NVIDIA RTX Virtual Workstation

Sizing Guide

Document History

nv-quadro-vgpu-deployment-guide-citrixonvmware-v2-082020

Version	Date	Authors	Description of Change
01	Aug 17, 2020	AFS, JJC, EA	Initial Release
02	Jan 08, 2021	CW	Version 2
03	Jan 14, 2021	AFS	Branding Update

Table of Contents

List of Figures

List of Tables

Chapter 1. Executive Summary

This document provides insights into how to deploy NVIDIA® RTX® Virtual Workstation (RTX vWS) software for creative and technical professionals. It 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 the types of users I will have?
- How do I appropriately size my Virtual Workstation environment?

Workloads will vary per user depending on many factors, including number of applications, the types of applications, file sizes, monitor resolution and number of monitors. 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 allows 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 allow 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 RTX vWS?

With NVIDIA RTX vWS software, you can deliver the most powerful virtual workstation from the data center. This frees the most innovative professionals to work from anywhere and on any device, with access to the familiar tools they trust. Certified with over 140 servers and supported by every major public cloud vendor, vWS is the industry standard for virtualized enterprises. NVIDIA RTX vWS is used to virtualize professional visualization applications, which benefit from the NVIDIA RTX Enterprise drivers and ISV certifications, support for NVIDIA CUDA® and OpenCL, higher resolution displays, and larger profile sizes.

Please refer to the NVIDIA vGPU Licensing Guide for additional information regarding feature entitlements that are included with the NVIDIA RTX vWS software license.

1.2 Why NVIDIA vGPU?

NVIDIA RTX vWS software is based upon NVIDIA virtual GPU (vGPU) technology and includes the NVIDIA RTX Enterprise driver that is required by graphic intensive applications. NVIDIA vGPU allows multiple virtual machines (VMs) to have simultaneous, direct access to a single physical GPU, or multiple physical GPUs can be aggregated and allocated to a single VM. 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 RTX vWS, you can gain access to the most powerful GPUs in a virtualized environment and gain vGPU software features such as:

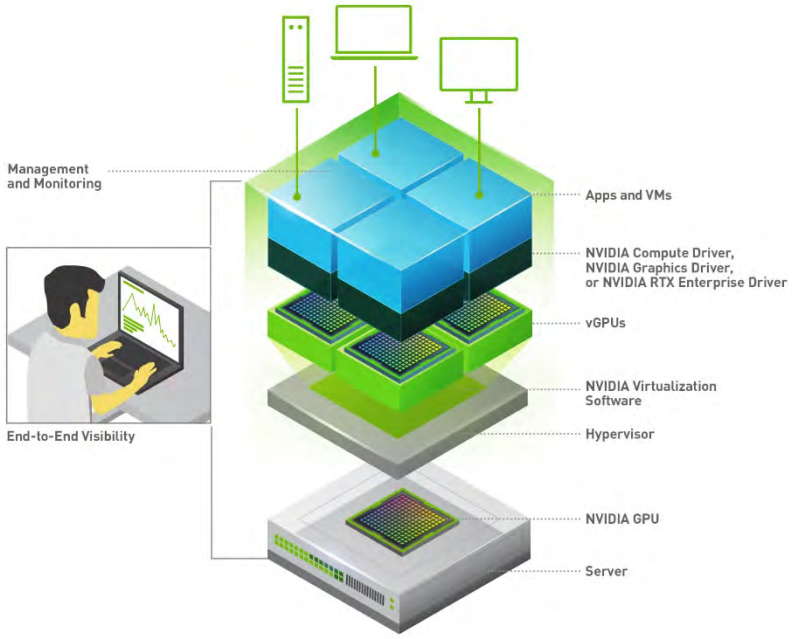
- Management and monitoring – streamline data center manageability by leveraging hypervisor-based tools.
- Live Migration – Live migrate GPU-accelerated VMs without disruption, easing maintenance and upgrades.
- Security – Extend the benefits of server virtualization to GPU workloads.
- Multi-Tenant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

Factors that should be considered during POC include items such as: 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 may have in your data center.

1.3 NVIDIA vGPU Architecture

The high-level architecture of an NVIDIA virtual GPU enabled environment is illustrated below in Figure 1.1. NVIDIA GPUs are installed 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 versus being delivered 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 identifies the appropriate vGPU license based upon the vGPU profile which is assigned to a VM.

Figure 1.1 NVIDIA vGPU 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.

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 RTX vWS

Table 1.1 lists the hardware specification for NVIDIA GPU's that are recommended for NVIDIA RTX Virtual Workstation.

Table 1.1 NVIDIA RTX Workstation GPUs

	V100S/V100	A40	RTX 8000	RTX 6000	T4	P6
GPUs / Board (Architecture)	Volta	Ampere	Turing	Turing	Turing	Pascal
Memory Size	32GB/16GB HBM2	48 GB GDDR6	48 GB GDDR6	24 GB GDDR6	16 GB GDDR6	16 GB GDDR5
vGPU Profiles	1GB, 2GB, 4GB, 8GB, 16GB, 32GB	1GB, 2GB, 3GB, 4GB, 6GB, 8GB, 12GB, 16GB, 24GB, 48GB	1GB, 2GB, 3GB, 4GB, 6GB, 8GB, 12GB, 16GB, 24GB, 48GB	1GB, 2GB, 3GB, 4GB, 6GB, 8GB, 12GB, 24GB	1GB, 2GB, 4GB, 8GB, 16GB	1GB, 2GB, 4GB, 8GB, 16GB
Form Factor	PCIe 3.0 Dual Slot and SXM2	PCIe 4.0 Dual Slot	PCIe 3.0 Dual Slot	PCIe 3.0 Dual Slot	PCIe 3.0 Single Slot	MXM (blade servers)
Power	250/300W	300W	250 W	250 W	70 W	90 W
Thermal	Passive	Passive	Passive	Passive	Passive	Bare Board
Use Case	Ultra-high-end rendering, simulation, 3D design with RTX vWS; ideal upgrade path for V100; compute-intensive AI, deep learning, HPS workloads with vCS.	Mid-range to High-end 3D design and creative workflows. AI with NVIDIA vCS; upgrade path for RTX8000, RTX6000	High-end rendering 3D design and creative workflows. AI and data science with vCS.	Mid-range to High-end 3D design and creative workflows.	Entry-level to mid-range 3D design and engineer workflows. High-density, low power GPU acceleration for knowledge workers.	For customers requiring GPUs in a blade server form factor; ideal upgrade path for M6.

For more information regarding how to select the right GPU for your virtualized workload, refer to the NVIDIA Virtual GPU Positioning Technical Brief.

NOTE: It is important to resize your environment when switching from Maxwell GPUs to newer GPUs like Pascal, Turing, and Ampere GPUs. For example, the NVIDIA T4 leverages ECC memory which 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. 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 to gain a better understanding of how your users work and how much GPU resource they really need. This includes analyzing the utilization of all resources, both physical and virtual, as well as gathering subjective feedback in order to optimize the configuration to meet the performance requirements of your users and for best scale. Benchmark examples like those highlighted in later sections within this guide can be used to help size a deployment, but they have some limitations.

Since user behavior varies and is a critical factor in determining the best GPU and profile size, sizing recommendations are typically made for three user types and are segmented as either light, medium or heavy based on type of workflow and the size of the model/data they are working with. Users with more advanced graphics requirements and using larger data sets are categorized as heavy users, for example. Light and medium users require less graphics and typically work with smaller model sizes. The following sections cover topics and methodology which should be considered for sizing.

2.1 vGPU Profiles

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

vGPU types have a fixed amount of frame buffer, number of supported display heads, and maximum resolutions. They are grouped into different series according to the different classes of workload for which they are optimized. Each series is identified by the last letter of the vGPU type name. The Q-series requires a NVIDIA RTX vWS license.

Table 2.1 NVIDIA vGPU Profiles

Profile	Optimal Workload
Q-profile	Virtual workstations for creative and technical professionals who require the performance and features of NVIDIA RTX Enterprise drivers
C-profile	Compute-intensive server workloads, such as artificial intelligence (AI), deep learning, or high-performance computing (HPC)
B-profile	Virtual desktops for business professionals and knowledge workers

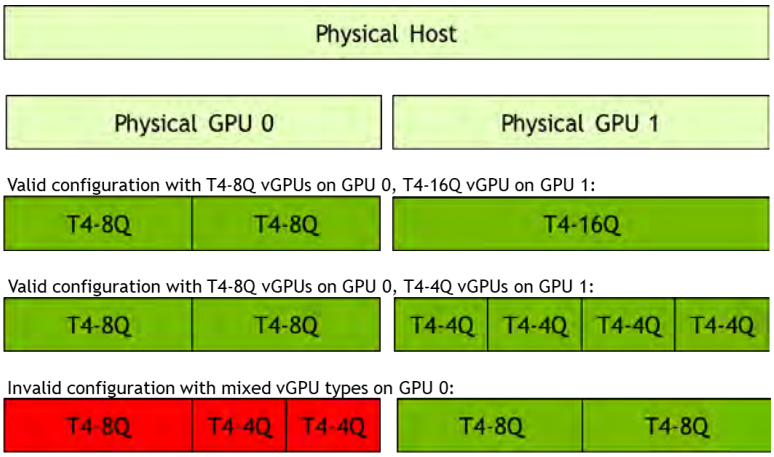
A-profile	App streaming or session-based solutions for virtual applications users

For more information regarding vGPU types, please refer the vGPU software user guide.

It is important to consider which vGPU profile will be used within a deployment since this will ultimately determine how many vGPU backed VMs can be deployed. All VMs using the shared GPU resource must be assigned the same fractionalized vGPU profile. Meaning, you cannot mix vGPU profiles on a single GPU using vGPU software.

In the image below, the right side illustrates valid configurations in green, where VMs are sharing a single GPU resource (GPU 1) on a T4 GPU and all VM's are assigned homogenous profiles, such as 8GB, 4GB, or 16GB Q profiles. Since there are two GPUs installed in the server, the other T4 (GPU 0) can be partitioned/fractionalized differently than GPU 1. An invalid configuration is shown in red, where a single GPU is being shared using 8Q, and 4Q profiles. Heterogenous profiles are not supported on vGPU, and VMs will not successfully power on.

Figure 2.1 Example vGPU Configurations



2.2 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 allow for "over-committing" or 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. In general, the oversubscribing ratio can have a dramatic impact on the performance and scalability of your NVIDIA RTX vWS implementation. In general, utilizing a 2:1 CPU oversubscription ratio can be a starting point. Actual oversubscription ratios may vary depending on your application and workflow.

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 will 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 accurately represent 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 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

Figure 3.1 GPU Profiler

3.2 NVIDIA System Management Interface (nvidia-smi)

The built in NVIDIA vGPU Manager provides extensive monitoring features to allow 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 used for providing RTX vWS VMs, 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 allows 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 VDI delivery.

Within a VDI environment, there are two tiers of metrics which can be captured: server level and VM level. Each tier has their own performance metrics, and all must be validated to ensure optimal performance and scalability.

4.1 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 which are useful during a POC or 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. A good rule of thumb to follow is that a VM's framebuffer usage should not exceed **90%** for a short time or average over **70%**. If high utilization is noted, then the vGPU backed VM is more prone to produce suboptimal user experience with potentially degraded performance and crashing. Since users interact and work differently within software applications, we recommend performing your own POC with your workload to determine framebuffer thresholds within your environment.

4.1.2 vCPU Usage

When using NVIDIA RTX vWS, 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 be used 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 Kepler generation GPU, 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. 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 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.

4.2.2 GPU Utilization

NVIDIA System Management Interface (nvidia-smi) is used for monitoring GPU Utilization rates, which report how busy each GPU is over time. It can be used to determine how much vGPU backed VMs are using the NVIDIA GPUs in the host server.

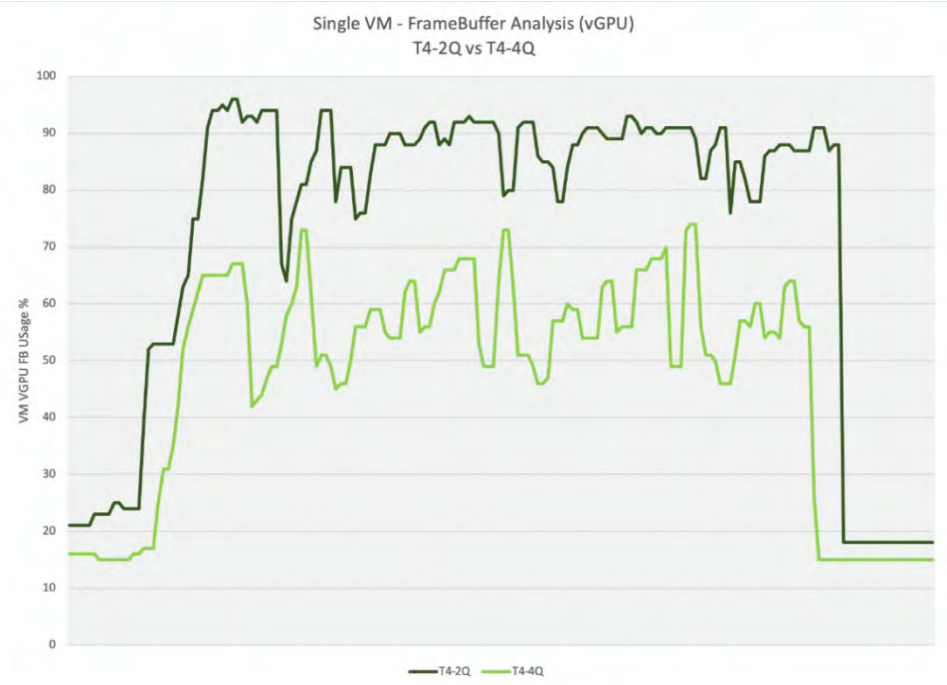
Chapter 5. Performance Analysis

5.1 Single VM Testing FB Analysis

Closely analyze the GPU framebuffer on the VDI VM to ensure correct sizing. As mentioned in the previous section, a good rule of thumb to follow is that a VM's framebuffer usage should not exceed **90%** for a short time or average over **70%**. If high utilization is noted, then the vGPU backed VM is more prone to produce suboptimal user experience with potentially degraded performance and crashing.

The graph below illustrates the vGPU FB usage within a VM when using a 2Q vGPU profile compared to 4Q profile. In this example, the benchmark was Esri ArcGIS Pro which is a professional geospatial software application and spatial navigating multi-patch 3D data. 2Q VM's reported longer rendering times and staggering software, while the 4Q VM maintained a rich and fluid end user experience with performant render times.

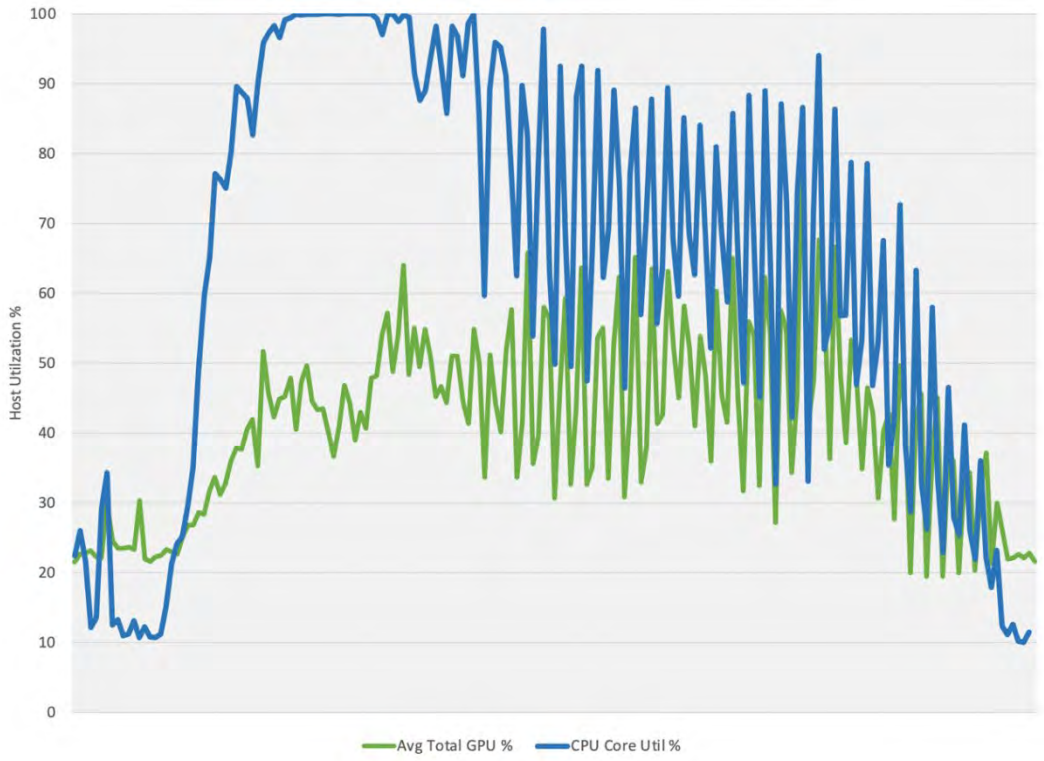
Figure 5.1 vGPU Framebuffer Usage within a VM



5.2 Host Utilization Analysis

Analyzing Host resource metrics to identify potential bottlenecks when multiple VMs are executing workloads is imperative for providing quality user experience. The most successful deployments are those that balance user density (scalability) with quality user experience. User experience will suffer when server resources are over utilized. The following chart illustrates the host utilization rates when a benchmark test is scaled across multiple VMs.

Figure 5.2 Host Utilization Rates Across Multiple VMs



GPU utilization rates illustrated in Figure 5.2 indicates there is not a GPU bottleneck. This means the server has plenty of head room within the GPU compute engine. GPU Util time is being reported by averaging utilization across the six T4 GPUs in the server. While GPU headroom is maintained throughout the duration of the test, CPU resources have become depleted, therefore VDI performance and user experience is negatively affected.

Choosing the correct server CPU for virtualization and proper configuration can have a direct effect on scalability even when a virtual GPU is present. Processor resources are often hyperthreaded and overprovisioned to a certain degree. In terms of CPU specs, you should evaluate the number of cores and clock speed. For NVIDIA RTX vWS choose higher clock speeds over higher core counts.

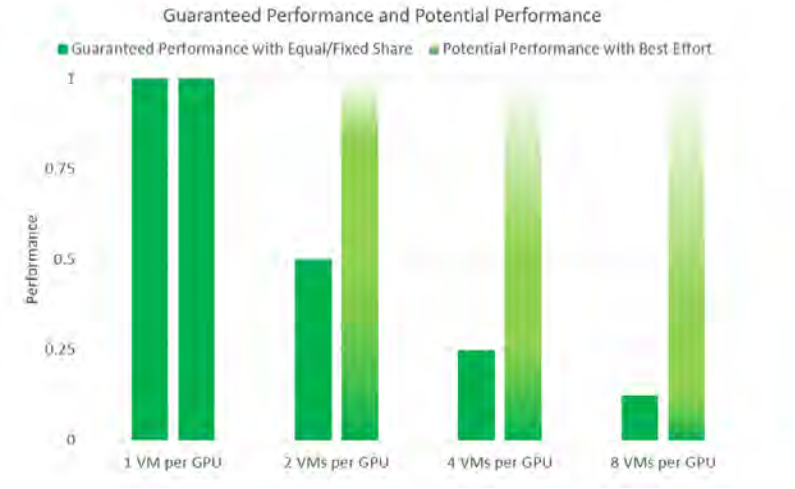
Chapter 6. Example VDI Deployment Configurations

Application-specific sizing guides utilize a combination of benchmark results and typical user configurations. These recommended best practices are based upon SPECviewperf benchmark testing and are optimized for both performance and user density. NVIDIA T4 is recommended for both light and medium users. For heavy users, the RTX 6000 or RTX 8000 is recommended. We also recommend that a larger profile size be used, 8Q for light users, 16Q for medium users and 24Q for heavy users. As a result, fewer users can be supported on each server. If only performance is important, it is recommended that the fixed share scheduler is utilized. Most customer deployments typically select the best effort GPU scheduler policy to achieve better utilization of the GPU, which usually results in supporting more users per server and better TCO per user. It is important to keep scheduling policy in mind when comparing the two options to one another.

For more on the GPU scheduling options, refer to Understanding the GPU Scheduler below.

The configuration for “performance-only” is based on running SPECviewperf across all virtual machines since it shows the impact of a peak workload on all resources of the server, including CPU, memory, GPU, and network, to best architect the solution. Tests are simultaneously executed on all virtual machines with no pauses or idle time. This workflow is not typical in a true production environment but provides a methodology for assessing dedicated performance during these worst-case scenarios.

Figure 6.1 Comparison of VMs Per GPU performance Utilization Based on Dedicated Performance vs Best Effort Configs



The following tables summarize the recommended configurations based on benchmarking data and customer best practices. These consider the performance requirements for different user types as well as optimizing for scale, or user density, on the server to achieve the best total cost of ownership.

DEDICATED PERFORMANCE

12 Users per Server T4-8Q 4vCPU 8GB RAM	6 Users per Server T4-16Q, RTX6000-16Q, RTX8000-16Q 8vCPU 16GB RAM	3-4 Users per Server RTX 6000-24Q or RTX8000-24Q 12vCPU 48GB RAM
Light User	**Medium User**	**Heavy User**

TYPICAL CUSTOMER DEPLOYMENT

16-24 Users per Server T4-2Q 4vCPU 8-16GB RAM	12-18 Users per Server T4-2Q/4Q,RTX6000-4Q,RTX8000-8Q 8vCPU 16-32GB RAM	6-9 Users per Server RTX 6000-8Q, RTX6000-12Q 12vCPU+ >96GB RAM
Light User	**Medium User**	**Heavy User**

The NVIDIA specific and third-party industry tools mentioned within this guide were used to capture VM and server level metrics to validate the optimal performance and scalability based upon

benchmark data. It is highly recommended that you run a proof of concept for each deployment type in order to validate using objective measurements and subjective feedback from your end users.

Chapter 7. Deployment Best Practices

7.1 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 data centers, 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 considerations such as 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. For further information regarding installation and server configuration steps, please refer to the NVIDIA vGPU on VMware vSphere Deployment Guide.

7.2 Run a Proof of Concept

The most successful deployments are those that balance user density (scalability) with quality user experience. This is achieved when NVIDIA RTX vWS 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

7.3 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.

7.4 Understand Your Users & Applications

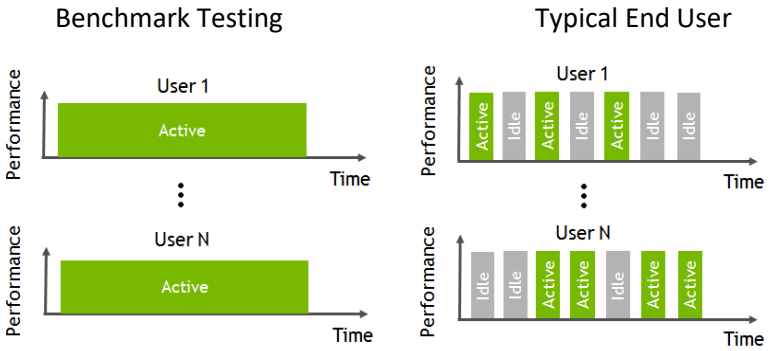
Another benefit of performing a POC prior to deployment is that it enables more accurate categorization of user behavior and GPU requirements for each virtual application. Customers often segment their end users into user types for each application and bundle similar user types on a host. Light users 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 a larger GPU like the RTX 8000. Work with your application ISV and NVIDIA representative to help you determine the correct license(s) and NVIDIA GPUs for your deployment needs.

7.5 Use Benchmark Testing

Benchmarks like SPECviewperf can be used to help size a deployment but they have some limitations. The 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 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.

Figure 7.1 Comparison of benchmarking versus typical end user



7.6 Understanding the GPU Scheduler

NVIDIA RTX vWS 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. The fixed share scheduling policies guarantee equal GPU performance across all vGPUs sharing the same physical GPU. Dedicated quality of service simplifies a POC since it allows the use of common benchmarks used to measure physical workstation performance such as SPECviewperf, to compare the performance with current physical or virtual workstations.
- **Best effort scheduling** provides consistent performance at a higher scale and therefore reduces the TCO per user. The best effort scheduler leverages a round-robin scheduling algorithm which shares GPU resources based on actual demand which results in optimal utilization of resources. This results in consistent performance with optimized user density. The best effort scheduling policy best utilizes the GPU during idle and not fully utilized times, allowing for optimized density and a good QoS.
- **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.

Organizations typically leverage the best effort GPU scheduler policy for their deployment to achieve better utilization of the GPU, which usually results in supporting more users per server with a lower quality of service (QoS) and better TCO per user.

Chapter 8. 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 allow administrators and IT staff to ensure their deployment is optimized for each user. Due to applications being used in different ways, we recommend performing your own POC with your workload.

8.1 Process for Success

Successful NVIDIA RTX vWS deployments follow these steps to deliver a rich accelerated end user experience.

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

8.2 Virtualize Any Application with an Amazing User Experience

From stunning industrial design to advanced special effects to complex scientific visualization, NVIDIA RTX is the world’s preeminent visual computing platform. By combining NVIDIA RTX Virtual Workstation (RTX vWS) software with NVIDIA GPUs, you can deliver the most powerful virtual workstation from the data center or cloud to any device. Millions of creative and technical professionals can access the most demanding applications from anywhere and tackle larger datasets, all while meeting the need for greater security. To see how you can virtualize any application with an amazing end user experience using NVIDIA RTX vWS software, try it for free.

Appendix A. NVIDIA Test Environment

Table A.1 Virtual Machine (VM) Configuration

VM Configuration	
Operating system	Windows 10 RS4
vCPUs	8 (single socket)
vMemory	16 GB
Internal Storage	100 GB
vGPU Driver Version	NVIDIA Virtual GPU Software 11.1
vGPU Software Edition	Quadro vDWS
vSync	Default
Frame Rate Limiter	Disabled
VDA Version	7.9
Direct Connect Version	7.9
Number of Screens	1
Screen Resolution	1920 x 1080

Table A.2 Hypervisor Configuration

Hypervisor Configuration	
Hypervisor	VMware ESXi, 6.7.0, 15160138
Remote Stack	VMware Horizon 7.9 with PCoIP
GPU Allocation Policy	Depth-First
vGPU Manager Version	NVIDIA Virtual GPU Software 11.1

Table A.3 Server Configuration

Server Configuration	
CPU	2 x Intel Xeon Gold 6154 CPUs (3.0 GHz)
Memory	512 GB
Hyperthreading	Enabled
Power Setting	High Performance
Storage Type	All-Flash SAN (iSCSI)
Network	10 GbE

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051

http://www.nvidia.com