NVIDIA QUADRO VIRTUAL DATA CENTER WORKSTATION APPLICATION SIZING GUIDE FOR SIEMENS NX

APPLICATION GUIDE

Ver 2.0

EXECUTIVE SUMMARY

This document provides insights into how to deploy NVIDIA® Quadro® Virtual Data Center Workstation (Quadro vDWS) software for SIEMENS NX users. Recommendations are based on actual customer deployments and benchmarking data and cover three common questions:

Which NVIDIA GPU should I use for my business needs?

How do I select the right 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 along with 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 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. Recommendations for each of those users within each level of service along with the server configuration are shown below.

The vGPU profiles listed in the Dedicated Performance and Typical Customer Deployment tables below were created by first understanding the graphic performance of a Quadro workstation GPU (for example, Quadro P2000). The benchmark scores of the physical workstation card were then aligned with the scores outputted for the virtual GPU. It is important to note; the Dedicated Performance table is based upon Equal Share scheduler and does not oversubscribe the GPU compute engine, resulting in the same GPU performance at all times. Similar to 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 utilizes 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 which results in 2-3x the number of users. The degree to which higher scalability is achieved is dependent on the typical day to day activities of your users, such as the number of meetings and the length of lunch or breaks, multi-tasking, etc. It is recommended to test and validate the appropriate GPU scheduling policy to meet the needs of your users.

DEDICATED PERFORMANCE

12 Users per Server	6 Users per Server	3 Users per Server
T4-8Q 4vCPU 8GB RAM	T4-16Q 8vCPU 16GB RAM	P40-24Q or RTX 6000-24Q 12vCPU 8GB RAM
Light User	**Medium User**	**Heavy User**

TYPICAL CUSTOMER DEPLOYMENT

16-24 Users per Server	12-18 Users per Server	6-9 Users per Server
T4-2Q 4vCPU 8-16GB RAM	T4-2Q/T4-4Q 8vCPU 16-32GB RAM	P40-8Q / P40-12Q or RTX 6000-8Q/RTX6000-12Q 12vCPU+ >96GB RAM
Light User	**Medium User**	**Heavy User**

REFERENCE SERVER LAB BUILDS

6x NVIDIA T4 GPUs	3x Tesla P40 GPUs	3x Quadro RTX6000 GPUs
2x Intel Xeon Gold 6154 128-512-768 GB RAM 10GbE Network Flash Based Storage	2x Intel Xeon Gold 6154 512-768+ GB RAM 10GbE Network Flash Based Storage	2x Intel Xeon Gold 6154 512-768+ GB RAM 10GbE Network Flash Based Storage
Light User **Medium User**		**Heavy User**

Table 1. Sample Siemens NX VDI Deployment Configurations

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. Beginning with a POC enables customers to understand the expectations and behavior of their users and optimize their deployment for the best user density while maintaining required performance levels. Continued maintenance is important because user behavior can change over the course of a project and as the role of an individual changes in the organization. An engineer that was once a light graphics user might become a heavy graphics user when they change teams or are assigned a different project. Management and monitoring tools enable administrators and IT staff to ensure their deployment is optimized for each user.

ABOUT SIEMENS NX

Siemens NX software is a flexible and powerful integrated solution that helps you deliver better products faster and more efficiently. NX delivers the next generation of design, simulation, and manufacturing solutions that enable companies to realize the value of the digital twin. Supporting every aspect of product development, from concept design through engineering and manufacturing, NX gives you an integrated toolset that coordinates disciplines, preserves data integrity and design intent, and streamlines the entire process.

Siemens worked closely with NVIDIA to certify the deployment of NX in the private cloud using VDI with NVIDIA Quadro vDWS software. VDI certification eliminates the need to install NX on a local client, which can help reduce IT support and maintenance costs and enable greater mobility and collaboration. This virtual workstation deployment option enhances flexibility and further expands the wide variety of platform choices available to NX customers.

ABOUT NVIDIA QUADRO VIRTUAL DATA CENTER WORKSTATION

NVIDIA virtual GPU (vGPU) software enables the delivery of graphics-rich virtual desktops and workstations accelerated by NVIDIA GPUs. NVIDIA Quadro vDWS software is based on NVIDIA virtual GPU technology and includes the Quadro graphics driver required by professional 3D applications. The Quadro vDWS license enables sharing the same NVIDIA GPU across multiple virtual machines running

any application, so every virtualized user has access to an experience that the application ISV has intended; only provided with NVIDIA Quadro.

NVIDIA Quadro is the world's preeminent visual computing platform, trusted by millions of creative and technical professionals to accelerate their workflows. With Quadro vDWS software, you can deliver the most powerful virtual workstation from the data center. This frees your 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, Quadro vDWS is the industry standard for virtualized enterprises.

To deploy an NVIDIA vGPU solution for Siemens NX , you will need NVIDIA GPUs and a Quadro vDWS software license for each user.

FINDINGS

To determine the optimal configuration of Quadro vDWS for Siemen's NX , both user performance and scalability were considered based on data from industry benchmarks as well as insights from customer best practices.

1. Benchmarking based on the industry standard SPECviewperf 13[1].
2. Documenting customer best practices using Siemens NX with Quadro vDWS

The following tables summarize the recommended configurations based on benchmarking data and customer best practices. These recommendations take into account 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. The performance of the equivalent physical Quadro workstation card was also measured and then analyzed. A 10% threshold was used to align the equivalent physical Quadro workstation card with the reported VDI performance score.

The dedicated performance table illustrates recommendations based upon the fixed share scheduler, which provides the most consistent dedicated performance at all times. However, 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 tables to one another.

For more on the GPU scheduling options, refer to Deployment Best Practices, Section 5 below:
UNDERSTANDING THE GPU SCHEDULER

DEDICATED PERFORMANCE

User Type	Equivalent Performance Level +/-10%	Users per Server	vCPUs	vGPU Profile	vMemory	CPUs	GPUs	Memory	Storage Type	Networking
Light	**Quadro P2000**	12	8	T4-8Q	8GB	2x Intel Xeon Gold 6154	6x T4	128GB	Flash-Based	10GbE
Medium	**Quadro P4000**	6	8	T4-16Q	8GB	2x Intel Xeon Gold 6154	6x T4	128GB	Flash-Based	10GbE
Heavy	**Quadro P5000**	3	12	P40-24Q	8GB	2x Intel Xeon Gold 6154	3x P40	128GB	Flash-Based	10GbE

[1] For more information about SPECviewperf 13 see www.spec.org

	Quadro RTX 6000	3	12	RTX6000-24Q	8GB	2x Intel Xeon Gold 6154	3x RTX 6000	128GB	Flash-Based	10GbE

TYPICAL CUSTOMER DEPLOYMENT

User Type	Equivalent Performance Level+/-10%	Users per Server	vCPUs	vGPU Profile	vMemory	CPUs	GPUs	Memory	Storage Type	Networking
Light	**Quadro P2000**	16 – 24	4	T4-2Q	12 – 16GB	2x Intel Xeon Gold 6154	6x T4	384GB	Flash-Based	10GbE
Medium	**Quadro P4000**	12 – 18	4	T4-4Q	16 – 32GB	2x Intel Xeon Gold 6154	6x T4	384-512GB	Flash-Based	10GbE
Heavy	**Quadro P5000**	6 - 9	8 - 12	P40-8Q P40-12Q	> 96GB	2x Intel Xeon Gold 6154	3x P40	384-1TB	Flash-Based	10GbE
	Quadro RTX 6000	6 - 9	8-12	RTX6000-8Q RTX6000-12Q	> 96GB	2x Intel Xeon Gold 6154	3x RTX 6000	384-1TB	Flash-Based	10GbE

Table 2. Server configurations for customer deployments of Siemens NX in a VDI environment.

NVIDIA T4 WITH QUADRO vDWS FOR LIGHT TO MEDIUM USERS

Quadro vDWS combined with NVIDIA T4 is recommended for virtualizing Siemens NX . The T4 GPU performance is in line with commonly used Quadro GPUs, such as the Quadro P4000, used in physical workstations for Siemens NX . When compared with the P4, the T4 offers double the frame buffer which enables professional users to work with even larger model sizes and provides about a 25% performance improvement versus the previous generation P4. The T4 adds support for new features like VP9 and H.265 which can be used for video play-back.

The T4 GPU is a single width, half height form factor and requires less power than other GPUs, allowing it to be powered via the standard PCIe bus. This results in a high-density solution accommodating up to six T4 GPUs per server. SPECviewperf 13 benchmark results show that six T4 GPUs in a server configured with two Intel Xeon Gold 6154 CPUs is a well-balanced configuration for Siemens NX .

Based on SPECviewperf 13 results, there are enough CPU resources available to host six T4 GPUs in a single 2 rack unit (RU), 2-socket server running Siemens NX on 24 virtual machines. The testing shows no performance degradation running 24 virtual machines (four per GPU) in comparison to 4 virtual machines (four per GPU) demonstrating a great balance between CPU and GPU resources.

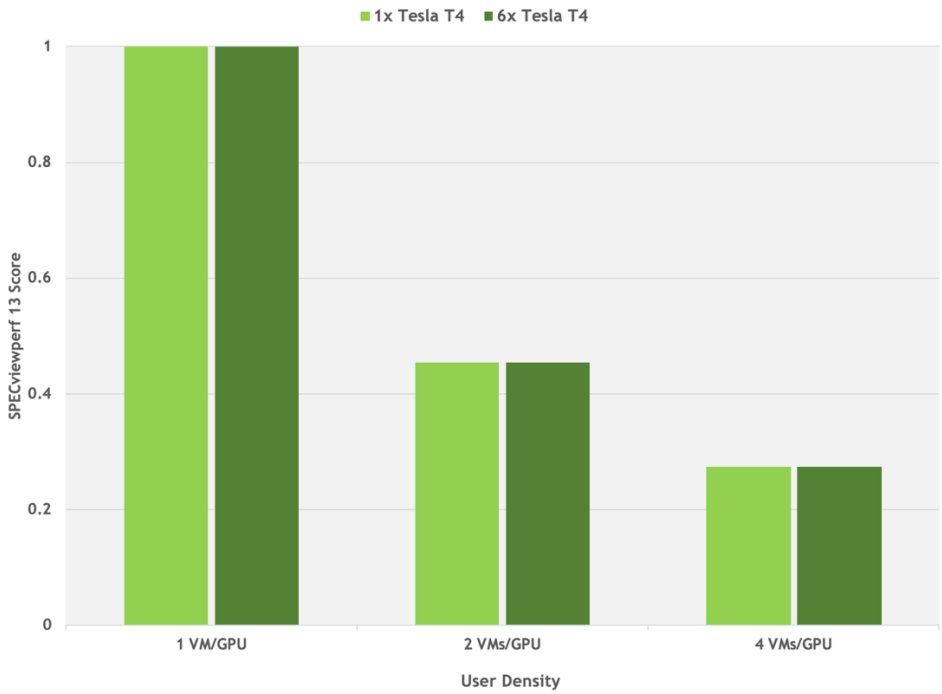

Table 3. Optimal TCO with six T4 GPUs

TESLA P40 WITH QUADRO vDWS FOR HEAVY USERS

Quadro vDWS combined with NVIDIA Tesla® P40 is recommended for heavy users that require the additional performance of a Tesla P40 over a T4. The Tesla P40 is the optimal choice for users that require more graphics and compute acceleration due to the additional performance and frame buffer of this GPU. The Tesla P40 is a dual slot card, which allows for up to three GPUs to be installed in a 2RU, 2-socket server.

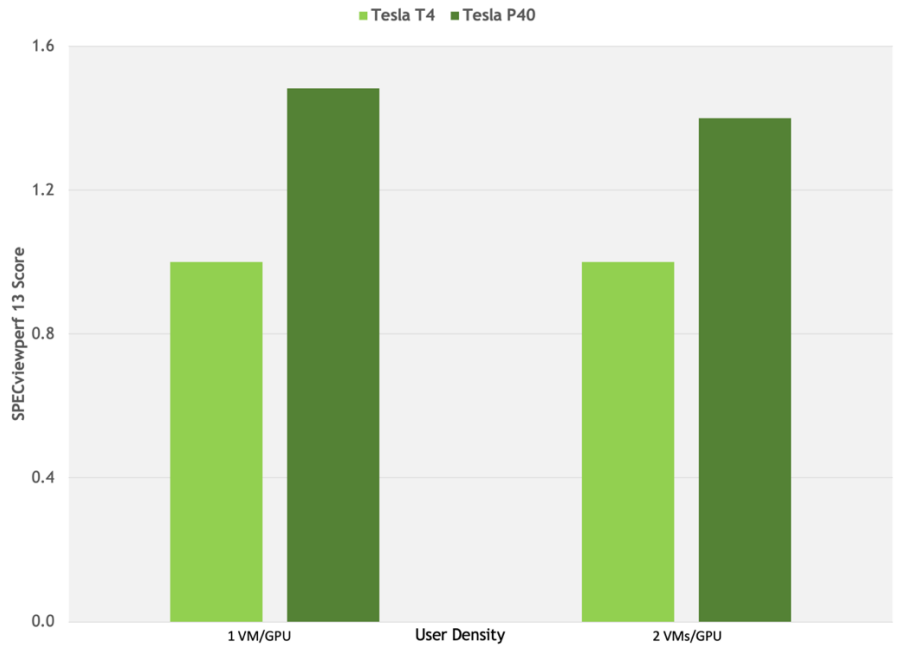

Table 4. Up to 1.5X performance achieved with Tesla P40 over T4

QUADRO RTX 6000 WITH QUADRO vDWS FOR HEAVY USERS

Quadro vDWS combined with Quadro RTX™ 6000 is recommended for heavy users that work with large assemblies with complex parts. NVIDIA Iray rendering was natively integrated in NX11, allowing Ray Traced Studio users to access NVIDIA's latest rendering technology. Quadro RTX 6000 offers Turing's RT cores which accelerate ray tracing and AI denoising. RTX 6000 can render complex models and scenes with physically accurate shadows, reflections, and refractions to empower users with instant insight. The RTX 6000 is a dual slot card, which allows for up to three GPUs to be installed in many 2RU, 2-socket servers.

SERVER RECOMMENDATION: DUAL SOCKET, 2U RACK SERVER

A two RU, 2-socket server configured with two Intel Xeon Gold 6154 processors is recommended. With a high-frequency 3.0 GHz combined with 18-cores, this CPU is well-suited for optimal performance for each end user while supporting the highest user scale, making it a cost-effective solution for Siemens NX .

SUFFICIENT SYSTEM MEMORY FOR EACH INDIVIDUAL USER

While SPECviewperf 13 performs optimally with 8 GB of system memory for each virtual machine, Siemens NX customers typically assign 16 - 32 GB of system memory to medium users for optimal performance. System memory requirements don't change with the transition to virtual workstations powered by Quadro vDWS, therefore the same amount of system memory that is used in a physical workstation should be assigned to the Quadro vDWS accelerated virtual machine.

FLASH BASED STORAGE FOR BEST PERFORMANCE

The use of flash-based storage, such as solid-state drives (SSDs) are recommended for optimal performance. Flash-based storage is the common choice with Siemens NX users using physical workstations and similar performance can be achieved in similarly configured virtual environments.

A typical configuration for non-persistent virtual machines is to use the direct attached storage (DAS) on the server in a RAID 5 or RAID 10 configuration. For persistent virtual machines, a high performing all-flash storage solution is the preferred option.

TYPICAL NETWORKING CONFIGURATION FOR QUADRO vDWS

There is no typical network configuration for in a Quadro vDWS powered virtual environment since this varies based on multiple factors including choice of hypervisor, persistent versus non-persistent virtual machines, and choice of storage solution. Most customers are using 10 GbE networking for optimal performance.

OPTIMIZING FOR DEDICATED QUALITY OF SERVICE

For comparative purposes, we also considered the requirements for a configuration optimized for performance only. This configuration option does not take into account the need to also optimize for scale, or user density. Additionally, this configuration option is based solely on performance results using the SPECviewperf benchmark.

As with the recommended best practice, which is optimized for both performance and user density, NVIDIA T4 is recommended for both light and medium users. For heavy users, the Tesla P40 is recommended. We also recommend that a larger profile size be used, a T4-8Q for light users, T4-16Q

for medium users and P40-24Q 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.

This 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. The dedicated performance data in this application sizing guide shows SPECviewperf 13 running at scale.

Tests are simultaneously executing 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.

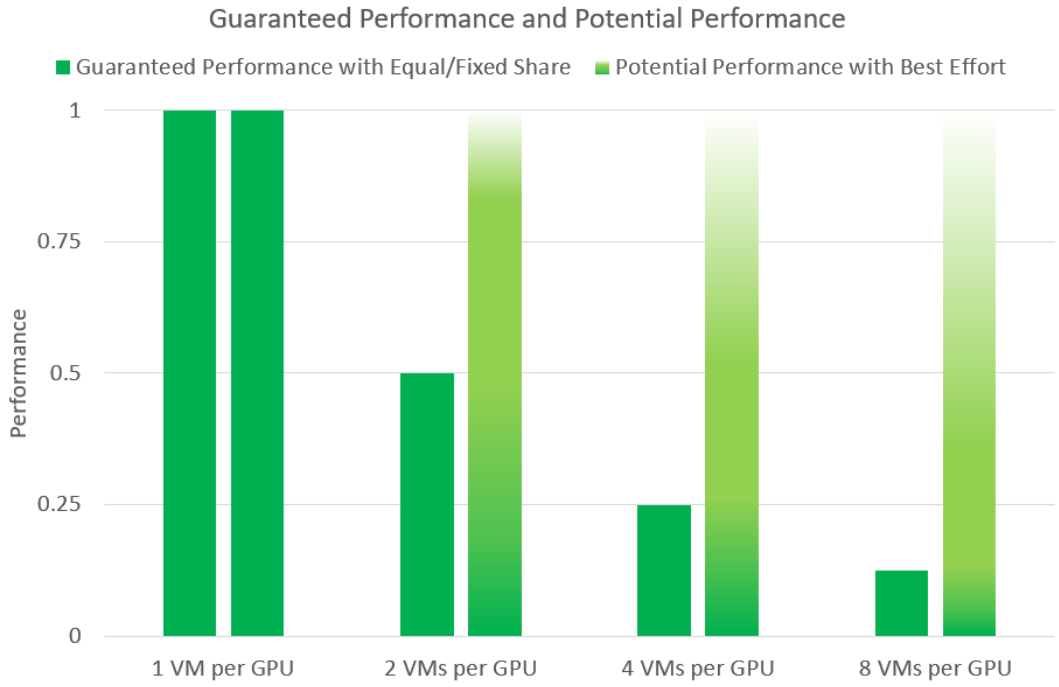

Table 5. Comparison of VMs Per GPU performance Utilization Based on Dedicated Performance vs Best Effort Configs

DEPLOYMENT BEST PRACTICES

1. RUN A PROOF OF CONCEPT

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

We highly recommend a proof of concept (POC) is run prior to doing a full deployment to provide a better understanding of how your users work and how many GPU resources they really need, analyzing the utilization of all resources, both physical and virtual. Consistently analyzing resource utilization and gathering subjective feedback allows for optimizing the configuration to meet the performance requirements of end users while optimizing the configuration for best scale.

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

Table 6. Example metrics for a successful POC

2. LEVERAGE MANAGEMENT AND MONITORING TOOLS

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

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

Virtual Machine Frame Buffer Utilization:
nvidia-smi vgpu -q -l 5 | grep -e "VM ID" -e "VM Name" -e "Total" -e "Used" -e "Free"

Virtual Machine GPU, Encoder and Decoder Utilization:
nvidia-smi vgpu -q -l 5 | grep -e "VM ID" -e "VM Name" -e "Utilization" -e "Gpu" -e "Encoder" -e "Decoder"

Physical GPU, Encoder and Decoder Utilization:
nvidia-smi -q -d UTILIZATION -l 5 | grep -v -e "Duration" -e "Number" -e "Max" -e "Min" -e "Avg" -e "Memory" -e "ENC" -e "DEC" -e "Samples"

3. UNDERSTAND YOUR USERS

Another benefit of performing a POC prior to deployment is that it enables more accurate categorization of user behavior and GPU requirements for each virtual workstation. Customers often segment their end users into user types for each application and bundle similar user types on a host. Light users can be supported on a smaller GPU and smaller profile size while heavy users require more GPU resources, a large profile size and, may be best supported on a larger GPU.

User Classification	Description
Light	• View-only and full application • Accessing individual parts or small assemblies
Medium	• View-only or full application • Accessing medium assemblies
Heavy	• Full application • Accessing larger assemblies or full model

Table 7. Common user types for Siemens NX

4. USE INDUSTRY STANDARD BENCHMARKS

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

The table 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.

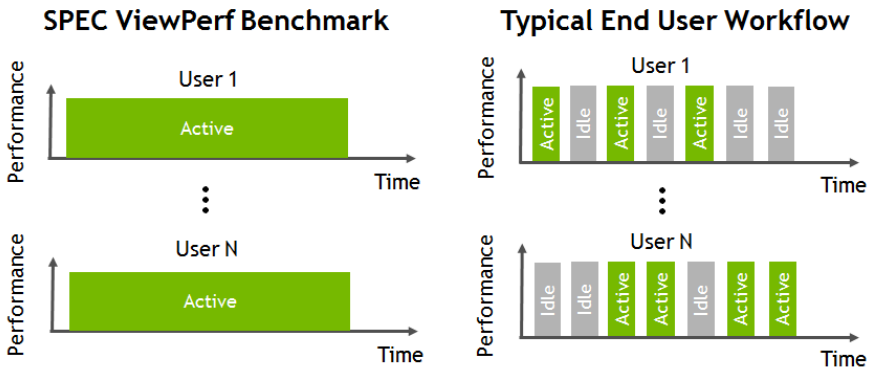

Table 8. Comparison of SPECviewperf benchmark utilization versus a typical end user workflow

NVIDIA used a custom-designed benchmarking engine to conduct vGPU testing at scale. This benchmarking engine automates the testing process from provisioning virtual machines, establishing remote connections, executing SPECviewperf, and analyzing the results across all virtual machines. Dedicated performance scores mentioned in this application guide are based on SPECviewperf 13 which was run in parallel on all virtual machines with scores averaged across three runs.

5. UNDERSTANDING THE GPU SCHEDULER

NVIDIA Quadro vDWS provides three GPU scheduling options to accommodate a variety of QoS requirements of customers.

1) **Fixed share scheduling** guarantees the same dedicated quality of service at all times.
2) **Best effort scheduling**[2] provides consistent performance at a higher scale and therefore reduces the TCO per user.
3) **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 select 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.

The below example demonstrates the different numbers of users per server that can be reached by applying different QoS thresholds via GPU Scheduling policies. Choosing the Fixed Share Scheduler always guarantees a particular QoS. In this example, two users on a T4 will always experience performance similar to a workstation with Quadro P2000 GPU. Using the Best Effort Scheduler, which is the most commonly chosen GPU scheduling option for enterprises and does not provide the same level of QoS, could allow more users to experience a Quadro P2000 level performance but user performance will vary depending on load from other users on the same T4 at any given time. A single user on a T4 will experience performance similar to a Quadro P4000 but as density increases to 3-4 users per GPU, the performance can be similar to a workstation with a Quadro P620 card. The below 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 (6 x NVIDIA T4)	12 (2 users per GPU with the performance of P2000 at all times)	16-24 (3 - 4 users per GPU with the performance of P620-P4000)

Table 9. T4 user density with Fixed Share Scheduler versus Best Effort Scheduler

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.

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.

The table below shows that when using the best effort GPU scheduling policy, performance for an individual user that shares a GPU with other users can be as good as having a dedicated GPU, if the other end users aren't executing GPU intensive tasks in parallel.

[2] Available since 2013 when NVIDIA virtual GPU technology was first introduced

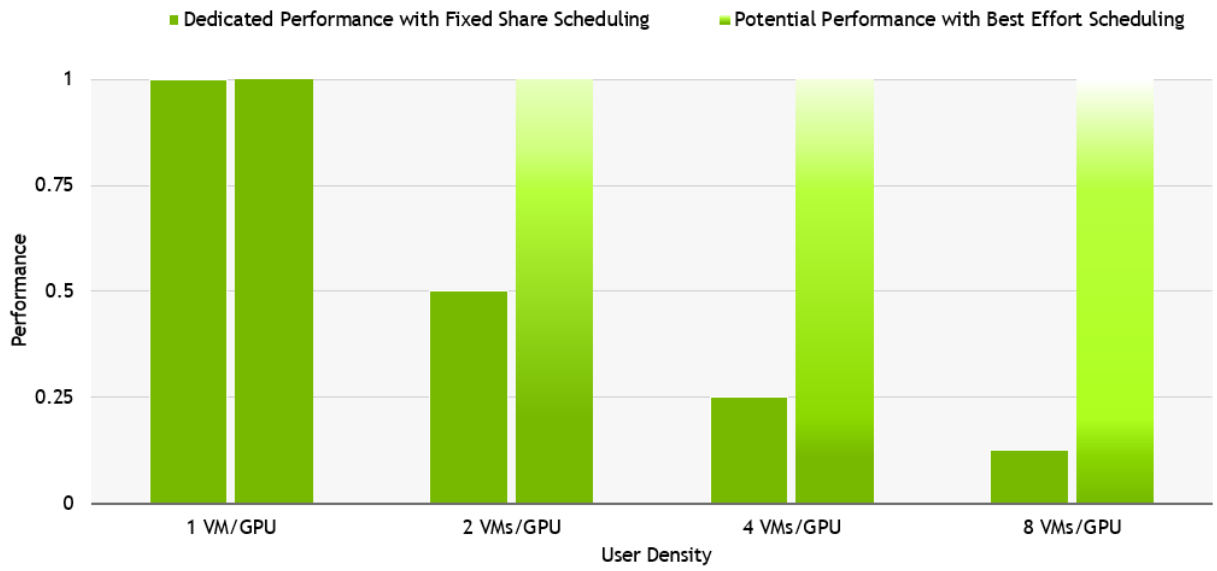

Table 10. Dedicated performance and potential performance comparison.

For details on the NVIDIA test environment used for this report, refer to the Appendix.

SUMMARY

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

To see how you can virtualize Siemens NX using Quadro vDWS software, try it for free. Or learn more about Quadro vDWS software.

APPENDIX

NVIDIA TEST ENVIRONMENT

VM Configuration	
Operating system	Windows 10 RS4
vCPUs	8
vMemory	16 GB
Internal Storage	100 GB
vGPU Driver Version	NVIDIA Virtual GPU Software 8.0 (418.98)
vGPU Software Edition	Quadro vDWS
vSync	Default
Frame Rate Limiter	Disabled
VDA Version	7.6
Direct Connect Version	7.6
Number of Screens	1
Screen Resolution	1920 x 1080

Table 11. Virtual Machine (VM) configuration details

Hypervisor Configuration	
Hypervisor	VMware vSphere 6.7.0
Remote Stack	VMware Horizon 7 with PCoIP
Remote Stack Version	7.16
VM Version	vmx-13
VM Tools	10336
GPU Allocation Policy	Depth-First
vGPU Manager Version	NVIDIA Virtual GPU Software 8.0 (418.40)

Table 12. Hypervisor configuration details

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

Table 13. Server configuration details