to replicate the most basic deployment possible in our testing, the test environment only implemented a single Citrix policy, "Optimize for 3D Graphics Workloads." Additional policy configuration may be needed to improve performance and scalability for your environment. For additional information on Citrix policies that can be used to improve user experience see the [Graphics Section of the Citrix Virtual Apps and Desktops Product Documentation](#).

The following table further describes the lab environment used for this testing:

| Host Configuration | Citrix VDA VM Configuration | Virtual Client |
|------------------------------------|--|---|
| PowerEdge R740xd Rack Server | vCPU: 4-16 | vCPU: 4 |
| Intel® Xeon® Gold 6240R @ 2.40 GHz | vRAM: 24-64 GB | vRAM: 4 GB |
| VMware ESXi, 6.7.0, 15160138 | NIC: 1 (vmxnet3) | NIC: 1 (vmxnet3) |
| Number of CPUs: 48 (2 x 24) | Hard disk: 120 GB | Hard disk: 40 GB |
| Memory: 768 GB | Virtual Hardware: vmx-13 | Virtual Hardware: vmx-13 |
| Storage: Local Flash | Citrix Virtual Apps and Desktops 7 1912 LTSR | Citrix Workspace App 1912 LTSR |
| Power Setting: High Performance | HDX3DPro | HDX3DPro |
| GPU: 4 x T4 | vGPU Driver: 11.0 (Windows Driver 451.48) | vGPU Driver: 11.0 (Windows Driver 451.48) |

| | | |
|--|--|--------------------------------------|
| Scheduling Policy: 0x00 (Default - Best Effort) | Guest OS: Windows Server 2016 Standard 1607 | Guest OS: Windows 10 Enterprise 1903 |
| Citrix Policy: Optimize for 3D Graphics Workloads | Resolution: 1920x1080 | Resolution: 1920x1080 |

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 OptiX, NVIDIA RTX, NVIDIA Turing, Quadro, Quadro RTX, and TensorRT 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

© 2020-2021 NVIDIA Corporation. All rights reserved.