NVIDIA RTX Virtual Workstation

Sizing Guide

Document History

nv-quadro-vgpu-vdws-generic-sizing-guide-v1-01142021.docx

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
04	Sept 17, 2021	AFS	Positioning update and benchmarks using 13.0

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 for each user depending on many factors, including the number of applications being used, the types of applications, file sizes, monitor resolution, and the 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 IT departments to understand the expectations and behavior of their users and optimize their deployment for the best user density while maintaining the required performance levels. Continued monitoring is essential because user behavior can change throughout a project and personnel changes can take place within the organization. Once light graphics users can become heavy graphics users when they change teams or are assigned a different task. Applications also have ever-increasing graphical requirements. Management and monitoring tools allow administrators and IT staff to ensure their deployment is optimized. Through this document, you will understand these tools and the critical 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 workstations 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, RTX 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, NVIDIA CUDA® and OpenCL, support for higher resolution displays, and larger GPU profile sizes.

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

1.2 Why NVIDIA vGPU?

NVIDIA RTX vWS software is based on NVIDIA virtual GPU (vGPU) technology and includes the NVIDIA RTX Enterprise driver 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. RTX vWS uses the same NVIDIA drivers that are deployed on non-virtualized operating systems. By doing so, NVIDIA RTX vWS provides VMs with high-performance graphics and application compatibility and cost-effectiveness and scalability since multiple VMs can be customized to specific tasks that may demand more or less GPU compute or memory.

NVIDIA RTX Virtual Workstations benefit from all of the enhancements of NVIDIA RTX technology, including real-time ray tracing, artificial intelligence, rasterization and simulation. With RTX technology, artists realize the dream of real-time cinematic-quality rendering of photorealistic environments with perfectly accurate shadows, reflections, and refractions so that they can create amazing content faster than ever before. NVIDIA RTX also brings the power of AI to visual computing to dramatically accelerate creativity by automating repetitive tasks, enabling all-new creative assistants, and optimizing compute-intensive processes.

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

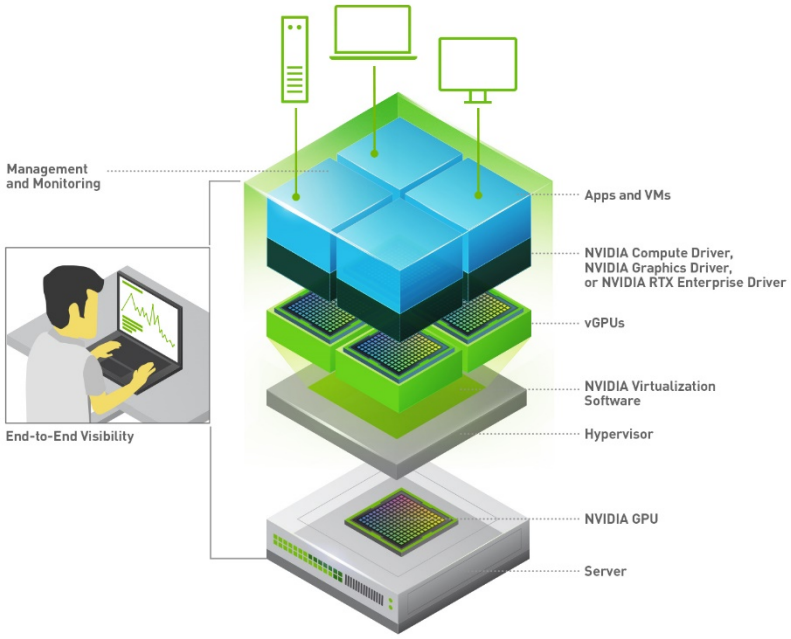
- Management and monitoring – streamline data center manageability by leveraging hypervisor-based tools.
- Live Migration – Live to migrate GPU-accelerated VMs without disruption, easing maintenance and upgrades.
- Security – Extend the benefits of server virtualization to GPU workloads.
- Multi-Tenancy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

Factors that should be considered during a 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 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



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

1.4 Recommended NVIDIA GPUs for NVIDIA RTX vWS

Table 1.1 lists the hardware specification for the most recent generation NVIDIA data center GPUs recommended for NVIDIA RTX Virtual Workstation.

Table 1.1 NVIDIA GPUs Recommended for RTX vWS

	A40	A16*
GPUs / Board (Architecture)	Ampere	Ampere
Memory Size	48 GB GDDR6	64 GB GDDR6 (4 x 16 GB per card)
vGPU Profiles	1GB, 2GB, 3GB, 4GB, 6GB, 8GB, 12GB, 16GB, 24GB, 48GB	1GB, 2GB, 4GB, 8GB, 16GB,
Form Factor	PCIe 4.0 Dual Slot Full Length Full Slot (FHFL)	PCIe 4.0 Dual Slot Full Length Full Slot (FHFL)
Power	300W	250W
Thermal	Passive	Passive
Use Case	Light to High-end 3D design and creative workflows. Flexibly runs mixed workloads for both virtual workstations and compute workloads; Upgrade path for RTX 8000, RTX 6000, T4	Entry level Virtual Workstations Upgrade path for T4 and M10

*NVIDIA A16 is recommended only for entry level virtual workstations with light weight users. A minimum 8GB (8Q) profile is recommended when deploying NVIDIA RTX Virtual Workstations with A16.

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

NOTE: It is essential to resize your environment when you switch from Maxwell GPUs to newer GPUs such as Pascal, Turing, and Ampere GPUs. For example, the NVIDIA T4 supports 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 used by vGPU is reduced. For additional information, refer to the vGPU software release notes.

Chapter 2. Sizing Methodology

It is highly recommended that a proof of concept is performed before full deployment to understand better how your users work and how much GPU resource they need. This includes analyzing the utilization of all resources, both physical and virtual, and gathering subjective feedback to optimize the configuration to meet the performance requirements of your users and for the best scale. Benchmark examples like those highlighted in later sections within this guide can 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 the kind of workflow and the size of the model/data they are working with. For example, users with more advanced graphics requirements and larger data sets are categorized as heavy users. 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 virtual machines (VMs) in the hypervisor management console using vGPU profiles. Virtual GPU profiles determine the amount of GPU frame buffer that can be allocated to your VMs. Choosing 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, several supported display heads, and maximum resolutions. They are grouped into different series according to the various classes of workload for which they are optimized. The Q-profile requires an NVIDIA RTX vWS license. The following table provides further details and lists the the other vGPU profiles available to all vGPU license levels.

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 to the vGPU software user guide.

It is essential 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. Note, that since the NVIDIA A16 has a quad-GPU board design, you can mix different profile sizes on a single A16 board.

In the image below, the right side illustrates valid configurations in green, where VMs share a single GPU resource (GPU 1) on an A40 GPU and all VM's are assigned homogenous profiles, such as 4GB, 12GB, or 24GB Q profiles. Since there are two GPUs installed in the server, the other A40 (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 24Q and 4Q profiles. Heterogenous profiles are not supported on vGPU, and VMs will not successfully power on.

Figure 2.1 Example vGPU Configurations for NVIDIA A40

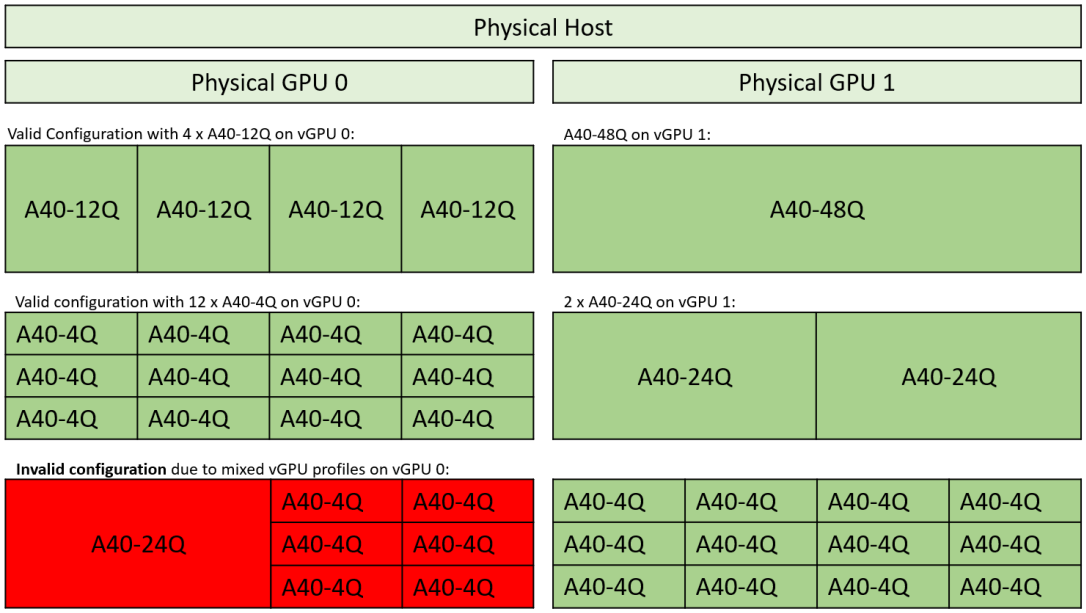
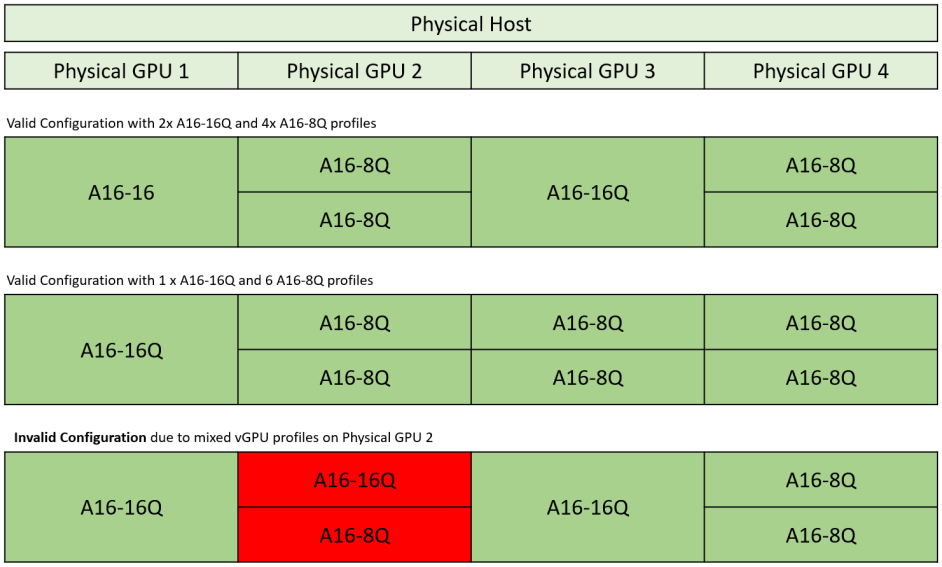


Figure 2.2 Example vGPU Configurations for NVIDIA A16



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

Several NVIDIA-specific and third-party industry tools 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 physical and virtual resources to optimize the configuration to meet the performance requirements of your users and for the best scale. These tools are helpful during your POC to ensure your test environment will accurately represent a live production environment. It is essential 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 that 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 understand better usage of the various engines of an NVIDIA vGPU. The utilization of the compute engine, the frame buffer, the encoder, and the 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 the physical GPU bottlenecks used to provide 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 physical host state information for each processor, the host's memory utilization, and 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, 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 essential to collect metrics during your POC and regularly in a production environment to ensure optimal VDI delivery.

Within a VDI environment, there are two tiers of metrics that can be captured: Server level and VM level. Each tier has its 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 great tools for understanding resource usage metrics within VMs. The following sections cover the metrics useful during a POC or monitor an existing deployment to understand potential performance bottlenecks further.

4.1.1 Framebuffer Usage

In a virtualized environment, the frame buffer is the amount of vGPU memory exposed to the guest operating system. A good rule of thumb to follow is that a VM's frame buffer usage should not exceed **90%** frequently or average over **70%**. If high utilization is noted, then the vGPU backed VM is more prone to produce a suboptimal user experience with potentially degraded performance and crashing. Since users interact and work differently within software applications, we recommend performing your POC with your workload to determine frame buffer thresholds within your environment.

4.1.2 vCPU Usage

Using NVIDIA RTX vWS, vCPU usage can be just as crucial as the VM's vGPU frame buffer usage. Since all workloads require CPU resources, vCPU usage should not bottleneck and is vital 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 fully accelerates hardware-based video decoding and encoding for several popular codecs. Complete encoding (which can be computationally complex) is offloaded from the CPU to the GPU using NVENC. A hardware-based decoder (referred to as NVDEC) provides faste 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 great tools for understanding resource usage metrics for a physical host. The following sections cover the metrics useful during a POC or monitoring an existing deployment to 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 determine how much vGPU backed VMs are using the NVIDIA GPUs in the host server.

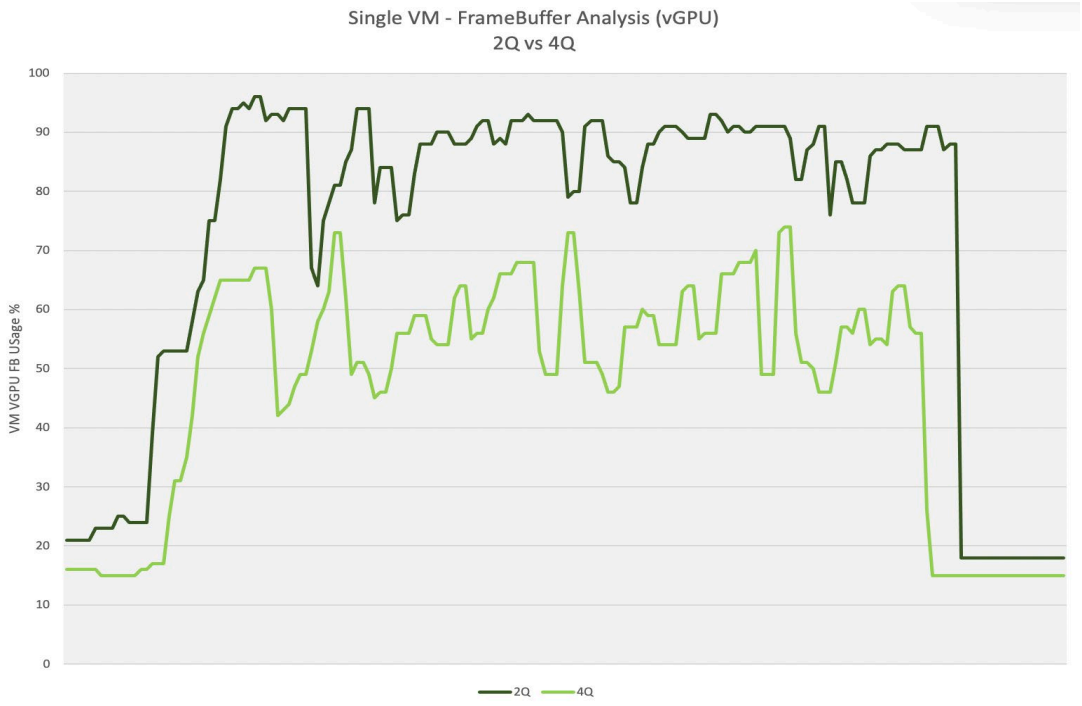
Chapter 5. Performance Analysis

5.1 Single VM Testing FB Analysis

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

The graph below illustrates the vGPU FB usage within a VM using a 2Q vGPU profile compared to a 4Q profile. In this example, the benchmark was Esri ArcGIS Pro, 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.

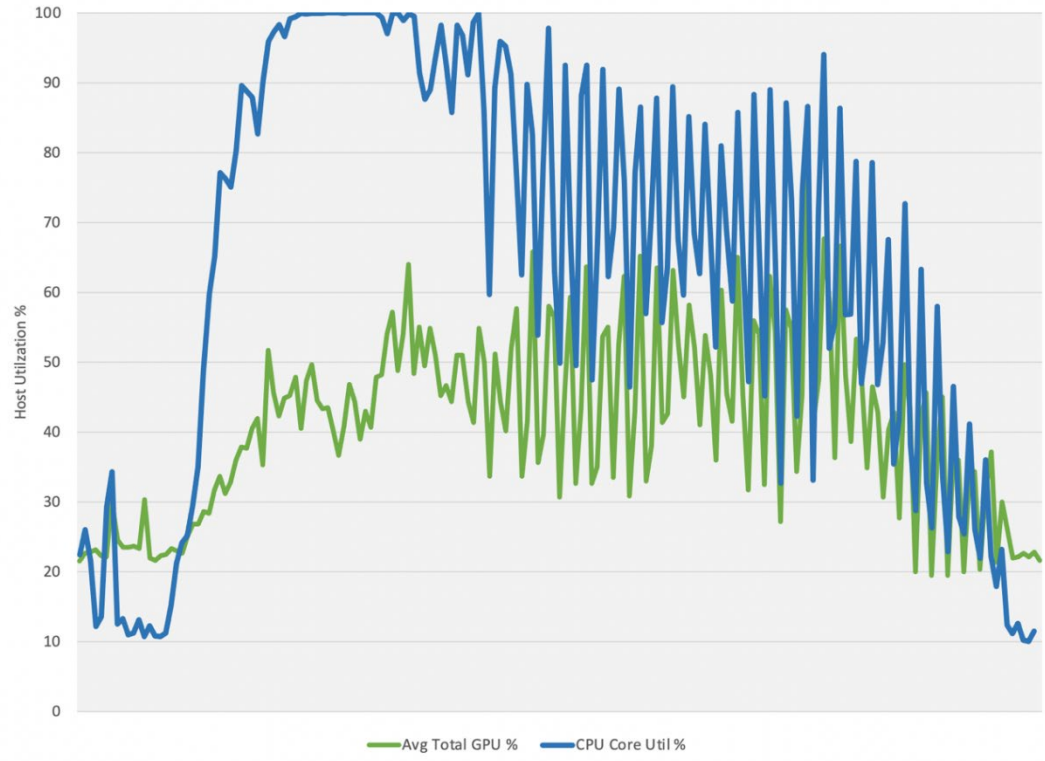
Figure 5.1 vGPU Framebuffer Usage within a VM



5.2 Host Utilization Analysis

Analyzing Host resource metrics to identify potential bottlenecks when multiple VMs execute workloads is imperative for providing a 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 indicate there is not a GPU bottleneck. This means the server has plenty of headroom within the GPU compute engine. GPU utilization time is being reported by averaging utilization across the three A40 GPUs in the server. While GPU headroom is maintained throughout the test, CPU resources have become depleted; therefore, VDI performance and user experience are negatively affected.

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

An example server configuration for NVIDIA RTX vWS is provided as an appendix item.

Chapter 6. Example VDI Deployment Configurations

Application-specific sizing utilizes a combination of benchmark results and typical user configurations. Recommendations are made and cover three common questions:

- Which NVIDIA Data Center GPU should I use for my business needs?
- How do I select the correct profile(s) for the types of users I will have?
- How many users can be supported (user density) per server?

Since user behavior varies and is a critical factor in determining the best GPU and profile size, recommendations are made for three user types and two levels of quality of service (QoS) for each user type: Dedicated Performance and Typical Customer Deployment. User types are segmented as either light, medium, or heavy-based on workflow and the size of the model/data they are working with. For example, users with more advanced graphics requirements and larger data sets are categorized as heavy users. Light and medium users require less graphics and typically work with smaller model sizes. Recommendations for each of those users within each level of service, along with the server configuration, are shown. These recommendations are meant to be a guide. The most successful customer deployments start with a proof of concept and are "tuned" throughout the lifecycle of the deployment.

If only performance is essential, it is recommended that the fixed share scheduler is utilized. We also recommend that a larger profile size be used; thus, fewer users can be supported on each server. Most customer deployments typically select the best effort GPU scheduler policy to better utilize the GPU, which usually supports 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.

The following table summarizes Dedicated Performance and Typical Customer Deployment findings:

DEDICATED PERFORMANCE

18 Users per Server	12 Users per Server	6 Users per Server
User VM Config: A40-8Q 4vCPU 8GB RAM	User VM Config: A40-12Q 8vCPU 16GB RAM	User VM Config: A40-24Q 12vCPU 32GB RAM
Light User	**Medium User**	**Heavy User**

TYPICAL CUSTOMER DEPLOYMENT

16-24 Users per Server	9-18 Users per Server	6-12 Users per Server
User VM Config: A40-4Q 4vCPU 8-16GB RAM	User VM Config: A40-4Q, A40-6Q, A40-8Q 8vCPU 16-32GB RAM	User VM Config: A40-12Q, A40-16Q, A40-24Q 12vCPU+ >96GB RAM
Light User	**Medium User**	**Heavy User**

While NVIDIA recommends the A40 for RTX vWS deployments, the A16 can be leveraged for lightweight entry level virtual workstation use cases. For best performance, it is recommended that a minimum 8GB profile be used when deploying virtual workstations on the NVIDIA A16. This would mean up to 2 users per GPU would be supported, and up to 8 users per A16 board.The A16 provides additional flexibility as it has 4 x GPUs per board, each equipped with 16 GB of memory (64 GB total). This enables IT to deploy multiple profile sizes per board, and also multiple vGPU software licenses on a single board. For example, one GPU on the A16 could support 2B profiles for vPC users, while another GPU on the same board could support 8Q profiles for RTX vWS users. It is highly recommended to conduct a POC to determine if the A16 meets the density and performance needs of your organization. For OEM considerations, please consult vGPU Certified Servers for more information.

REFERENCE SERVER LAB BUILDS

3x NVIDIA A40 GPUs 2x Intel Xeon Gold 6354 128-512 GB RAM 10GbE Network (min) Flash Based Storage	3x NVIDIA A40 GPUs 2x Intel Xeon Gold 6354 512-768+ GB RAM 10GbE Network (min) Flash Based Storage	3x NVIDIA A40 GPUs 2x Intel Xeon Gold 6354 512-768+ GB RAM 10GbE Network (min) Flash Based Storage
Light User	**Medium User**	**Heavy User**

It is important to note; the Dedicated Performance table is based upon the Equal Share scheduler and does not oversubscribe the GPU compute engine, resulting in the same GPU performance at all times. Like vCPU to physical core oversubscription, many virtual GPUs can utilize the same physical GPU compute engine. The GPU compute engine can be oversubscribed by selecting the Best Effort GPU

scheduler policy, which best uses the GPU during idle and not fully utilized times. For many customer deployments, it is not typical that 12 users will be executing rendering requests simultaneously or even to the degree which were replicated in dedicated performance testing. Therefore, selecting the best effort scheduler often results in a 2–3x oversubscription of the GPU compute engine, resulting in 2-3x the number of users. The degree to which higher scalability is achieved depends on your users' typical day-to-day activities, such as the number of meetings and the length of lunch or breaks, multi-tasking, etc. It is recommended to test and validate the appropriate GPU scheduling policy to meet the needs of your users.

The vGPU profiles listed are recommendations based upon Dedicated Performance and by first understanding the graphics performance of a workstation GPU (for example, RTX A4000). The benchmark scores of the physical workstation card were then aligned with the scores achieved for the virtual GPU. The following table summarizes these findings:

DEDICATED PERFORMANCE

User Type	Equivalent Performance Level +/-10%	Users per Server	vCPUs	vGPU Profile	vMemory	CPUs	GPUs	Memory	Storage Type
Light	P1000	18	8	A40-8Q	8GB	2x Intel Xeon Gold 6354	3 x A40	128GB	Flash-Based
Medium	RTX 3060	12	8	A40-12Q	8GB	2x Intel Xeon Gold 6354	3 x A40	128GB	Flash-Based
Heavy	RTX A4000	6	12	A40-24Q	8GB	2x Intel Xeon Gold 6354	3x A40	128GB	Flash-Based

The following example demonstrates the different numbers of users per server that can be achieved by applying different Quality of Service (QoS) thresholds through GPU scheduling policies. Choosing the Fixed Share Scheduler always guarantees a particular QoS. In this example, six users on an A40 will always experience performance similar to a workstation with an NVIDIA P1000 GPU. Using the Best Effort Scheduler, the most commonly chosen GPU scheduling option for enterprises,does not provide the same QoS level, but could allow more users to experience an NVIDIA P1000 level performance. Still, user performance will vary depending on the load from other users on the same A40 at any given time. A single user on an A40 will experience performance similar to an NVIDIA RTX A6000. Still, as density increases to 3-8 users per GPU, the performance can be similar to a workstation with a Quadro P620 card. The following example assumes sufficient frame buffer at all scales to demonstrate options on how GPU scheduling policies can impact scale.

	Dedicated Performance (Fixed Share scheduler)	Typical Customer Configuration (Best Effort Scheduler)
Users/Server Host (3 x NVIDIA A40)	18 (6 users per GPU with the performance of P1000 at all times)	16-24 (3 - 8 users per GPU with the performance of P620-A6000)

For more on the GPU scheduling options and how to configure the server, refer to NVIDIA's VMware or Citrix Hypervisor vGPU Deployment Guide.

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 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 or Citrix Hypervisor 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

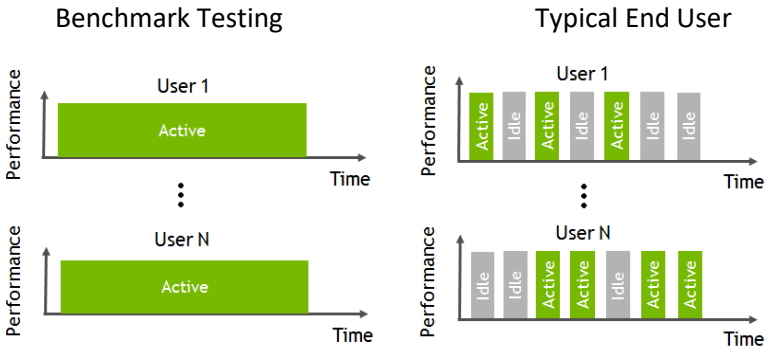
Another benefit of performing a POC prior to deployment is that it enables more accurate categorization of user behavior and GPU requirements for each virtual application. Customers often segment their end users into user types for each application and bundle similar user types on a host. Light users can be supported on a smaller vGPU profile size while heavy users require more GPU resources, and a large profile size like what can be achieved with the A40. Note, that while the NVIDIA A16 board has a total framebuffer size of 64GB, each GPU on the A16 is 16GB so the largest profile size supported on an A16 is 16Q. However, the A40 has one GPU on a board and therefore it supports up to a 48Q profile size. 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 are used, and for the best effort scheduling policy that can be leveraged 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



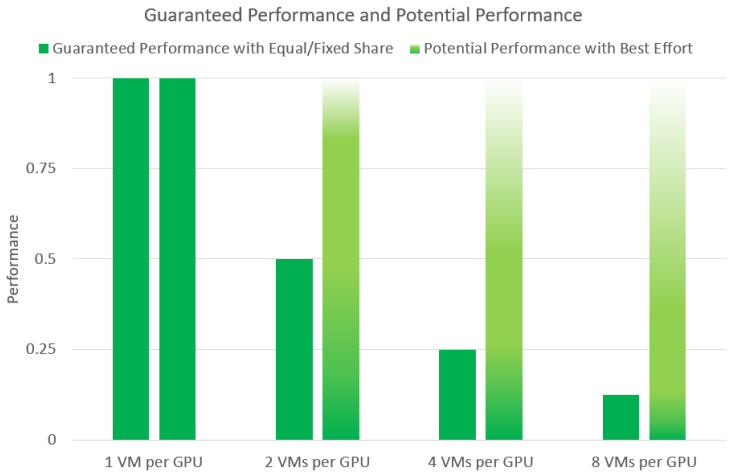
7.6 Understanding the GPU Scheduler

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.

Figure 7.2 Comparison of VMs Per GPU performance Utilization Based on Dedicated Performance vs Best Effort Configs



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 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 exceptional 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
vCPUs	8 (single socket)
vMemory	16 GB
Internal Storage	100 GB
vGPU Driver Version	NVIDIA Virtual GPU Software 13.0
vGPU Software Edition	RTX vWS
vSync	Default
Frame Rate Limiter	Disabled
Number of Screens	1
Screen Resolution	1920 x 1080

Table A.2 Hypervisor Configuration

Hypervisor Configuration	
Hypervisor	VMware ESXi, 7.0 U2
Remote Stack	VMware Horizon with PCoIP
GPU Allocation Policy	Depth-First
vGPU Manager Version	NVIDIA Virtual GPU Software 13.0

Table A.3 Server Configuration

Server Configuration	
CPU	2 x Intel Xeon Gold 6354 (3.0 GHz)
GPU	NVIDIA A40
Memory	512 GB
Hyperthreading	Enabled
Power Setting	High Performance
Storage Type	All-flash SAN
Network	10 GbE

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

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