NVIDIA Virtual GPU Positioning

Selecting the Right GPU for Your Virtualized Workload

Technical Brief

TB-09867-001_v02 | July 2020

Table of Contents

Intent of this Technical Brief

The flexibility of the NVIDIA vGPU solution sometimes leads to the question, "How do I select the right software license and GPU combination to best meet the needs of my workloads?"

In this technical brief, you will find guidance on how to select the best virtual GPU software license and graphics processing unit (GPU) combination, based on your workload. This guidance is based on variables such as performance and performance per dollar[1]. Other factors that should be considered include things like which NVIDIA vGPU certified OEM server you've selected, which NVIDIA GPUs are supported in that platform, as well as any power and cooling constraints.

Note:

[1]Performance per dollar assumes estimated GPU street price plus NVIDIA virtual GPU software license cost with 3-year subscription divided by the number of users.

Executive Summary

It is recommended that you test your unique workloads to determine the best NVIDIA virtual GPU solution to meet your needs. However, this technical brief provides general guidance based on performance and price performance, for virtualized workloads using NVIDIA virtual GPU software.

Table 1 summarizes the recommended GPU for running a specific virtualized workload, based only on performance. For this testing, we selected a representative benchmark for each workload, described in Table 5. For the specific benchmarks run with NVIDIA virtual GPU software, NVIDIA® Quadro RTX™ 6000 and Quadro RTX 8000 GPUs provided the best performance for professional graphics and rendering workloads, while the V100S provided the best performance for artificial intelligence (AI) and high-performance computing (HPC).

In many cases, raw performance is not the only factor considered when selecting the right virtual GPU solution for your workload. Cost is often also considered. Table 2 summarizes the recommended GPU if only performance per dollar is considered. If the infrastructure will support only a knowledge worker VDI workload, the M10 GPU provides the best performance per dollar, while also providing great user density. The T4 GPU is flexible enough to run knowledge worker VDI and professional graphics workloads, and it also provides the best performance per dollar for professional graphics applications. Because the NVIDIA RTX™ platform was designed to accelerate photorealistic rendering, it is no surprise that it provides the best performance per dollar for rendering workloads. For high performance computing, the NVIDIA Volta™ architecture of the V100S has hardware to accelerate double precision (FP64) workloads, giving it the best performance and performance per dollar. It is important to note, for AI training workloads, time-to-solution is extremely important, and for that reason, costs outside of just infrastructure should be considered. As such, V100S would be recommended for this workload when considering these other cost factors.

Table 1. Best Performance GPU per Workload

Workload	Professional Graphics	Rendering	AI Deep Learning Training	AI Deep Learning Inference	High Performance Computing	Knowledge Workers
GPU	Quadro RTX 6000 / Quadro RTX 8000	Quadro RTX 6000 / Quadro RTX 8000	V100S	V100S	V100S	All GPUs perform the same

Table 2. Best Performance per Dollar GPU per Workload

Workload	Professional Graphics	Rendering	AI Deep Learning Training	AI Deep Learning Inference	High Performance Computing	Knowledge Workers
GPU	T4	Quadro RTX 6000	-	Quadro RTX 6000	V100S	M10

Introduction

The NVIDIA virtual GPU (vGPU) solution provides a flexible way to accelerate virtualized workloads – from AI to VDI. The solution includes NVIDIA virtual GPU software and NVIDIA data center GPUs. There are three unique NVIDIA virtual GPU software licenses, each priced and designed to address a specific use case:

- NVIDIA GRID Virtual PC/Virtual Applications (NVIDIA GRID) – accelerates office productivity applications, streaming video, Windows 10, RDSH, multiple and high-resolution monitors and 2D electric design automation (EDA).
- NVIDIA Quadro Virtual Data Center Workstation (Quadro vDWS) – accelerates professional design and visualization applications including Autodesk Revit, Maya, Dassault Systèmes CATIA, Solidworks, Esri ArcGIS Pro, Petrel, and more.
- NVIDIA Virtual Compute Server (vCS) – accelerates artificial intelligence (AI), deep learning (DL), data science and high-performance computing (HPC) workloads run in a virtualized environment.

Decoupling the GPU hardware and virtual GPU software options enables customers to benefit from innovative features delivered in the software at a regular cadence, without a dependency on purchasing new GPU hardware. It also provides the flexibility for IT to architect the optimal solution to meet the specific needs of users in their environment.

Selecting the Right Virtual GPU Software

Select your NVIDIA virtual GPU software license based on the workload(s) your users are running. Table 3 shows the feature differences between the NVIDIA vGPU software license options. NVIDIA GRID® vPC software is selected for knowledge worker VDI to run office productivity applications. NVIDIA® Quadro® vDWS is selected to virtualize professional visualization applications which benefit from the Quadro platform drivers and ISV certifications, support for NVIDIA® CUDA® and OpenCL, higher resolution displays, and larger profile sizes. For server virtualization to run compute workloads such as AI, data science and HPC, the NVIDIA Virtual Compute Server (vCS) license, which includes a driver that has been tested to run these compute workloads, would be selected. Customers should evaluate whether they require any of the Quadro platform graphics and visualization features. If these are not required, the vCS license could be leveraged.

Table 3. NVIDIA Virtual GPU Software Features

Configuration and Deployment	**Quadro vDWS**	**NVIDIA GRID vPC**	**NVIDIA vCS**
Windows OS Support	✓	✓	
Linux OS Support	✓	✓	✓
NVIDIA Graphics Driver	✓	✓	
NVIDIA Quadro Driver	✓		
NVIDIA Compute Driver			✓
Multi-vGPU/NVLink	✓		✓
ECC Reporting and Handling	✓		✓
Page Retirement	✓		✓
Display	**Quadro vDWS**	**NVIDIA GRID vPC**	**NVIDIA vCS**
Maximum Hardware Rendered Display	Four 5K, Two 8K	Four QHD, Two 4K, One 5K	One 4K
Maximum Resolution	7680x4302	5120x2880	4096x2160
Maximum Pixel Count	66,355,200	17,694,720	8,847,360
Advanced Professional Features	**Quadro vDWS**	**NVIDIA GRID vPC**	**NVIDIA vCS**
ISV Certifications	✓		
NVIDIA CUDA/OpenCL	✓		✓
Graphics Features and APIs	**Quadro vDWS**	**NVIDIA GRID vPC**	**NVIDIA vCS**
NVENC	✓	✓	✓
OpenGL Extensions (WebGL)	✓	✓	
Insitu Graphics/GL Support			✓
Quadro Optimizations	✓		

DirectX	✓	✓	
Vulkan Support	✓		✓
Profiles	**Quadro vDWS**	**NVIDIA GRID vPC**	**NVIDIA vCS**
Max Frame Buffer Supported	48GB	2GB	48GB
Available Profiles	0Q, 1Q, 2Q, 3Q, 4Q, 6Q, 8Q, 12Q, 16Q, 24Q, 32Q, 48Q	0B, 1B, 2B	4C, 6C, 8C, 12C, 16C, 24C, 32C, 48C

NVIDIA GPUs Recommended for Virtualization

Table 4 shows the NVIDIA GPUs recommended for virtualization workloads. The GPUs in this table are tested and supported with NVIDIA virtual GPU software. Refer to the NVIDIA virtual GPU product documentation for the full support matrix details.

Table 4. NVIDIA GPUs Recommended for Virtualization

	V100S/V100 NVLink	**Quadro RTX 8000**	**Quadro RTX 6000**	**T4**	**M10**	**P6**
GPUs/Board (Architecture)	1 (Volta)	1 (Turing)	1 (Turing)	1 (Turing)	4 (Maxwell)	1 (Pascal)
CUDA Cores	5,120	4,608	4,608	2,560	2,560 (640 per GPU)	2,048
Tensor Cores	640	576	576	320	--	--
RT Cores	--	72	72	40	--	--
Memory Size	32GB/16GB HBM2	48GB GDDR6	24GB GDDR6	16GB GDDR6	32GB GDDR5 (8GB per GPU)	16GB 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, 24GB	1GB, 2GB, 4GB, 8GB, 16GB	0.5GB, 1GB, 2GB, 4GB, 8GB	1GB, 2GB, 4GB, 8GB, 16GB
Form Factor	PCIe 3.0 Dual Slot and SXM2	PCIe 3.0 Dual Slot	PCIe 3.0 Dual Slot	PCIe 3.0 Single Slot	PCIe 3.0 Dual Slot	MXM (blade servers)

	V100S/V100 NVLink	Quadro RTX 8000	Quadro RTX 6000	T4	M10	P6
Power	250W/300W	250W	250W	70W	225W	90W
Thermal	passive	passive	passive	passive	passive	bare board
Optimized For	performance	performance	performance	performance and density	density	blade

The NVIDIA GPUs recommended for virtualization are divided into three categories:

- **Performance Optimized GPUs** are typically recommended for high-end virtual workstations running professional visualization applications, artificial intelligence, deep learning, data science or HPC workloads.
- **Density Optimized GPUs** are typically recommended for knowledge worker virtual desktop infrastructure (VDI) to run office productivity applications, streaming video and Windows 10. They are designed to maximize the number of VDI users supported in a server.
- **Blade Optimized GPUs** are designed to fit in the compact, blade server form factor and leverage a Mobile PCI Express Module (MXM) interconnect instead of the standard PCIe interconnect used for rack servers. Currently, NVIDIA offers just one MXM form factor GPU for blade servers, the P6. The P6 GPU should be selected to run any workload where a blade server form factor is preferred.

The NVIDIA T4 GPU is a compact, single slot card that consumes just 70W of power. By comparison, the NVIDIA V100S and V100, Quadro RTX 6000, Quadro RTX 8000, and M10 GPUs are dual slot PCIe cards, which consume twice as much space (two PCIe slots) inside the server and more than three times the power. This means that you can fit two NVIDIA T4 GPUs in the same space that you had fit a single V100S or V100, Quadro RTX 6000, Quadro RTX 8000, or M10 GPU.

Built on the innovative NVIDIA RTX platform, the Quadro RTX 6000 and Quadro RTX 8000 GPUs are uniquely positioned to power the most demanding professional visualization workloads. They are an integral part of the NVIDIA RTX Server solution, which can run various workloads including powerful virtual workstations. You will find that the performance of the Quadro RTX 6000 and Quadro RTX 8000 GPUs is very comparable, and the key differences between these two cards are the memory size and price. The Quadro RTX 8000 GPU should be selected over the Quadro RTX 6000 GPU if there is a requirement for larger memory to power virtual workstations that support very large animations, files, or models.

The NVIDIA V100S is the most advanced data center GPU ever built to accelerate AI, high performance computing, and data science. Customers who train or use neural networks, use computationally intensive applications, or run simulations requiring double precision accuracy (FP64 performance) should be using the V100S, which provides the best time-to-solution. V100 is available in two form factors, PCIe and SXM module. The SXM module is available with servers that support NVIDIA® NVLink®, provide the best performance and strong-scaling for hyperscale and HPC data centers running applications that scale to multiple GPUs, such as deep learning.

Selecting the Right GPU

While many organizations seek the highest performing GPU or the GPU that provides the best performance per dollar, there are other factors like performance per watt or form-factor that can be taken into consideration.

Workloads have been executed on an industry standard dual socket server with VMware vSphere 6.7 U3 and NVIDIA vGPU 10.0 using vGPU 1:1 profile unless otherwise stated. 1:1 vGPU profiles correspond to the full GPU allocated to a single virtual machine. This was chosen as the impact of scaling does not differ between GPUs[1]. See "Impact of GPU Sharing" section for more details.

Note that the comparisons should be used as general guidance when choosing GPUs based on performance or performance per dollar. All recommendations are based on the workloads listed in Table 5 which could differ from the applications being used in production.

> **Note**:
>
> [1]Assumption is that enough frame buffer is available on all vGPUs across all GPUs.

Table 5. Description of Benchmarks Used

Workload	Description	vGPU Software Edition
Professional Graphics	**SPECviewperf 13 (1920x1080)** The SPECviewperf 13 is a standard benchmark for measuring graphics performance based on professional applications. The benchmark measures the 3D graphics performance of systems running under the OpenGL and Direct X application programming interfaces.	Quadro vDWS
Rendering	**Autodesk Arnold 6.0.1.0 (SOL Dataset)** Arnold is an advanced Monte Carlo ray tracing renderer built for the demands of feature-length animation and visual effects. Arnold is used by several prominent organizations in film, television, and animation.	Quadro vDWS
AI Deep Learning Training	**ResNet-50 V1.5, TensorFlow = 19.10_py3, Batch Size: 256, Precision: Mixed** ResNet-50 TensorFlow is a model based on deep residual learning for image recognition trained with mixed precision using Tensor Cores on NVIDIA Volta and NVIDIA Turing GPUs.	NVIDIA vCS
AI Deep Learning Inference	**ResNet-50 V1.5, TensorRT 6.0.1, Batch Size = 128, 19.12-py3, Precision: Mixed** ResNet-50 NVIDIA® TensorRT™ is a model for high-performance deep learning inference.	NVIDIA vCS
High Performance Computing	**LAMMPS Atomic Fluid (Lennard Jones Dataset)** LAMMPS is a classical molecular dynamics code with a focus on materials modeling. It is an acronym for Large-scale Atomic/Molecular Massively Parallel Simulator.	NVIDIA vCS
Knowledge Worker	**NVIDIA nVector Digital Worker Workload** NVIDIA's nVector benchmarking tool that simulates the end user workflow and measures key aspects of the user experience, including end-user latency, framerate, image quality and resource utilization.	NVIDIA GRID vPC

Professional Graphics

The Quadro RTX 6000 and Quadro RTX 8000 GPUs are based on the NVIDIA Turing™ architecture, which enables major advances in efficiency and performance and is well suited for professional graphics workloads. The significantly higher power budget of the Quadro RTX 6000 and Quadro RTX 8000 cards enable them to provide higher performance than the T4. However, for those that do not require the highest performance, the T4 provides the best performance per dollar for professional graphics workloads.

Figure 1 represents SPECviewperf13 results tested on a server with Intel Xeon Gold 6154 (18C, 3.0GHz), Quadro vDWS software, VMware ESXi 6.7.0 U3, host/guest driver 440.44/441.66, VM config, Windows 10, 8 vCPU, 16GB memory.

Figure 1. Quadro vDWS SPECviewperf13 Performance

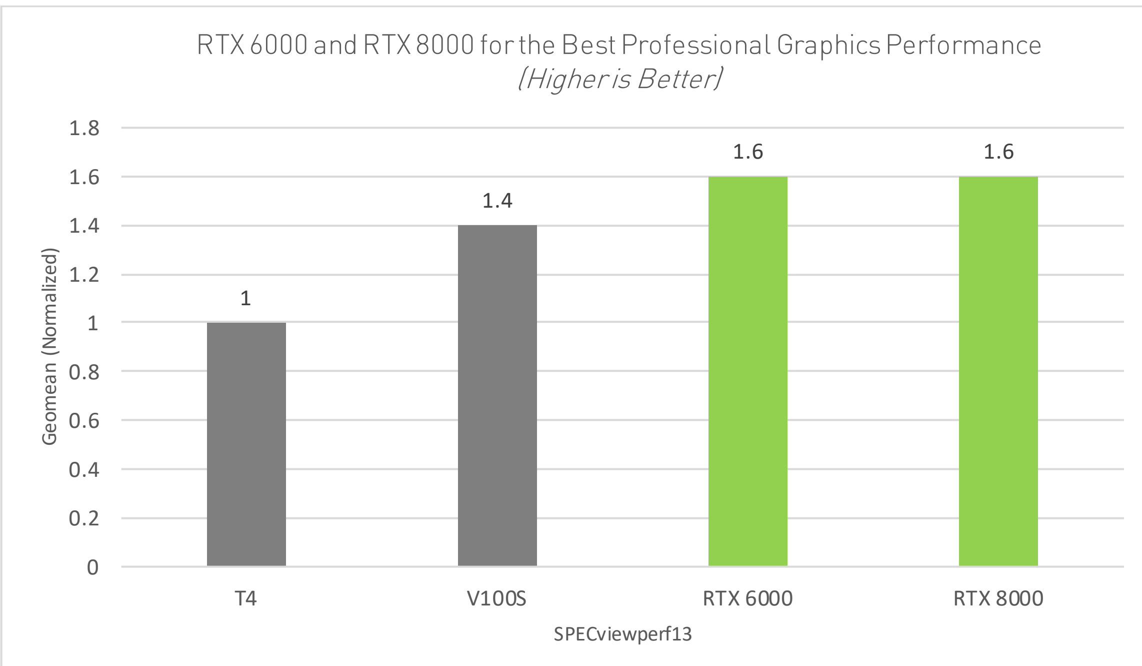
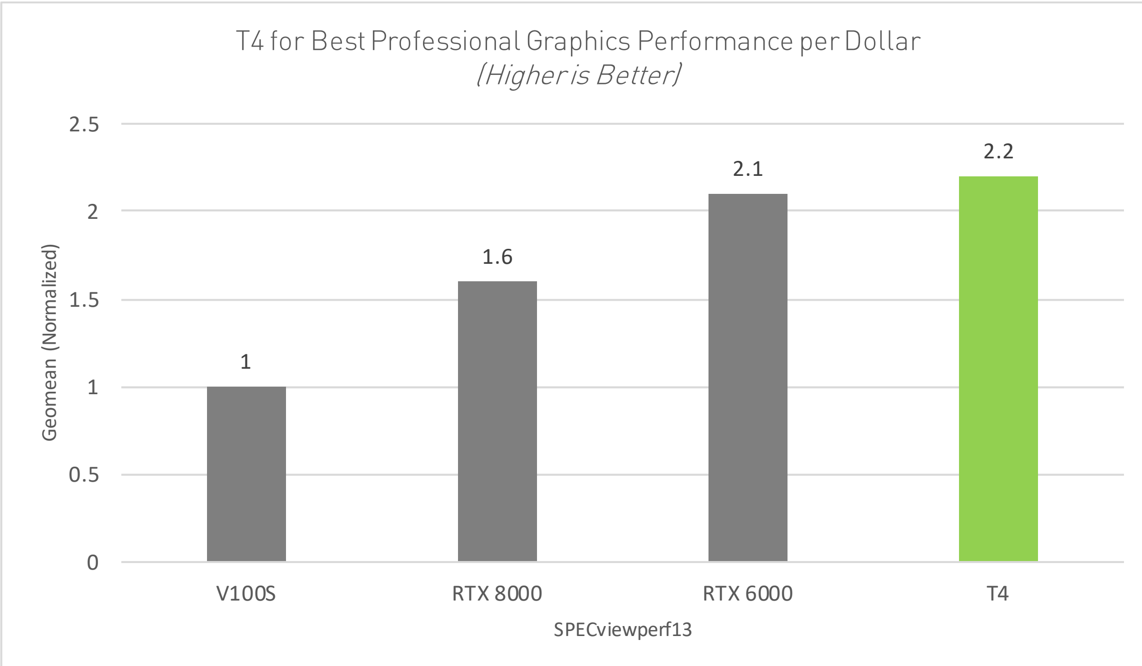


Figure 2 assumes estimated GPU street price plus NVIDIA Quadro vDWS software cost with 3-year subscription.

Figure 2. Quadro vDWS SPECviewperf13 Performance per Dollar



Rendering

Quadro RTX 6000 and Quadro RTX 8000 GPUs have RT Cores, accelerator units that are dedicated to performing ray tracing operations with extraordinary efficiency, making them the optimal choice for providing the highest rendering performance. The Quadro RTX 6000 and Quadro RTX 8000 GPUs also have a significantly higher power budget versus the T4, resulting in higher performance. The Quadro RTX 8000 would be selected over Quadro RTX 6000 if there is a requirement to support larger models or scenes. Because the scenes used in our tests did not require the additional frame buffer of the Quadro RTX 8000, you will see that the performance results between Quadro RTX 6000 and Quadro RTX 8000 were comparable for this test. However, the attractive price point of the Quadro RTX 6000 makes it ideal for those who wish to achieve the best performance per dollar.

Figure 3 represents testing on a server with Intel Xeon Gold 6154 (18C, 3.0GHz), Quadro vDWS, VMware ESXi 6.7.0 U3, host/guest driver 440.44/441.66, VM config, Windows 10, 8 vCPU, 16GB.

Figure 3. Quadro vDWS Autodesk Arnold Rendering Performance

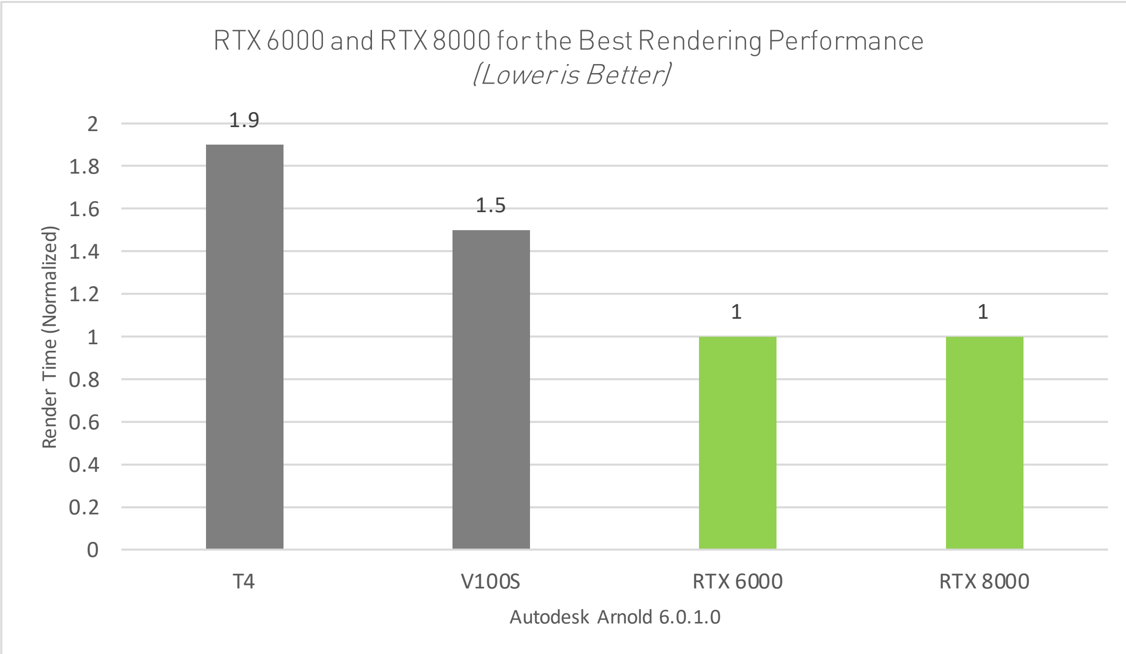
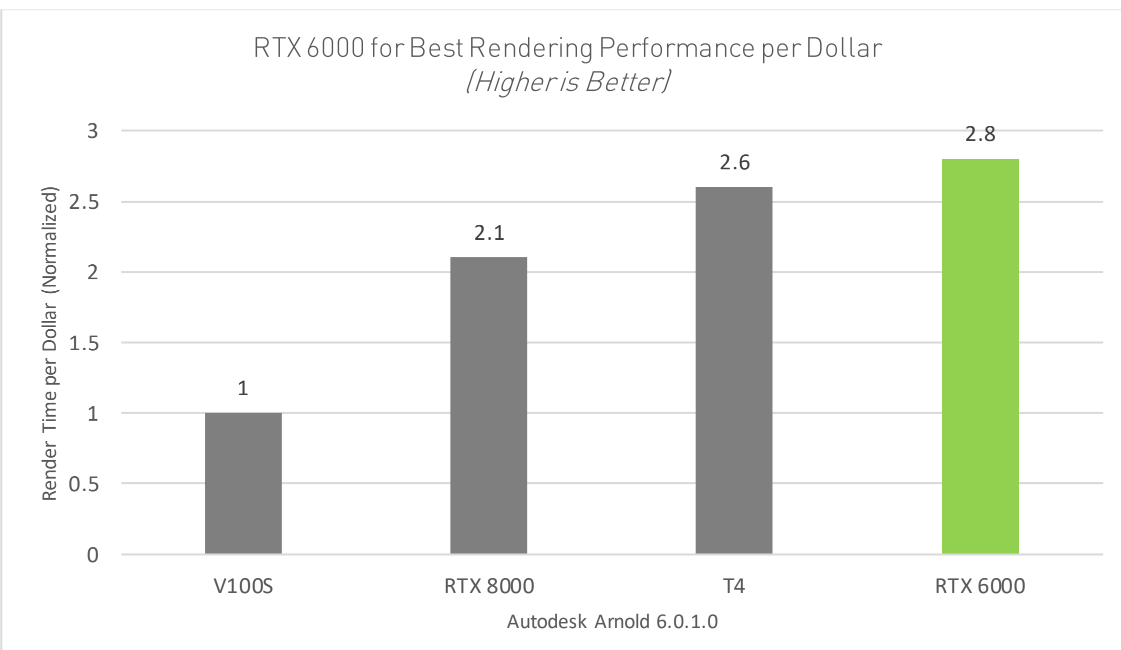


Figure 4 assumes estimated GPU street price plus NVIDIA Quadro vDWS software cost with 3-year subscription.

Figure 4. Quadro vDWS Arnold Rendering Performance per Dollar

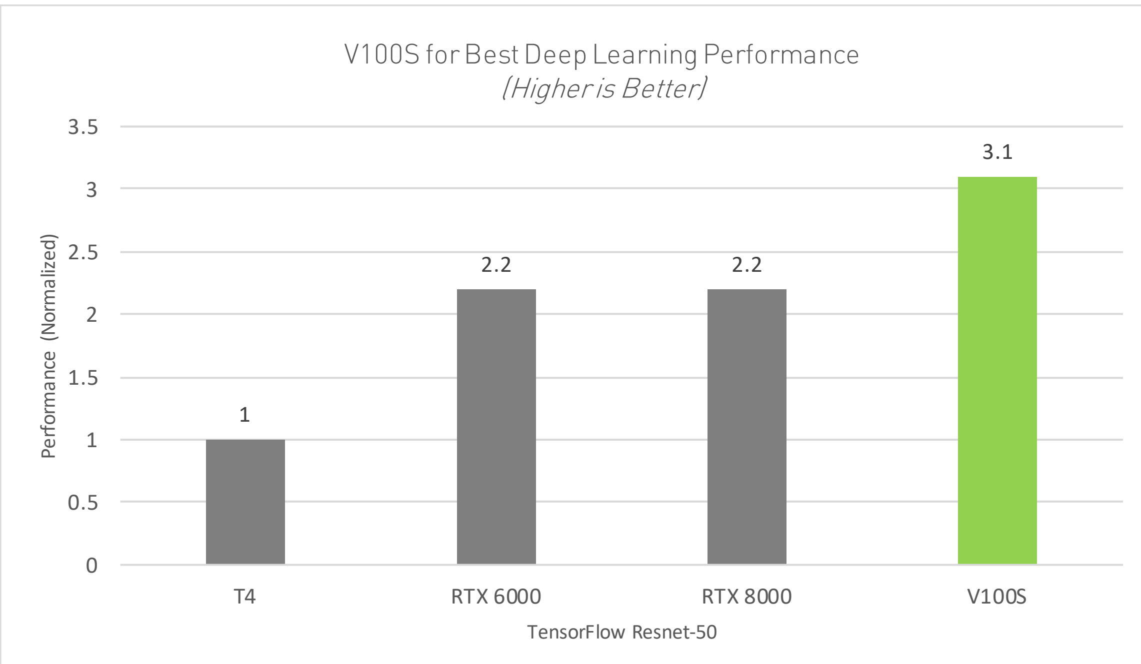


AI Deep Learning Training

V100S, based on the NVIDIA Volta architecture, is designed to bring AI to every industry. The V100S is built to accelerate AI, and it is no surprise that it provides the highest performance for deep learning training workloads. It is important to note, for deep learning training workloads, time-to-solution is extremely important. For example, the cost of having highly paid data scientists wait for results could outweigh the benefits of a slightly lower cost solution, so V100S would be recommended when considering these other cost factors.

Figure 5 represents Resnet-50 V1.5 | TensorFlow = 19.10_py3 | Batch Size: 256 | Precision: Mixed. Tested on a server with Intel Xeon Gold Skylake 6140, ESXi 6.7.0, 72 vGPU, 384 GB memory.

Figure 5. NVIDIA vCS Deep Learning Training Performance



AI Deep Learning Inference

For deep learning inference workloads, cost is often an important consideration. Therefore, the NVIDIA T4 and Quadro RTX 6000 are typically the preferred solutions. In environments where cost is the most important factor, T4 is an ideal solution. Environments which require more performance but are still looking for great performance per dollar would select Quadro RTX 6000. Environments that prioritize performance as the most important consideration would select the V100S.

Figure 6 represents Resnet-50 V1.5 | TensorRT 6.0.1 | Batch Size = 128 | 19.12-py3 | Precision: Mixed. Tested on a server with Intel Xeon Gold Skylake 6140, ESXi 6.7.0, 72 vGPU, 384 GB.

Figure 6. NVIDIA vCS Deep Learning Inference Performance

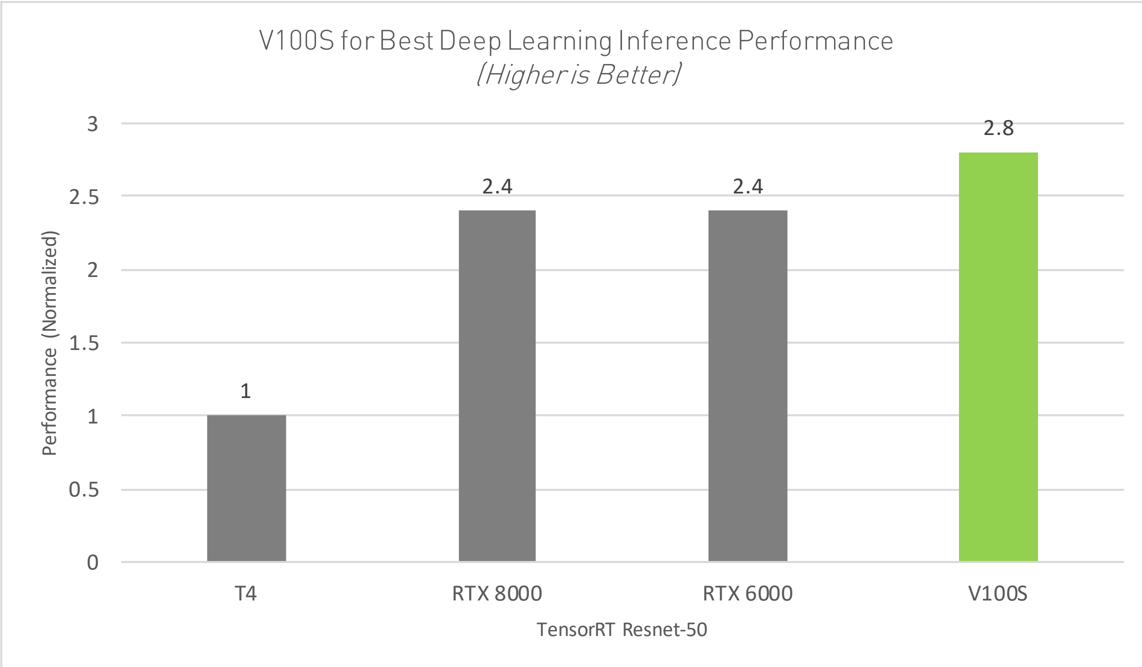
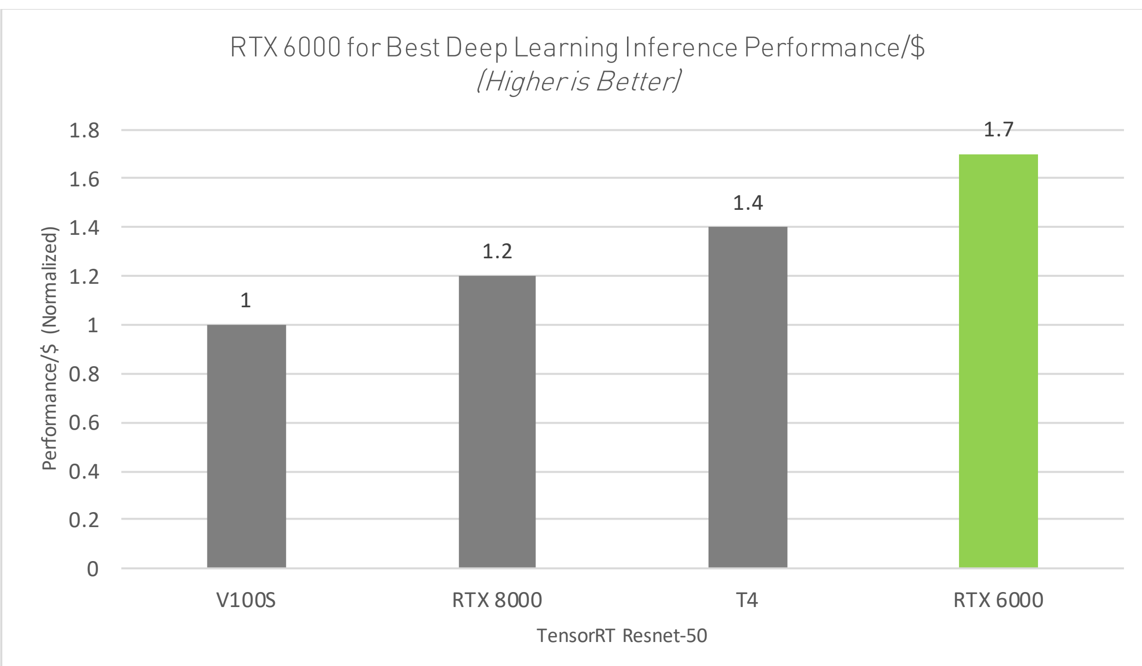


Figure 7 assumes estimated GPU street price plus NVIDIA vCS software cost.

Figure 7. NVIDIA vCS DL Inference Performance per Dollar



High Performance Computing

V100S is the best choice for scientific computing centers and higher education and research institutions running HPC workloads. The V100S provides the best performance, best performance per dollar and is optimized for double precision (FP64) workloads.

Figure 8 represents Atomic Fluid Lennard Jones. Tested on a server with Intel Xeon Gold Skylake 6140, VMware ESXi 6.7.0, 16 vCPU, 64GB memory.

Figure 8. NVIDIA vCS HPC Performance

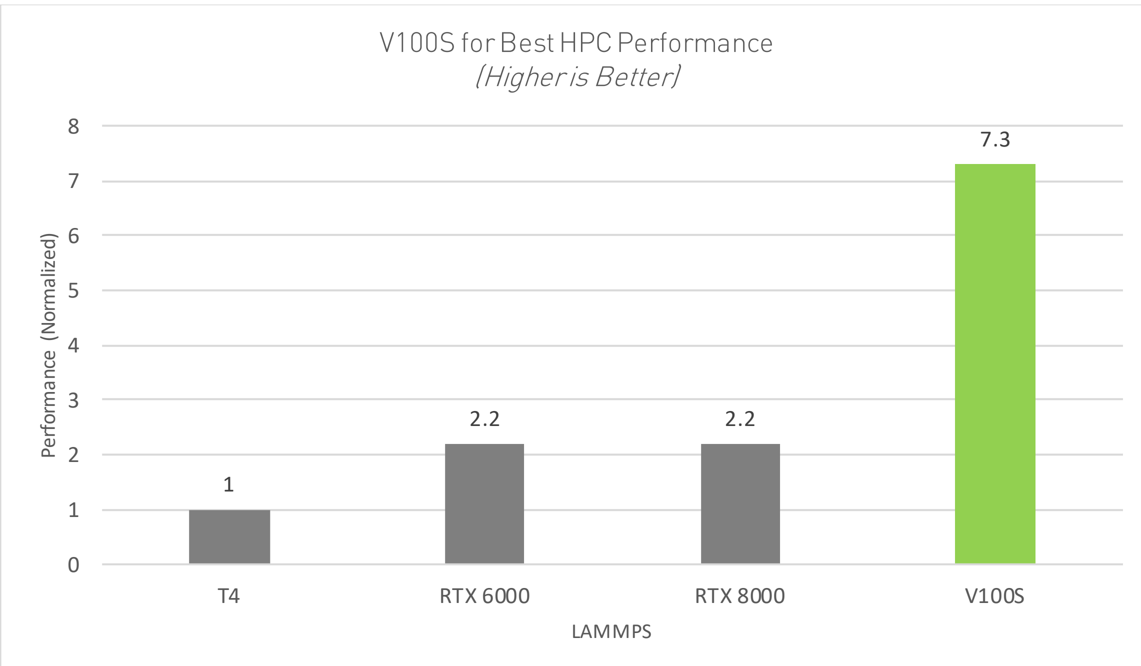
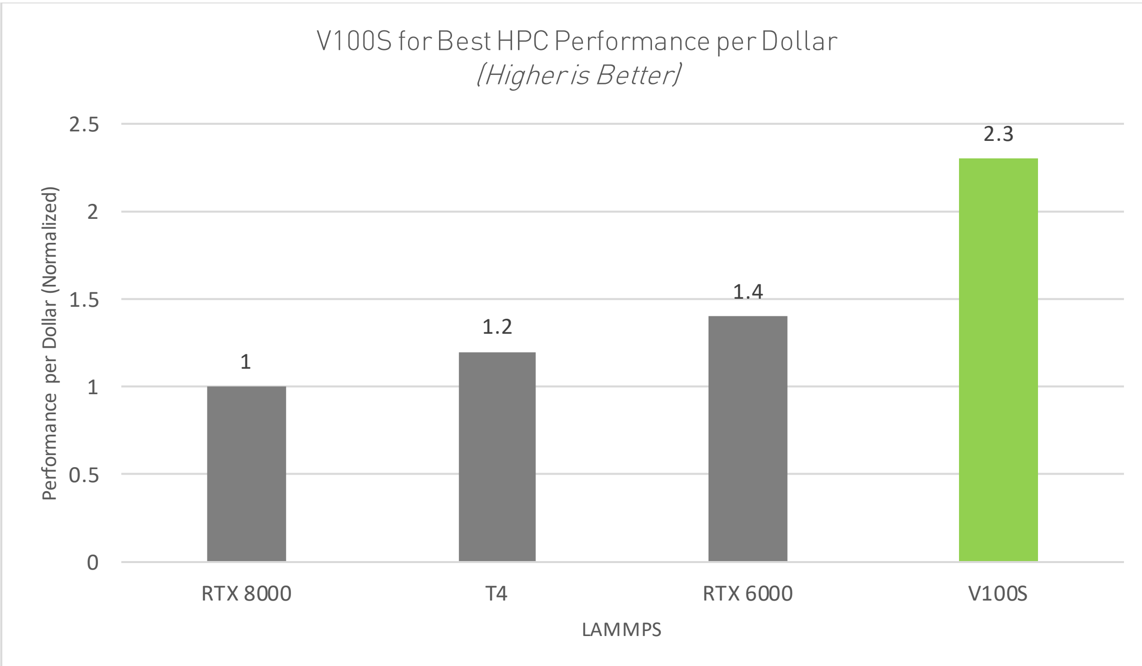


Figure 9 assumes estimated GPU street price plus NVIDIA vCS software cost.

Figure 9. NVIDIA vCS HPC Performance per Dollar



Knowledge Workers

As more knowledge worker users are added on a server, the server runs out of CPU resources. Adding an NVIDIA GPU for this workload offloads constraints on the CPU resulting in improved user experience and performance for end users. The NVIDIA nVector knowledge worker VDI workload was used to test user experience and performance with NVIDIA GPUs. NVIDIA M10, T4, Quadro RTX 6000, Quadro RTX 8000 and V100S achieve similar performance for this workload.

Customers are realizing the benefits of increased resource utilization by leveraging common virtualized GPU accelerated server resources to run virtual desktops and workstations but leveraging these same resources to run compute when users are logged off. Customers who want to be able to run compute workloads on the same infrastructure that they run VDI, might leverage a V100S to do so. Learn more about Using NVIDIA Virtual GPUs to Power Mixed Workloads in our whitepaper.

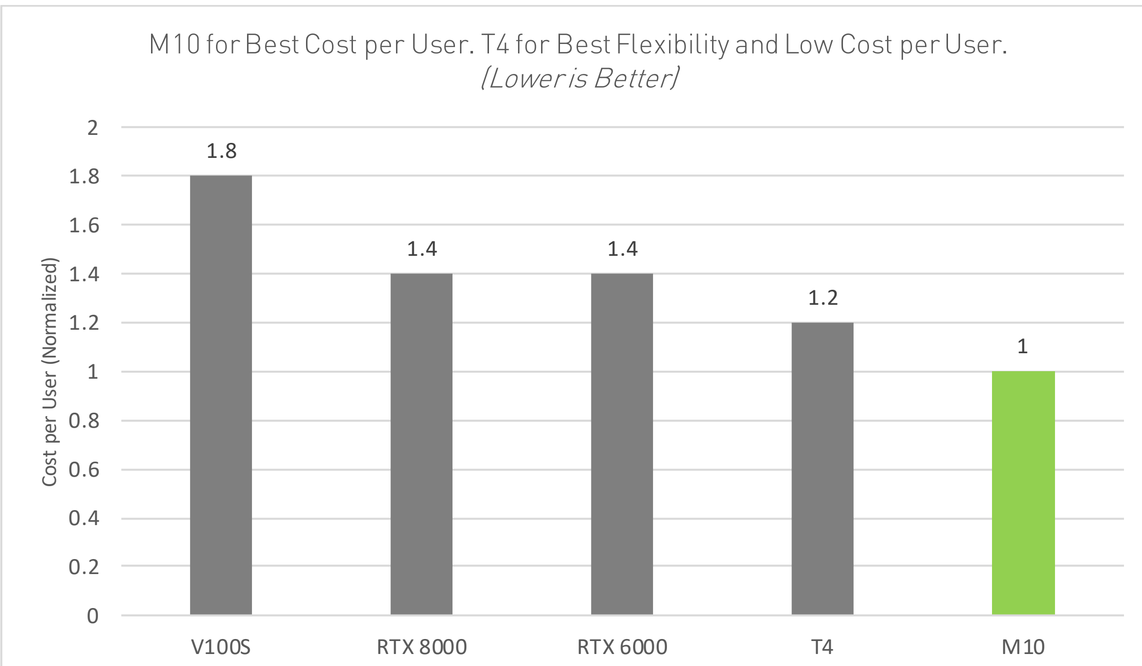
Despite having 48GB of frame buffer, the Quadro RTX 8000 supports a maximum of only 32 users due to reaching the context switching limit per GPU. Refer to Table 6 to see how many VDI users can be supported for each GPU (with 1GB profile size).

Table 6. Maximum Number of Supported NVIDIA GRID vPC Knowledge Workers (with 1GB Profile Size)

GPU	M10	T4	Quadro RTX 6000	Quadro RTX 8000	V100S
Max. Users	32	16	24	32	32

Figure 10 assumes estimated GPU street price plus NVIDIA GRID software cost with 3-year subscription divided by number of users.

Figure 10. NVIDIA GRID VDI Cost per User

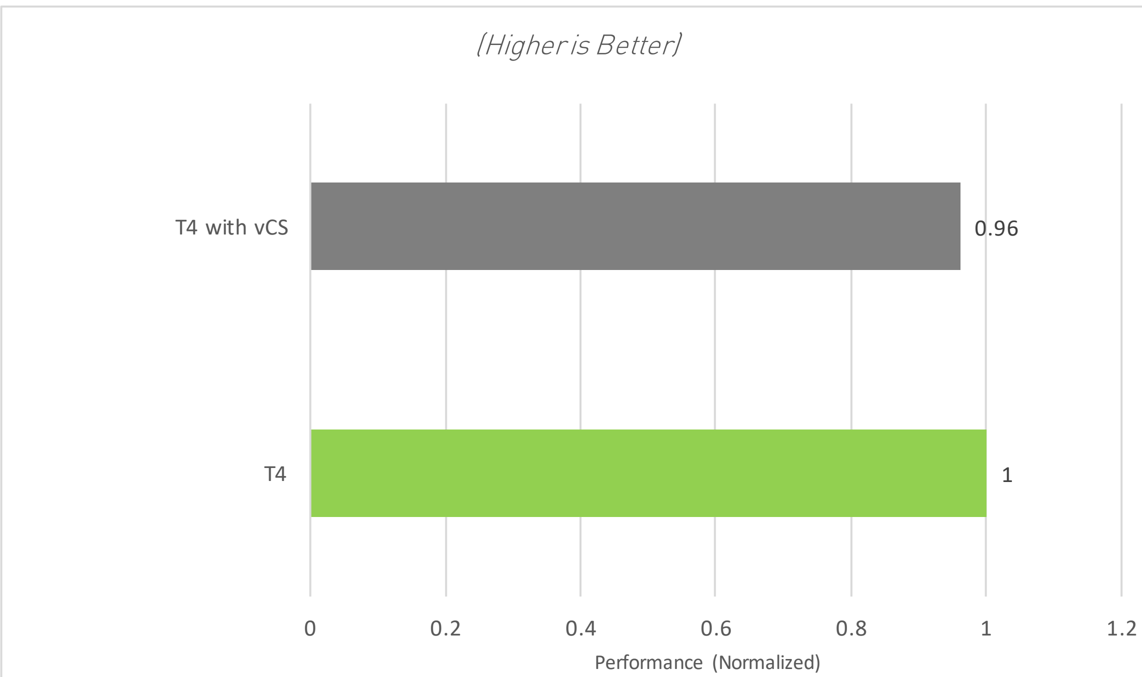


NVIDIA vGPU vs. Bare Metal

Organizations chose to virtualize servers and applications for various reasons (manageability, flexibility, and security to name a few) and are often willing to sacrifice performance. When allocating a full GPU to a workload in a virtualized environment, there is a performance difference. However, the performance difference of using NVIDIA vGPU is negligible and will depend on the workload, as well as various other configuration variables. The following example illustrates 4% lower performance with NVIDIA vGPU in comparison to a bare metal server running an AI Inference benchmark in a 1:1 configuration.

Figure 11 represents Resnet-50 V1.5 | TensorRT 6.0.1 | Batch Size = 128 | 19.12-py3 | Precision: Mixed.

Figure 11. Inference Benchmark



Impact of GPU Sharing

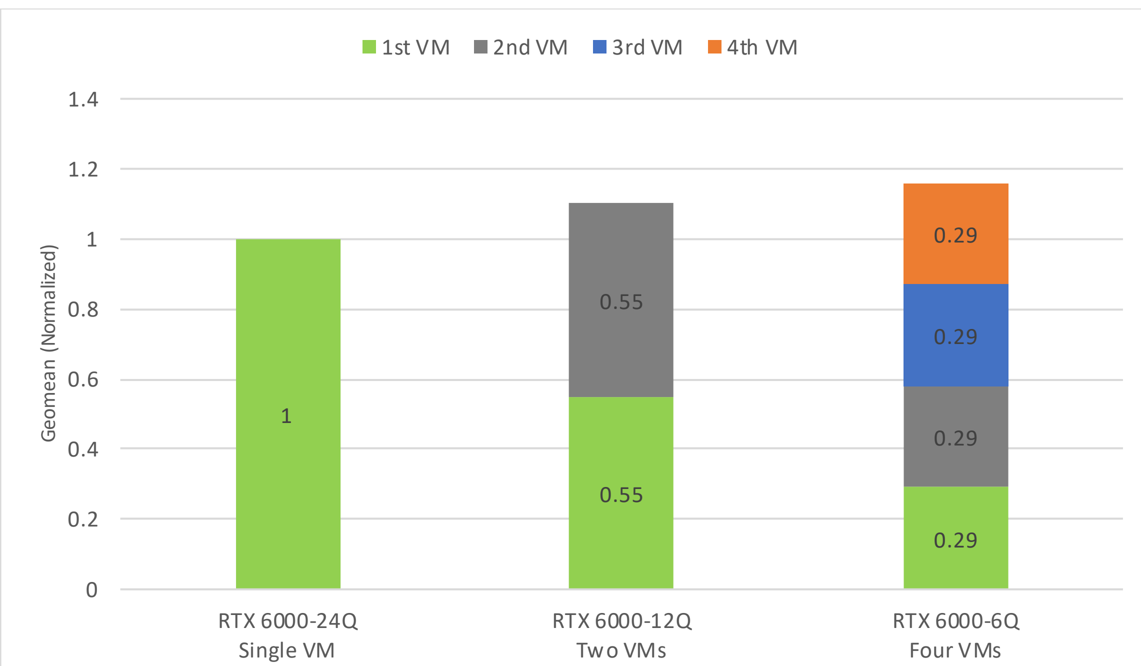
Improving overall utilization through sharing a GPU across multiple virtual machines with NVIDIA vGPU software is implemented by scheduling the time which each virtual machine can use the GPU. NVIDIA vGPU software provides multiple GPU scheduling options to accommodate a variety of Quality of Service (QoS) levels for sharing the GPU. View the NVIDIA vGPU product documentation for more information about GPU scheduling options.

In general, the performance per virtual machine when sharing a GPU with n virtual machines will be 1/n of the total performance of the GPU. Therefore, two virtual machines sharing a GPU will result in approximately 50 percent of the overall performance per virtual machine and four virtual machines will result in approximately 25 percent of the overall performance per virtual machine.

Figure 12 is an illustration of multiple virtual machines with an overall throughput increase of 16%.

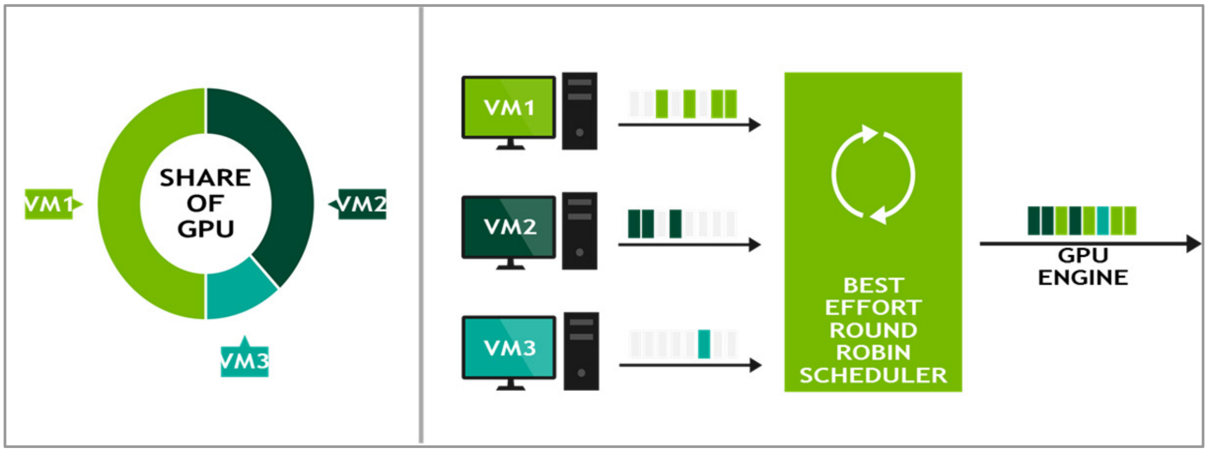
Figure 12 represents SPECviewperf13 results tested on a server with Intel Xeon Gold (18C, 3.0GHz), Quadro vDWS with RTX 8000 with Equal Share scheduler, VMware ESXi 6.7.0 U3, host/guest driver 440.44/441.66, VM config, Windows 10, 8 vCPU, 16GB memory.

Figure 12. Virtual GPU Sharing



However, when workloads across virtual machines are not executed at the same time, or aren't always GPU bound, the performance can exceed the expected performance. The default GPU scheduling policy, "Best Effort," will be selected for this to happen as it leverages unused GPU time of other virtual machines. See Figure 13 for a simplified view of how the "Best Effort" GPU scheduler works.

Figure 13. Best Effort GPU Scheduler

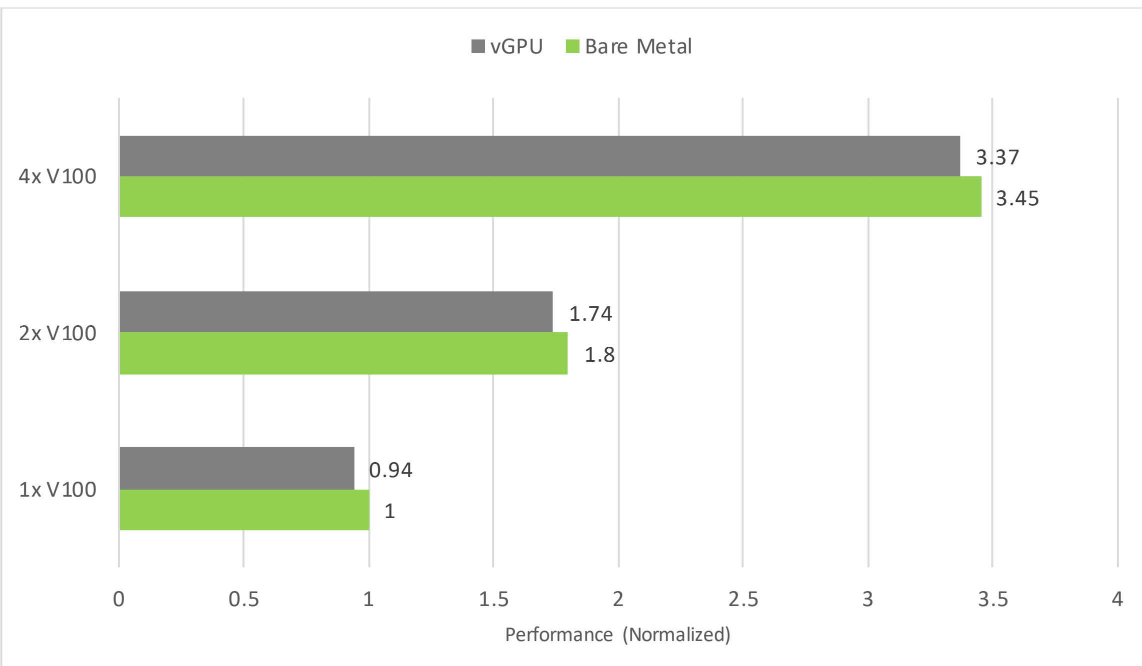


NVIDIA vGPU Aggregation

The scaling factor of virtual machines with vGPU aggregation is like the scaling factor using non-virtualized configurations. NVIDIA virtual GPU technology supports aggregating vGPUs for highest performance within a virtual machine via NVLink and traditional PCIe-based solutions. NVLink enables a high-speed, direct GPU-to-GPU interconnect that provides higher bandwidth for multi GPU system configurations than traditional PCIe-based solutions.

Figure 14 represents Server Config: 2x Intel Xeon Gold (6140, 3.2GHz), VMware ESXi 6.7 U3, NVIDIA vCS 9.1 RC, NVIDIA V100 (32C profile), Driver 430.18, TensorFlow Resnet-50 V1, NGC 19.01, FP16 BS: 256.

Figure 14. NVIDIA vGPU Aggregation Performance



Conclusion

While this technical brief provides general guidance on how to select the right NVIDIA GPU for your workload, actual results may vary depending on the specific application being virtualized.

The most successful deployments are those that balance virtual machine density (scalability) with required performance. This is achieved when a proof of concept (POC) with production workloads is conducted while analyzing the utilization of all resources of a system and gathering subjective feedback from all stakeholders. Consistently analyzing resource utilization and gathering subjective feedback allows for optimizing the configuration to meet the performance requirements while optimizing the configuration for best scale.

Resources Links

NVIDIA GRID Resources:

NVIDIA GRID Windows 10 Profile Sizing Guidance

Quantifying the Impact of NVIDIA Virtual GPUs

NVIDIA GRID Solution Overview

NVIDIA GRID webpage

NVIDIA Quadro Virtual Workstation Resources:

NVIDIA Quadro Virtual Workstation Application Sizing Guide for Dassault Systèmes CATIA

NVIDIA Quadro Virtual Workstation Application Sizing Guide for Esri ArcGIS Pro

NVIDIA Quadro Virtual Workstation Application Sizing Guide for Siemens NX

NVIDIA Quadro vDWS Solution Overview

NVIDIA Quadro vDWS webpage

NVIDIA Virtual Compute Server (vCS) Resources:

NVIDIA Virtual Compute Server webpage

NVIDIA Virtual Compute Server Solution Overview

Webinar: Introducing the Modern Data Center Powered by NVIDIA Virtual Compute Server

Other Resources:

Try NVIDIA vGPU for free

Using NVIDIA Virtual GPUs to Power Mixed Workloads

NVIDIA Virtual GPU Software Documentation

NVIDIA vGPU Certified Servers

Notice

OpenCL

Trademarks

Copyright