White Paper
Cisco public

The bridge to possible

Deploy Cisco UCS M5 Servers with NVIDIA GRID 10 on VMware vSphere 6.7 and Remote Desktop Session Hosts and Desktops

Contents

What you will learn

Using the increased processing power of Cisco Unified Computing System™ (Cisco UCS®) C-Series Rack Servers and B-Series Blade Servers, applications with the most demanding graphics requirements are being virtualized. To enhance the capability to deliver these high-performance and graphics-intensive applications in Virtual Desktop Infrastructure (VDI), Cisco offers support for the NVIDIA GRID RTX 6000, RTX 8000, P6, and M10 cards in the Cisco UCS portfolio of PCI Express (PCIe) and mezzanine form-factor cards for the Cisco UCS C-Series Rack Servers and B-Series Blade Servers respectively.

With the addition of the new graphics processing cards, the engineering, design, imaging, and marketing departments of organizations can now experience the benefits that desktop virtualization brings to the applications they use at higher user densities per server. Users of Microsoft Windows 10 and Office 2016 or later versions can benefit from the NVIDIA M10 high-density graphics card, deployable on Cisco UCS C240 M5 and C480 M5 Rack Servers.

These graphics capabilities help enable organizations to centralize their graphics workloads and data in the data center. This capability greatly benefits organizations that need to be able to shift work or collaborate geographically. Until now, graphics files have been too large to move, and the files have had to be local to the person using them to be usable.

The PCIe graphics cards in Cisco UCS C-Series servers offer these benefits:

- Support for two or four full-length, full-power NVIDIA GRID cards in 2-Rack-Unit (2RU) or 4RU form-factor servers
- Support for up to 48 high-performance Graphics Processing Unit (GPU)–supported graphics workstation users on a single rack server
- Cisco UCS Manager integration for management of the servers and NVIDIA GRID cards
- End-to-end integration with Cisco UCS management solutions, including Cisco UCS Central Software and Cisco UCS Director
- More efficient use of rack space with Cisco UCS rack servers with two or four NVIDIA GRID cards

The modular LAN-on-motherboard (mLOM) form-factor NVIDIA graphics cards in the Cisco UCS B-Series servers offers these benefits:

- Support for two NVIDIA P6 mLOM form-factor cards per server
- Up to four times the GPU-supported user density on B200 M5 Servers in a half-width blade server compared to the previous-generation B200 M4 Servers with a single NVIDIA M6 card (32 instead of 8)
- Cisco UCS Manager integration for management of the servers and the NVIDIA GRID GPU card
- End-to-end integration with Cisco UCS management solutions, including Cisco UCS Central Software and Cisco UCS Director
- More efficient use of rack space with Cisco UCS blade servers with two or four NVIDIA GRID cards

An important element of this document's design is VMware's support for the NVIDIA GRID virtual Graphics Processing Unit (vGPU) feature in VMware vSphere 6. Prior versions of vSphere supported only virtual direct graphics acceleration (vDGA) and virtual Shared Graphics Acceleration (vSGA), so support for vGPU in vSphere 6 greatly expands the range of deployment scenarios using the most versatile and efficient configuration of the GRID cards.

New with vSphere 6.5, you can configure host graphics settings, customize vGPU graphics settings on a per-virtual machine basis, and view GPU statistics using the Advanced option on the Performance tab

The purpose of this document is to help our partners and customers integrate NVIDIA GRID graphics processing cards on Cisco UCS B200 M5 Blade Servers and C240 M5 Rack Servers on VMware vSphere and VMware Remote Desktop Services (RDS) server sessions and desktops in vGPU mode.

Please contact our partners NVIDIA and VMware for lists of applications that are supported by the card, hypervisor, and desktop broker in each mode.

The objective here is to provide the reader with specific methods for integrating Cisco UCS servers with NVIDIA GRID M10 and RTX 6000 and 8000 cards with VMware vSphere and VMware products so that the servers, hypervisors, and virtual desktops are ready for installation of graphics applications.

GRID vGPU profiles

In any given enterprise, the needs of individual users vary widely. One of the main benefits of the GRID vGPU is the flexibility to use various vGPU profiles designed to serve the needs of different classes of end users.

Although the needs of end users can be diverse, for simplicity users can be grouped into the following categories: knowledge workers, power users, and designers.

- For knowledge workers, the main areas of importance include office productivity applications, a robust web experience, and fluid video playback. Knowledge workers have the least-intensive graphics demands, but they expect the same smooth, fluid experience that exists natively on today's graphics-accelerated devices such as desktop PCs, notebooks, tablets, and smartphones.
- Power users are users who need to run more demanding office applications, such as office productivity software, image editing software such as Adobe Photoshop, mainstream Computer-Aided Design (CAD) software such as Autodesk AutoCAD, and Product Lifecycle Management (PLM) applications. These applications are more demanding and require additional graphics resources with full support for APIs such as OpenGL and Direct3D.
- Designers are users in an organization who run demanding professional applications such as high-end CAD software and professional Digital Content Creation (DCC) tools. Examples include Autodesk Inventor, PTC Creo, Autodesk Revit, and Adobe Premiere. Historically, designers have used desktop workstations and have been a difficult group to incorporate into virtual deployments because of their need for high-end graphics and the certification requirements of professional CAD and DCC software.

vGPU profiles allow the GPU hardware to be time-sliced to deliver exceptional shared virtualized graphics performance (Figure 1).

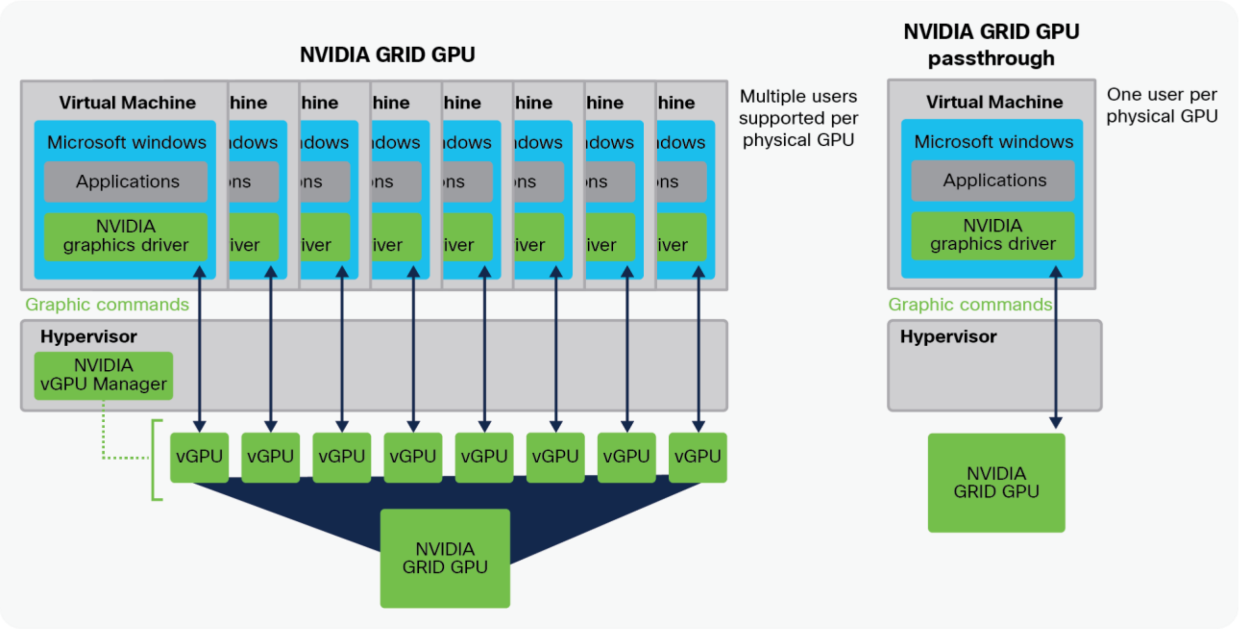


Figure 1.
NVIDIA GRID vGPU GPU system architecture

Cisco Unified Computing System

Cisco UCS is a next-generation data center platform that unites computing, networking, and storage access. The platform, optimized for virtual environments, is designed using open industry-standard technologies and aims to reduce Total Cost of Ownership (TCO) and increase business agility. The system integrates a low-latency; lossless 40 Gigabit Ethernet unified network fabric with enterprise-class, x86-architecture servers. It is an integrated, scalable, multichassis platform in which all resources participate in a unified management domain (Figure 2).

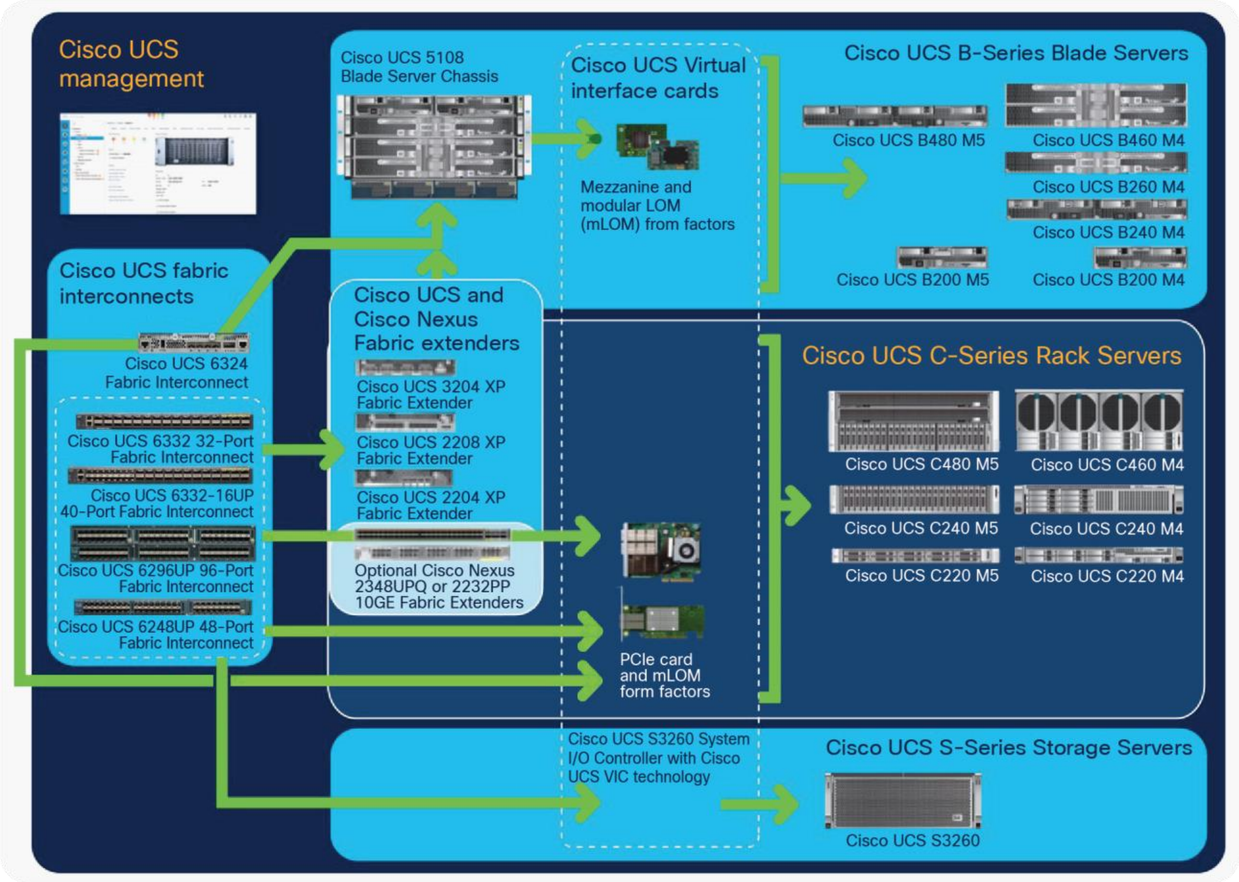

Cisco UCS management
Cisco UCS 5108 Blade Server Chassis
Cisco UCS Virtual interface cards
Mezzanine and modular LOM (mLOM) from factors
Cisco UCS B-Series Blade Servers
Cisco UCS B480 M5
Cisco UCS B460 M4
Cisco UCS B260 M4
Cisco UCS B240 M4
Cisco UCS B200 M5
Cisco UCS B200 M4
Cisco UCS fabric interconnects
Cisco UCS 6324 Fabric Interconnect
Cisco UCS 6332 32-Port Fabric Interconnect
Cisco UCS 6332-16UP 40-Port Fabric Interconnect
Cisco UCS 6296UP 96-Port Fabric Interconnect
Cisco UCS 6248UP 48-Port Fabric Interconnect
Cisco UCS and Cisco Nexus Fabric extenders
Cisco UCS 3204 XP Fabric Extender
Cisco UCS 2208 XP Fabric Extender
Cisco UCS 2204 XP Fabric Extender
Optional Cisco Nexus 2348UPQ or 2232PP 10GE Fabric Extenders
PCIe card and mLOM form factors
Cisco UCS C-Series Rack Servers
Cisco UCS C480 M5
Cisco UCS C460 M4
Cisco UCS C240 M5
Cisco UCS C240 M4
Cisco UCS C220 M5
Cisco UCS C220 M4
Cisco UCS S3260 System I/O Controller with Cisco UCS VIC technology
Cisco UCS S-Series Storage Servers
Cisco UCS S3260

Figure 2.
Cisco UCS components

The main components of Cisco UCS are described here:

- Computing: The system is based on an entirely new class of computing system that incorporates blade servers and modular servers based on Intel processors.
- Network: The system is integrated onto a low-latency, lossless, 40-Gbps unified network fabric. This network foundation consolidates LANs, SANs, and High-Performance Computing (HPC) networks, which are separate networks today. The unified fabric lowers costs by reducing the number of network adapters, switches, and cables and by decreasing power and cooling requirements.
- Virtualization: The system unleashes the full potential of virtualization by enhancing the scalability, performance, and operational control of virtual environments. Cisco® security, policy enforcement, and diagnostic features are now extended into virtualized environments to better support changing business and IT requirements.

- Storage access: The system provides consolidated access to local storage, SAN storage, and Network-Attached Storage (NAS) over the unified fabric. With storage access unified, Cisco UCS can access storage over Ethernet, Fibre Channel, Fibre Channel over Ethernet (FCoE), and Small Computer System Interface over IP (iSCSI) protocols. This capability provides customers with choice for storage access and investment protection. In addition, server administrators can preassign storage-access policies for system connectivity to storage resources, simplifying storage connectivity and management and helping increase productivity.
- Management: Cisco UCS uniquely integrates all system components, enabling the entire solution to be managed as a single entity by Cisco UCS Manager. The manager has an intuitive GUI, a Command-Line Interface (CLI), and a robust API for managing all system configuration processes and operations.

Cisco UCS is designed to deliver these benefits:

- Reduced TCO and increased business agility
- Increased IT staff productivity through just-in-time provisioning and mobility support
- A cohesive, integrated system that unifies the technology in the data center; the system is managed, serviced, and tested as a whole
- Scalability through a design for hundreds of discrete servers and thousands of virtual machines and the capability to scale I/O bandwidth to match demand
- Industry standards supported by a partner ecosystem of industry leaders

Cisco UCS Manager

Cisco UCS Manager provides unified, embedded management of all software and hardware components of Cisco UCS through an intuitive GUI, a CLI, and an XML API. The manager provides a unified management domain with centralized management capabilities and can control multiple chassis and thousands of virtual machines. Tightly integrated Cisco UCS manager and NVIDIA GPU cards provides better management of firmware and graphics card configuration.

Cisco UCS 6454 Fabric Interconnect

The Cisco UCS 6454 Fabric Interconnect (Figure 3) is a 1RU 10/25/40/100 Gigabit Ethernet, FCoE and Fibre Channel switch offering up to 3.82 Tbps throughput and up to 54 ports. The switch has twenty-eight 10/25-Gbps Ethernet ports, four 1/10/25-Gbps Ethernet ports, six 40/100-Gbps Ethernet uplink ports, and 16 unified ports that can support 10/25-Gbps Ethernet ports or 8/16/32-Gbps Fibre Channel ports. All Ethernet ports are capable of supporting FCoE. Cisco HyperFlex™ nodes can connect at 10- or 25-Gbps speeds depending on the model of Cisco Virtual Interface Card (VIC) in the nodes and the Small Form-Factor Pluggable (SFP) optics or cables chosen.

Figure 3.
Cisco UCS 6454 Fabric Interconnect

Cisco UCS C-Series Rack Servers

Cisco UCS C-Series Rack Servers keep pace with Intel® Xeon® processor innovation by offering the latest processors with increased processor frequencies and improved security and availability features. With the increased performance provided by the Intel Xeon processor Scalable product family, C-Series servers offer an improved price-to-performance ratio. They also extend Cisco UCS innovations to an industry-standard rack-mount form factor, including a standards-based unified network fabric, Cisco VN-Link virtualization support, and Cisco Extended Memory Technology.

Designed to operate both in standalone environments and as part of Cisco UCS managed configuration, these servers enable organizations to deploy systems incrementally—using as many or as few servers as needed—on a schedule that best meets the organization's timing and budget. C-Series servers offer investment protection through the capability to deploy them either as standalone servers or as part of Cisco UCS.

One compelling reason that many organizations prefer rack-mount servers is the wide range of I/O options available in the form of PCIe adapters. C-Series servers support a broad range of I/O options, including interfaces supported by Cisco as well as adapters from third parties.

Cisco UCS C240 M5 Rack Server

The Cisco UCS C240 M5 Rack Server (Figure 4) is designed for both performance and expandability over a wide range of storage-intensive infrastructure workloads, from big data to collaboration.

The C240 M5 Small Form-Factor (SFF) server extends the capabilities of the Cisco UCS portfolio in a 2RU form factor with the addition of the Intel Xeon processor Scalable family, 24 DIMM slots for 2666-MHz DDR4 DIMMs, and up to 128 GB of capacity, up to 6 PCI Express (PCIe) 3.0 slots, and up to 26 internal SFF drives. The C240 M5 SFF server also includes one dedicated internal slot for a 12-Gbps SAS storage controller card. The C240 M5 server includes a dedicated internal mLOM slot for installation of a Cisco VIC or third-party Network Interface Card (NIC) without consuming a PCI slot, in addition to two 10GBASE-T Intel x550 embedded (on the motherboard) LOM ports.

In addition, the C240 M5 offers outstanding levels of internal memory and storage expandability with exceptional performance. It offers these features:

- Up to 24 DDR4 DIMMs at speeds up to 2666 MHz for improved performance and lower power consumption
- One or two Intel Xeon processor Scalable family CPUs
- Up to six PCIe 3.0 slots (four full-height, full-length for GPU)
- Six hot-swappable fans for front-to-rear cooling
- Twenty-four SFF front-facing SAS/SATA Hard-Disk Drives (HDDs) or SAS/SATA Solid-State Disks (SSDs)
- Optionally, up to two front-facing SFF Non-Volatile Memory Express (NVMe) PCIe SSDs (replacing SAS/SATA drives); these drives must be placed in front drive bays 1 and 2 only and are controlled from Riser 2 option C
- Optionally, up to two SFF, rear-facing SAS/SATA HDDs or SSDs or up to two rear-facing SFF NVMe PCIe SSDs; rear-facing SFF NVMe drives are connected from Riser 2, Option B or C
- Support for 12-Gbps SAS drives

- Flexible support on the dedicated mLOM slot on the motherboard, accommodating the following cards:
    - Cisco VICs
    - Quad-port Intel i350 1 Gigabit Ethernet RJ-45 mLOM NIC
- Two 1 Gigabit Ethernet embedded LOM ports
- Support for up to two double-wide NVIDIA Graphics Processing Units (GPUs), providing a robust graphics experience to more virtual users
- Excellent Reliability, Availability, and Serviceability (RAS) features with tool-free CPU insertion, easy-to-use latching lid, and hot-swappable and hot-pluggable components
- One slot for a micro-Secure Digital (SD) card on PCIe Riser 1 (Options 1 and 1B)
    - The micro-SD card serves as a dedicated local resource for utilities such as the Cisco Host Upgrade Utility (HUU).
    - Images can be pulled from a file share (Network File System [NFS] or Common Internet File System [CIFS]) and uploaded to the cards for future use.
- A mini-storage module connector on the motherboard supports either:
    - An SD card module with two SD card slots. Mixing different-capacity SD cards is not supported.
    - An M.2 module with two SATA M.2 SSD slots. Mixing different-capacity M.2 modules is not supported.

Note: SD cards and M.2 modules cannot be mixed. M.2 does not support RAID-1 with VMware. Only Microsoft Windows and Linux are supported.

The C240 M5 also increases performance and customer choice over many types of storage-intensive applications such as the following:

- Collaboration
- Small and medium-sized business (SMB) databases
- Big data infrastructure
- Virtualization and consolidation
- Storage servers
- High-performance appliances

The C240 M5 can be deployed as a standalone server or as part of Cisco UCS. Cisco UCS unifies computing, networking, management, virtualization, and storage access into a single integrated architecture that enables end-to-end server visibility, management, and control in both bare-metal and virtualized environments. Within a Cisco UCS deployment, the C240 M5 takes advantage of Cisco's standards-based unified computing innovations, which significantly reduce customers' TCO and increase business agility.

For more information about the Cisco UCS C240 M5 Rack Server, see https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/c240m5-sff-specsheet.pdf.

Figure 4.
Cisco UCS C240 M5 Rack Server

Cisco UCS VIC 1457

The Cisco UCS VIC 1457 (Figure 5) is a quad-port SFP28 mLOM card designed for the M5 generation of Cisco UCS C-Series Rack Servers. The card supports 10- and 25-Gbps Ethernet and FCoE, where the speed of the link is determined by the model of SFP optics or cables used. The card can be configured to use a pair of single links, or optionally to use all four links as a pair of bonded links. The VIC 1457 is used in conjunction with the Cisco UCS 6454 Fabric Interconnect.

The mLOM is used to install a Cisco VIC without consuming a PCIe slot, which provides greater I/O expandability. It incorporates next-generation Converged Network Adapter (CNA) technology from Cisco, providing investment protection for future feature releases. The card enables a policy-based, stateless, agile server infrastructure that can present up to 256 PCIe standards-compliant interfaces to the host, each dynamically configured as either a NIC or a Host Bus Adapter (HBA). The personality of the interfaces is set programmatically using the service profile associated with the server. The number, type (NIC or HBA), identity (MAC address and World Wide Name [WWN]), failover policy, adapter settings, bandwidth, and Quality-of-Service (QoS) policy of the PCIe interfaces are all specified using the service profile.

For more information about the VIC, see https://www.cisco.com/c/en/us/products/collateral/interfaces-modules/unified-computing-system-adapters/datasheet-c78-741130.html.

Figure 5.
Cisco UCS VIC 1457 CNA

Cisco UCS B200 M5 Blade Server

Delivering performance, versatility, and density without compromise, the Cisco UCS B200 M5 Blade Server (Figure 6) addresses the broadest set of workloads, including IT and web infrastructure and distributed databases. The enterprise-class Cisco UCS B200 M5 extends the capabilities of the Cisco UCS portfolio in a half-width blade form factor. The Cisco UCS B200 M5 harnesses the power of the latest Intel Xeon processor Scalable family CPUs with up to 3072 GB of RAM (using 128-GB DIMMs), two SSDs or HDDs, and connectivity with up to 80-Gbps throughput.

The B200 M5 server mounts in a Cisco UCS 5100 Series Blade Server Chassis or Cisco UCS Mini blade server chassis. It has 24 total slots for Error-Correcting Code (ECC) Registered DIMMs (RDIMMs) or Load-Reduced DIMMs (LR DIMMs). It supports one connector for the Cisco UCS VIC 1340 adapter, which provides Ethernet and FCoE.

The UCS B200 M5 has one rear mezzanine adapter slot, which can be configured with a Cisco UCS port expander card for the VIC. This hardware option can enable an additional four ports of the VIC 1340, bringing the total capability of the VIC 1340 to a dual native 40-Gbps interface or a dual four 10 Gigabit Ethernet port-channeled interface. Alternatively, the same rear mezzanine adapter slot can be configured with an NVIDIA P6 GPU.

The UCS B200 M5 has one front mezzanine slot. The UCS B200 M5 can be ordered with or without the front mezzanine card. The front mezzanine card can accommodate a storage controller or an NVIDIA P6 GPU.

For more information, see https://www.cisco.com/c/en/us/products/collateral/interfaces-modules/unified-computing-system-adapters/datasheet-c78-741130.html.

Figure 6.
Cisco UCS B200 M5 Blade Server (front view)

Cisco VIC 1440

The Cisco UCS VIC 1440 (Figure 7) is a single-port 40-Gbps or 4 x 10-Gbps Ethernet and FCoE-capable mLOM designed exclusively for the M5 generation of Cisco UCS B-Series Blade Servers. When used in combination with an optional port expander, the Cisco UCS VIC 1440 capabilities are enabled for two ports of 40-Gbps Ethernet. The Cisco UCS VIC 1440 enables a policy-based, stateless, agile server infrastructure that can present to the host PCIe standards-compliant interfaces that can be dynamically configured as either NICs or HBAs.

For more information, see https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-b-series-blade-servers/b200m5-specsheet.pdf.

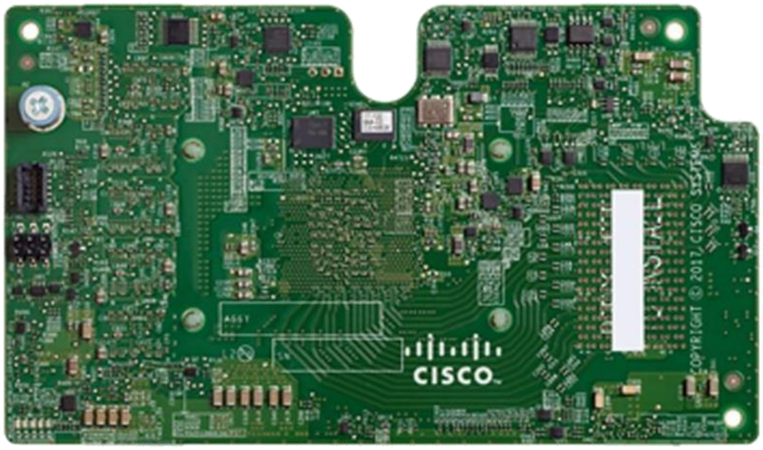


Figure 7.
Cisco UCS VIC 1440

NVIDIA

This section provides an overview of the NVIDIA components used in the solution described in this document.

NVIDIA Quadro RTX 6000 GPU

The NVIDIA Quadro RTX 6000 GPU (Figure 8) provides the following features:

- Used in combination with NVIDIA Quadro Virtual Data Center Workstation (vDWS) software, enables design engineers to work from high-powered virtual design workstations and render nodes to accelerate design workflows and arrive at their best creations faster
- With ray-tracing cores, a large frame buffer, and multiple profile sizes, gives engineers and designers the flexibility to run demanding workloads from the data center
- Supports up to two NVIDIA RTX 6000 cards per Cisco UCS C240 M5 rack-mount server (July 2020) or Cisco HyperFlex HX240c M5 rack-mount server (Q4 2020)

Figure 8.
NVIDIA Quadro RTX 6000 GPU

NVIDIA Quadro RTX 8000 GPU

The NVIDIA Quadro RTX 8000 GPU (Figure 9) provides the following features:

- Includes 48 GB of memory for the largest models: twice the frame buffer size of the RTX 6000
- Enables designers to work with the largest and most complex ray-tracing and visual computing workloads
- Delivers exceptional flexibility with Quadro vDWS software, powering virtual design workstations and render nodes to propel creative workflows
- Supports up to two NVIDIA RTX 8000 cards per Cisco UCS C240 M5 rack-mount server (July 2020) or Cisco HyperFlex HX240c M5 rack-mount server (Q4 2020)

Figure 9.
NVIDIA Quadro RTX 8000 GPU

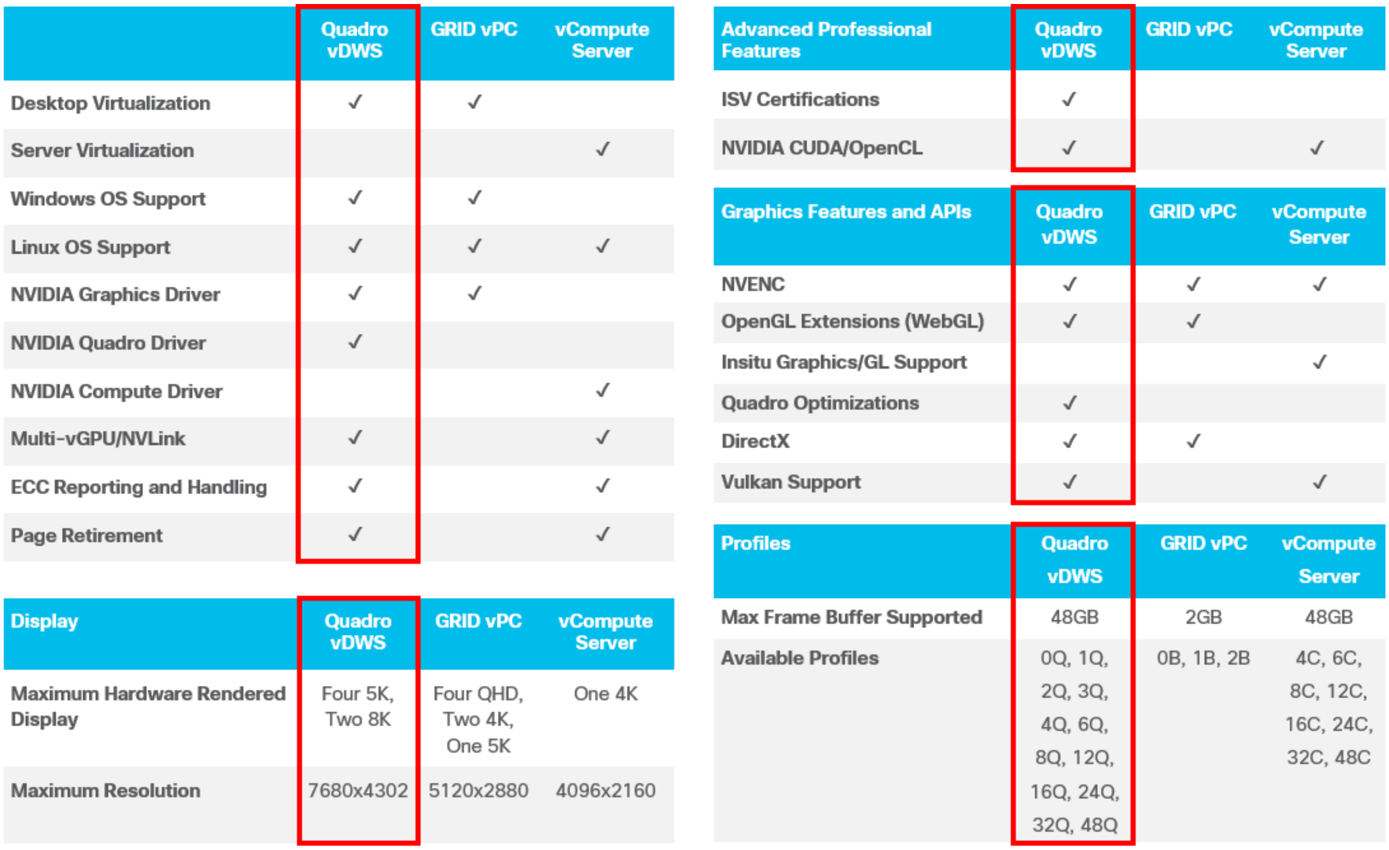
NVIDIA Quadro Virtual Data Center Workstation Software

For stunning industrial design, advanced special effects, complex scientific visualization, and more, Quadro provides the world's preeminent visual computing platform. Now, with NVIDIA Quadro vDWS software, you can deliver the most powerful virtual workstation from the data center or cloud to any device, anywhere. Millions of creative and technical professionals can access the most demanding applications from any device, work from anywhere, and tackle larger data sets, all while meeting the need for greater security.

Virtual workstations free users from the confines of physical location, delivering resources from the data center and providing secure access on any device, anywhere. NVIDIA Quadro vDWS extends the trusted benefits of Quadro to deliver a true GPU-accelerated data center. It lets IT virtualize any application from the data center with a workstation-class user experience. Now your business can eliminate constrained workflows that inhibit agility, and users can securely collaborate in real time without borders or limits. You can efficiently centralize all your applications and data, achieving dramatically lower IT operating expenses.

Table 1 summarizes the main features of Quadro vDWS software. These features provide outstanding flexibility for supporting critical capabilities and selecting the virtual workstation operating system.

Table 1. NVIDIA Quadro vDWS features



	Quadro vDWS	GRID vPC	vCompute Server
Desktop Virtualization	✓	✓	
Server Virtualization			✓
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	GRID vPC	vCompute Server
Maximum Hardware Rendered Display	Four 5K, Two 8K	Four QHD, Two 4K, One 5K	One 4K
Maximum Resolution	7680x4302	5120x2880	4096x2160

Advanced Professional Features	Quadro vDWS	GRID vPC	vCompute Server
ISV Certifications	✓		
NVIDIA CUDA/OpenCL	✓		✓

Graphics Features and APIs	Quadro vDWS	GRID vPC	vCompute Server
NVENC	✓	✓	✓
OpenGL Extensions (WebGL)	✓	✓	
Insitu Graphics/GL Support			✓
Quadro Optimizations	✓		
DirectX	✓	✓	
Vulkan Support	✓		✓

Profiles	Quadro vDWS	GRID vPC	vCompute Server
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

Sizing considerations

You need to consider four main factors when sizing a graphics workload deployment:

- User types and requirements
- GPU used
- Scheduling engine
- Frame buffer

User types and requirements

Grouping users by their vGPU use pattern is the first step in a successful virtual graphics workstation deployment. As a general rule, Cisco uses the classifications listed in Table 2.

Table 2. User classifications and characteristics

User classification	Characteristics
Light	• Read-only for review and documentation • Project management • Small parts or subassembly
Medium	• Read-only and full application • Medium assemblies
Heavy	• Full application • Large assemblies or full product

For this document, virtual workstations are constructed with the three user classifications shown in Table 2. Cisco recommends running a proof-of-concept study with your users to develop user classifications based on your specific use case.

NVIDIA GRID

NVIDIA GRID is the industry's most advanced technology for sharing vGPUs across multiple virtual desktop and application instances. You can now use the full power of NVIDIA data center GPUs to deliver a superior virtual graphics experience to any device anywhere. The NVIDIA GRID platform provides exceptional levels of performance, flexibility, manageability, and security—offering the right level of user experience for any virtual workflow.

For more information about NVIDIA GRID technology, see https://www.nvidia.com/object/nvidia-grid.html.

NVIDIA GRID GPU

The NVIDIA GRID solution runs on top of award-winning NVIDIA Maxwell-powered GPUs. These GPUs come in two server form factors: the NVIDIA Tesla P6 for blade servers and converged infrastructure, and the NVIDIA Tesla M10 and P40 for rack and tower servers.

NVIDIA GRID license requirements

GRID 5.0 requires concurrent user licenses and an on-premises NVIDIA license server to manage the licenses. When the guest OS boots, it contacts the NVIDIA license server and consumes one concurrent license. When the guest OS shuts down, the license is returned to the pool.

GRID 5.0 also requires the purchase of a 1:1 ratio of concurrent licenses to NVIDIA Support, Update, and Maintenance Subscription (SUMS) instances.

The following NVIDIA GRID products are available as licensed products on NVIDIA Tesla GPUs:

- Virtual workstation
- Virtual PC
- Virtual applications

For complete details about GRID license requirements, see https://images.nvidia.com/content/grid/pdf/GRID-Licensing-Guide.pdf.

VMware vSphere 6.7

VMware provides virtualization software. VMware's enterprise software hypervisors for servers—VMware vSphere ESX, vSphere ESXi, and vSphere—are bare-metal hypervisors that run directly on server hardware without requiring an additional underlying operating system. VMware vCenter Server for vSphere provides central management and complete control and visibility into clusters, hosts, virtual machines, storage, networking, and other critical elements of your virtual infrastructure.

vSphere 6.5 introduces many enhancements to the vSphere hypervisor, VMware virtual machines, vCenter Server, virtual storage, and virtual networking, further extending the core capabilities of the vSphere platform.

Graphics acceleration in VMware RDS host server sessions and desktops

The software- and hardware-accelerated graphics features available with the PC over IP (PCoIP) display protocol enable remote desktop users to run 3D applications ranging from Google Earth to CAD and other graphics-intensive applications.

- NVIDIA GRID vGPU (shared GPU hardware acceleration): Available with vSphere 6.0 and later, this feature allows a physical GPU on an ESXi host to be shared among virtual machines. Use this feature if you require high-end, hardware-accelerated workstation graphics.
- Virtual dedicated graphics acceleration: Available with vSphere 5.5 and later, the vDGA feature dedicates a single physical GPU on an ESXi host to a single virtual machine. Use this feature if you require high-end, hardware-accelerated workstation graphics.
- Virtual shared graphics acceleration: Available with vSphere 5.1 and later, the vSGA feature allows multiple virtual machines to share the physical GPUs on ESXi hosts. You can use 3D applications for design, modeling, and multimedia.
- Soft 3D: Software-accelerated graphics, available with vSphere 5.0 and later, allows you to run DirectX 9 and OpenGL 2.1 applications without requiring a physical GPU. Use this feature for less-demanding 3D applications such as Windows Aero themes, Microsoft Office 2010, and Google Earth.

For these features, up to two monitors are supported, and the maximum screen resolution is 1920 x 1200.

With VMware Horizon 6.2, NVIDIA GRID vGPU and vDGA are now also supported in remote applications running on Microsoft RDS hosts.

GPU acceleration for Microsoft Windows desktops

Engineers, designers, and scientists have traditionally relied on dedicated graphics workstations to perform the most demanding tasks, such as manipulating 3D models and visually analyzing large data sets. These standalone workstations carry high acquisition and maintenance costs. In addition, in areas such as oil and gas, space exploration, aerospace, engineering, science, and manufacturing, individuals with these advanced requirements must be located in the same physical location as the workstation.

This document describes hardware-accelerated graphics in VMware virtual desktops in VMware Horizon. It begins with typical use cases and matches these use cases to the three types of graphics acceleration, explaining the differences. Later sections provide installation and configuration instructions, as well as best practices and troubleshooting guidance.

Note: This guide describes hardware-accelerated graphics in a VMware Horizon environment that uses a VMware vSphere infrastructure.

Purpose of hardware-accelerated graphics

Moving the graphics acceleration hardware from the workstation to a server is a critical architectural innovation. This shift changes the computing metaphor for graphics processing, putting the additional computing, memory, networking, and security advantages of the data center at the disposal of the user, so that complex models and very large data sets can be accessed and manipulated from almost anywhere (Figure 10).

With appropriate network bandwidth and suitable remote client devices, IT can now offer the most advanced users an immersive 3D-graphics experience while freeing them from the limitations of the old computing metaphor:

- Fewer physical resources are needed.
- The wait time to open complex models or run simulations is greatly reduced.
- Users are no longer tied to a single physical location.

In addition to handling the most demanding graphical workloads, hardware acceleration can reduce CPU use for less-demanding basic desktop and published applications and for video encoding and decoding, which includes the default Blast Extreme remote display protocol.

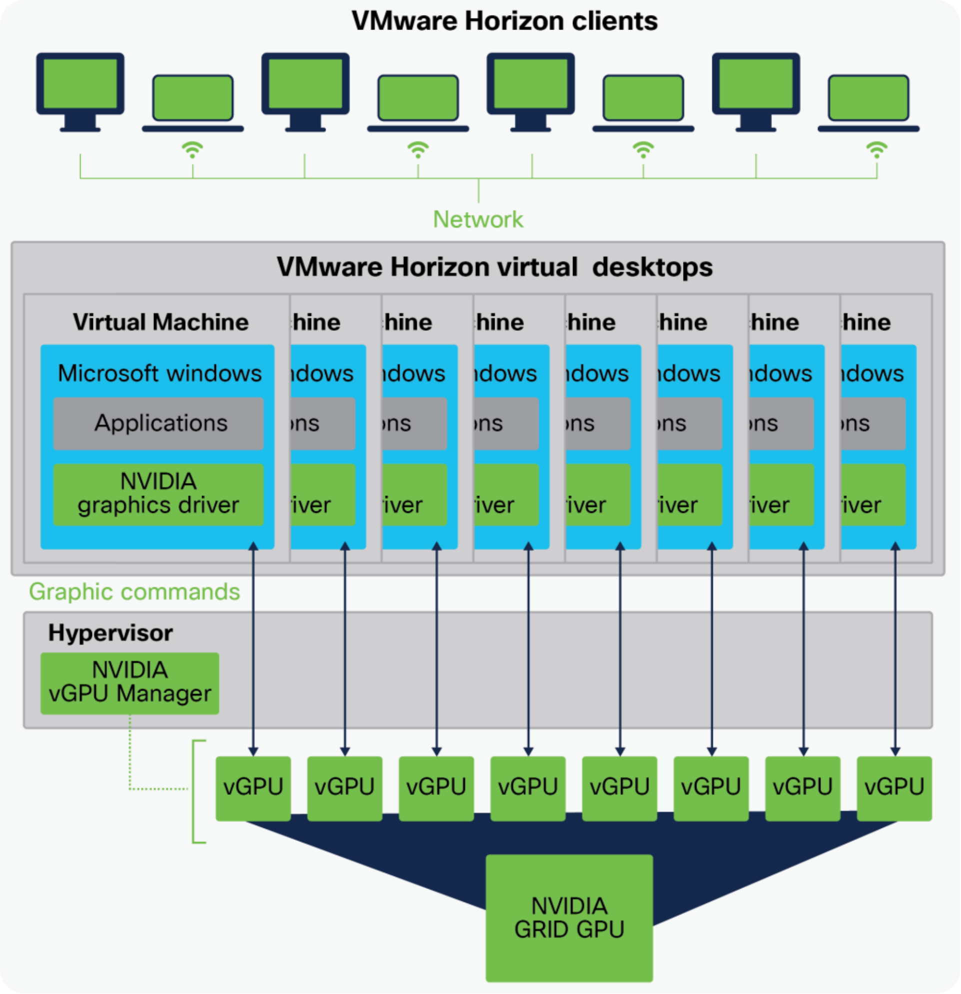


Figure 10.
VMware GPU 3D Pro process flow

GPU acceleration for Microsoft Windows Server

With 3D graphics configured for RDS hosts, both applications in application pools and applications running on RDS desktops can display 3D graphics.

- The following 3D graphics options are available:
- NVIDIA GRID vGPU (shared GPU hardware acceleration): A physical GPU on an ESXi host is shared among multiple virtual machines. This option requires ESXi 6.0 or later.
- AMD multiuser GPU using vDGA: A physical GPU on an ESXi host is shared among multiple virtual machines. This option requires ESXi 6.0 or later.
- vDGA: A physical GPU on an ESXi host is dedicated to a single virtual machine. This option requires ESXi 5.5 or later.

GPU sharing for VMware RDS workloads

With 3D graphics configured for RDS hosts, both applications in application pools and applications running on RDS desktops can display 3D graphics.

The following 3D graphics options are available:

- NVIDIA GRID vGPU (shared GPU hardware acceleration): A physical GPU on an ESXi host is shared among multiple virtual machines. This option requires ESXi 6.0 or later.
- vDGA: A physical GPU on an ESXi host is dedicated to a single virtual machine. This option requires ESXi 5.5 or later.

With vDGA, you get better performance, but the RDS host must be in a manual farm.

With vGPU, each graphics card can support multiple RDS hosts, and the RDS hosts must be in a manual farm. If an ESXi host has multiple physical GPUs, you can also configure the way the ESXi host assigns virtual machines to the GPUs. By default, the ESXi host assigns virtual machines to the physical GPU with the fewest virtual machines already assigned. This approach is called performance mode. You can also choose consolidation mode, in which the ESXi host assigns virtual machines to the same physical GPU until the maximum number of virtual machines is reached before placing virtual machines on the next physical GPU.

To confirm that GPU acceleration is occurring, use a third-party tool such as GPU-Z. GPU-Z is available at https://www.techpowerup.com/gpuz/.

VMware Horizon 3D Pro requirements

When you enable 3D graphics, you can select a hardware or software graphics renderer and optimize the graphics memory allocated to the virtual machine. You can increase the number of displays in multiple-monitor configurations and change the video card settings to meet your graphics requirements.

The default setting for total video RAM is adequate for minimal desktop resolution. For more complex situations, you can change the default memory. Typically, 3D applications require video memory of 64 to 512 MB.

Fault tolerance is not supported for virtual machines that have 3D graphics enabled.

Prerequisites

Before using Horizon 3D Pro, verify the following:

- Verify that the virtual machine is powered off.
- Verify that virtual machine compatibility is ESXi 5.0 or later.
- To enable 3D graphics in virtual machines with Windows 8 guest operating systems, verify that virtual machine compatibility is ESXi 5.1 or later.
- To use a hardware 3D renderer, verify that graphics hardware is available. See Configuring 3D Graphics.
- If you update the virtual machine compatibility from ESXi 5.1 or later to ESXi 5.5 or later, reinstall VMware Tools to get the latest sVGA virtual graphics driver and Windows Display Driver Model driver.
- Verify that you have the Virtual machine.Configuration.Modify device settings privilege on the virtual machine.

Procedure

Use the following procedure to set up Horizon 3D Pro:

1. Right-click a virtual machine in the inventory and select Edit Settings.
2. On the Virtual Hardware tab, expand Video Card.
3. Select custom or automatic settings for your displays from the drop-down menu:
    - Auto-detect settings: Applies common video settings to the guest operating system
    - Specify custom settings: Lets you select the number of displays and the total video memory
4. Select the number of displays from the drop-down menu. You can set the number of displays and extend the screen across them.
5. Enter the required video memory.
6. **(Optional)** Select Enable 3D support. This check box is active only for guest operating systems on which VMware supports 3D.
7. **(Optional)** Select a 3D renderer:
    - Automatic: Selects the appropriate option (software or hardware) for this virtual machine
    - Software: Uses normal CPU processing for 3D calculations
    - Hardware: Requires graphics hardware (GPU) for faster 3D calculations

Note: The virtual machine will not power on if graphics hardware is not available.

8. Click OK.

Results

Sufficient memory allocation is set for this virtual machine's graphics.

The NVIDIA vGPU software includes two separate components:

- The NVIDIA vGPU manager that is loaded as a VMware Installation Bundle (VIB) into the vSphere ESXi hypervisor
- A separate guest OS NVIDIA vGPU driver that is installed within the guest operating system of your virtual machine (the guest virtual machine driver)

Using the NVIDIA vGPU technology with vSphere allows you to choose between dedicating a full GPU device to one virtual machine or allowing partial sharing of a GPU device by more than one virtual machine. You may want to choose the partial-sharing option in the following circumstances:

- You know that the applications in your virtual machines do not need the power of the full GPU.
- You have a limited number of GPU devices and you want them to be available to more than one team of users simultaneously.
- You sometimes want to dedicate a full GPU device to one virtual machine, but at other times you want to allow partial use of a GPU by a virtual machine.

Solution configuration

Figure 11 provides an overview of the solution configuration.

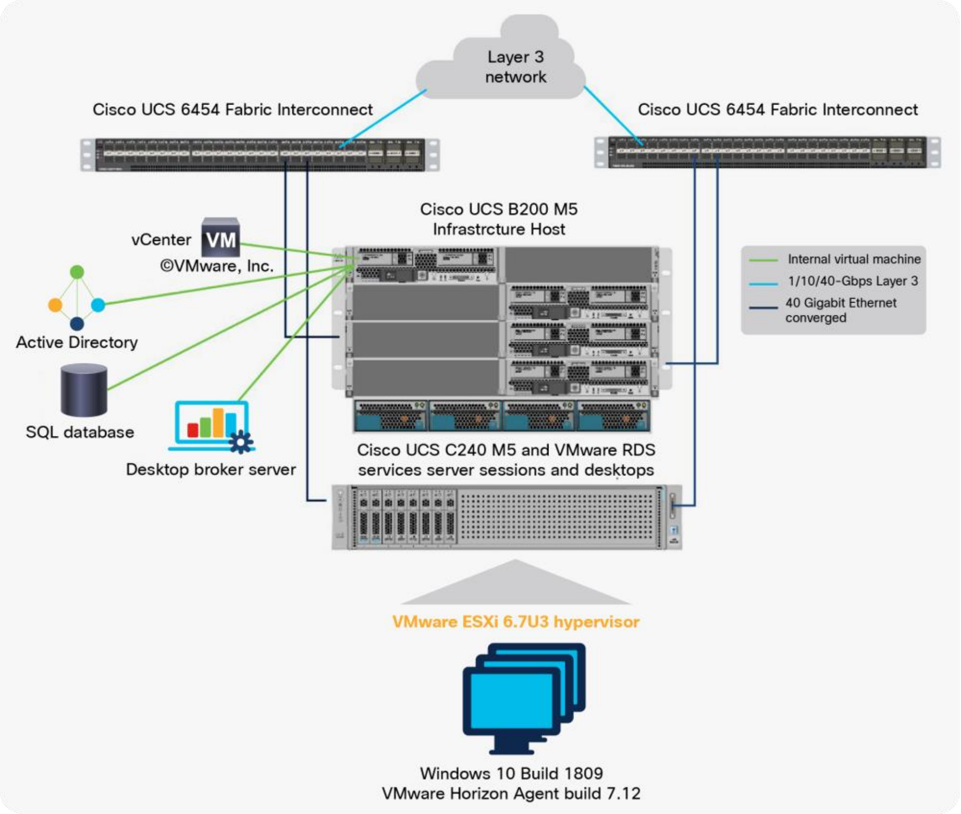

Layer 3 network
Cisco UCS 6454 Fabric Interconnect
Cisco UCS 6454 Fabric Interconnect
Cisco UCS B200 M5 Infrastrcture Host
vCenter
VM
©VMware, Inc.
Active Directory
SQL database
Desktop broker server
Internal virtual machine
1/10/40-Gbps Layer 3
40 Gigabit Ethernet converged
Cisco UCS C240 M5 and VMware RDS services server sessions and desktops
VMware ESXi 6.7U3 hypervisor
Windows 10 Build 1809
VMware Horizon Agent build 7.12

Figure 11.
Reference architecture

The hardware components in the solution are listed here:

- Cisco UCS C240 M5 Rack Server (two Intel Xeon Scalable Platinum 6454 CPUs at 2.10 GHz) with 1.5 TB of memory (64 GB x 24 DIMMs at 2666 MHz)
- Cisco UCS VIC 1457 mLOM (Cisco UCS C240 M5)
- Two Cisco UCS 6454 fourth-generation fabric interconnects
- NVIDIA RTX 8000 and 6000 cards
- Two Cisco Nexus® 9372 Switches (optional access switches)

The software components of the solution are listed here:

- Cisco UCS Firmware Release 4.04(g)
- VMware ESXi 6.7 for VDI hosts
- VMware Horizon Agent 7.12
- Microsoft Windows 10 64-bit
- Microsoft Server 2019
- NVIDIA GRID software and licenses:
    - NVIDIA_bootbank_NVIDIA-VMware_ESXi_6.7_Host_Driver_418.165.01-1OEM.670.0.0.8169922.vib
    - 426.94_grid_win10_server2016_server2019_64bit_international.exe

Configure Cisco UCS

This section describes the Cisco UCS configuration.

Install NVIDIA Tesla GPU card on Cisco UCS C240 M5

Install the RTX 8000 or 6000 GPU card on the Cisco UCS C240 M5 server.

Table 3 lists the minimum firmware required for the GPU cards.

Table 3. Minimum server firmware versions required for GPU cards

Cisco Integrated Management Controller (IMC)	BIOS minimum version
NVIDIA RTX 8000	Release 4.1(2a)
NVIDIA RTX 6000	Release 4.1(2a)

The rules for mixing NVIDIA GPU cards are as follows:

- Do not mix GRID GPU cards with Tesla GPU cards in the same server.
- Do not mix different models of Tesla GPU cards in the same server.

The rules for configuring the server with GPUs differ, depending on the server version and other factors. Table 4 lists rules for populating the Cisco UCS C240 M5 with NVIDIA GPUs.

Figure 12 shows a one-GPU installation, and Figure 13 shows a two-GPU installation.

Table 4. NVIDIA GPU population rules for Cisco UCS C240 M5 Rack Server

Single GPU	Dual GPU
Riser 1A, slot 2 **or** Riser 2A or 2B, slot 5	Riser 1A, slot 2 **and** Riser 2A or 2B, slot 5

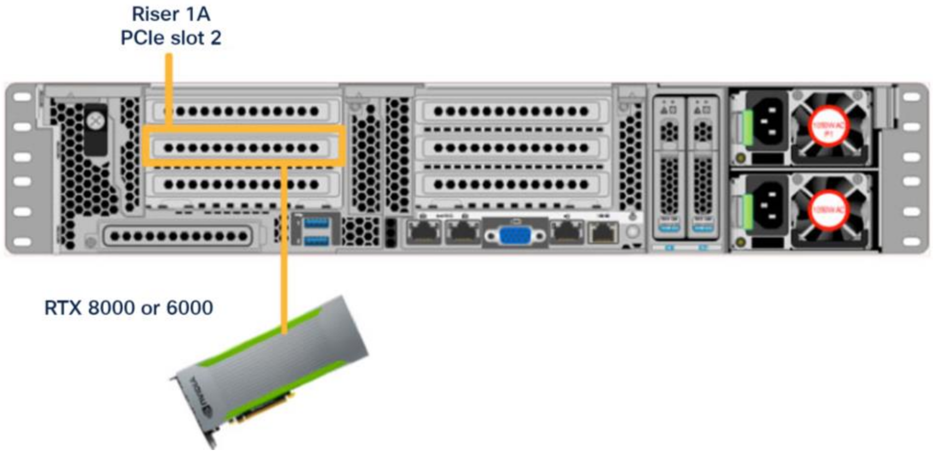


Figure 12.
One-GPU scenario

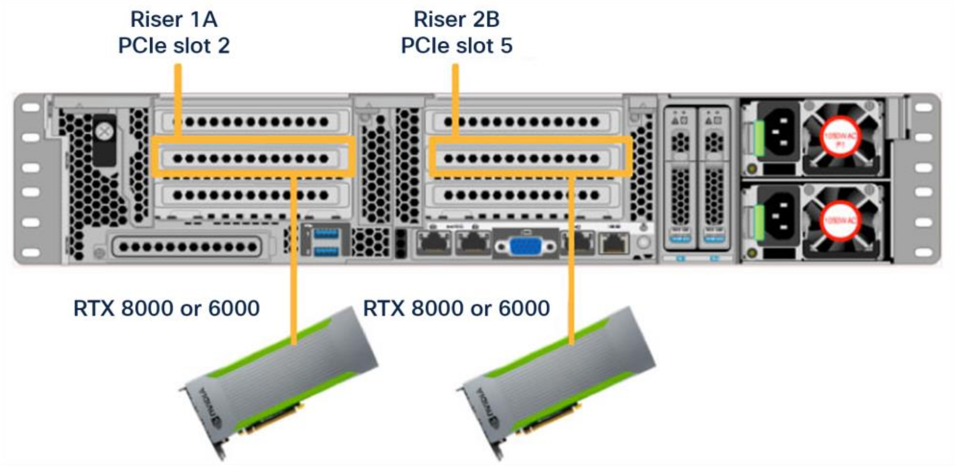


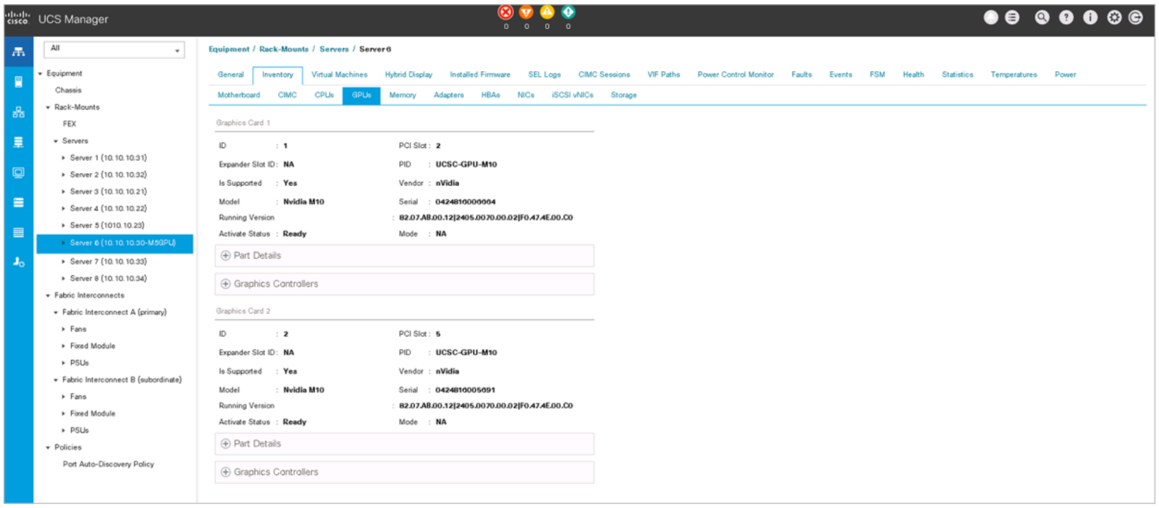
Figure 13.
Two-GPU scenario

For more information, see the Cisco UCS C240 M5 server installation and configuration document at https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/c/hw/C240M5/install/C240M5.pdf.

Configure the GPU card

Follow these steps to configure the GPU card:

1. After the NVIDIA RTX 6000 and 8000 GPU cards are physically installed and the Cisco UCS C240 M5 Rack Server is discovered in Cisco UCS Manager, select the server and choose Inventory > GPUs. As shown in Figure 14, PCIe slots 2 and 5 are used with two RTX 6000 and 8000 cards.



Equipment / Rack-Mounts / Servers / Server 6

Graphics Card 1

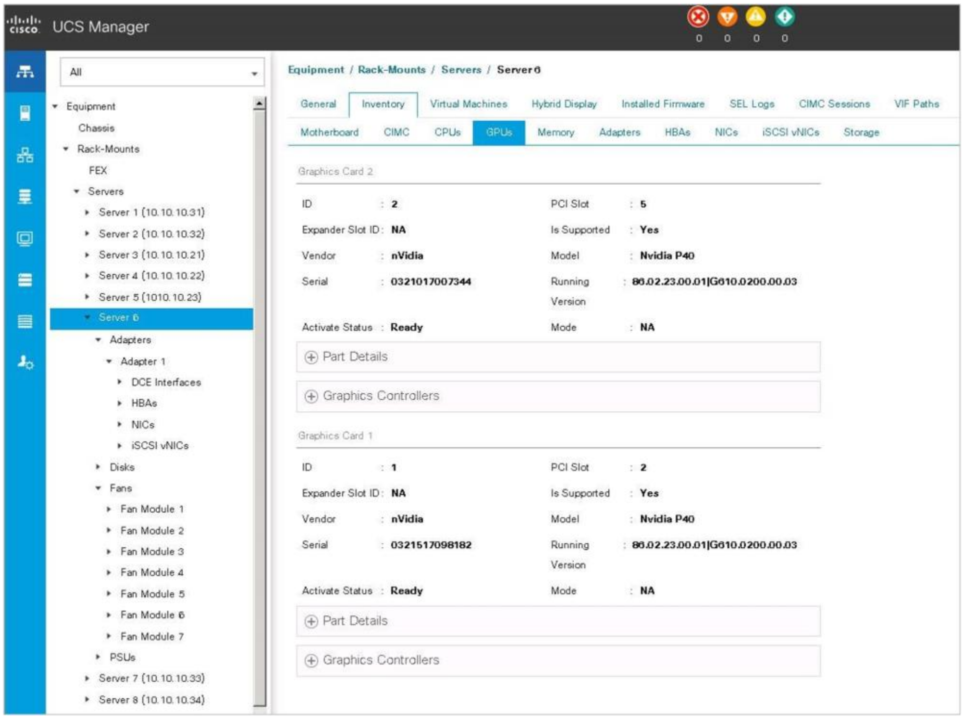
ID : 1 PCI Slot : 2
Expander Slot ID : NA PID : UCSC-GPU-M10
Is Supported : Yes Vendor : nVidia
Model : Nvidia M10 Serial : 0424810000004
Running Version : 82.07.AB.00.12|2405.0070.00.02|F0.47.4E.00.C0
Activate Status : Ready Mode : NA

Graphics Card 2

ID : 2 PCI Slot : 5
Expander Slot ID : NA PID : UCSC-GPU-M10
Is Supported : Yes Vendor : nVidia
Model : Nvidia M10 Serial : 0424810005091
Running Version : 82.07.AB.00.12|2405.0070.00.02|F0.47.4E.00.C0
Activate Status : Ready Mode : NA

Figure 14.
NVIDIA GRID M10 card inventory displayed in Cisco UCS Manager

2. After the NVIDIA RTX 6000 and 8000 GPU cards are physically installed and the Cisco UCS C240 M5 Rack Server is discovered in Cisco UCS Manager, select the server and choose Inventory > GPUs. As shown in Figure 15, PCIe slots 2 and 5 are used with the two RTX 6000 and 8000 cards.



Equipment / Rack-Mounts / Servers / Server 6

Graphics Card 2

ID : 2 PCI Slot : 5
Expander Slot ID : NA Is Supported : Yes
Vendor : nVidia Model : Nvidia P40
Serial : 0321017007344 Running Version : 86.02.23.00.01|G610.0200.00.03
Activate Status : Ready Mode : NA

Graphics Card 1

ID : 1 PCI Slot : 2
Expander Slot ID : NA Is Supported : Yes
Vendor : nVidia Model : Nvidia P40
Serial : 0321517098182 Running Version : 86.02.23.00.01|G610.0200.00.03
Activate Status : Ready Mode : NA

Figure 15.
NVIDIA RTX 6000 and 8000 card inventory displayed in Cisco UCS Manager

You can use Cisco UCS Manager upgrade the firmware for the NVIDIA GPU cards in managed Cisco UCS C240 M5 servers.

Note: VMware ESXi virtual machine hardware Version 9 or later is required for vGPU and vDGA configuration. Virtual machines with hardware Version 9 or later should have their settings managed through the VMware vSphere Web Client.

Install the NVIDIA GRID license server

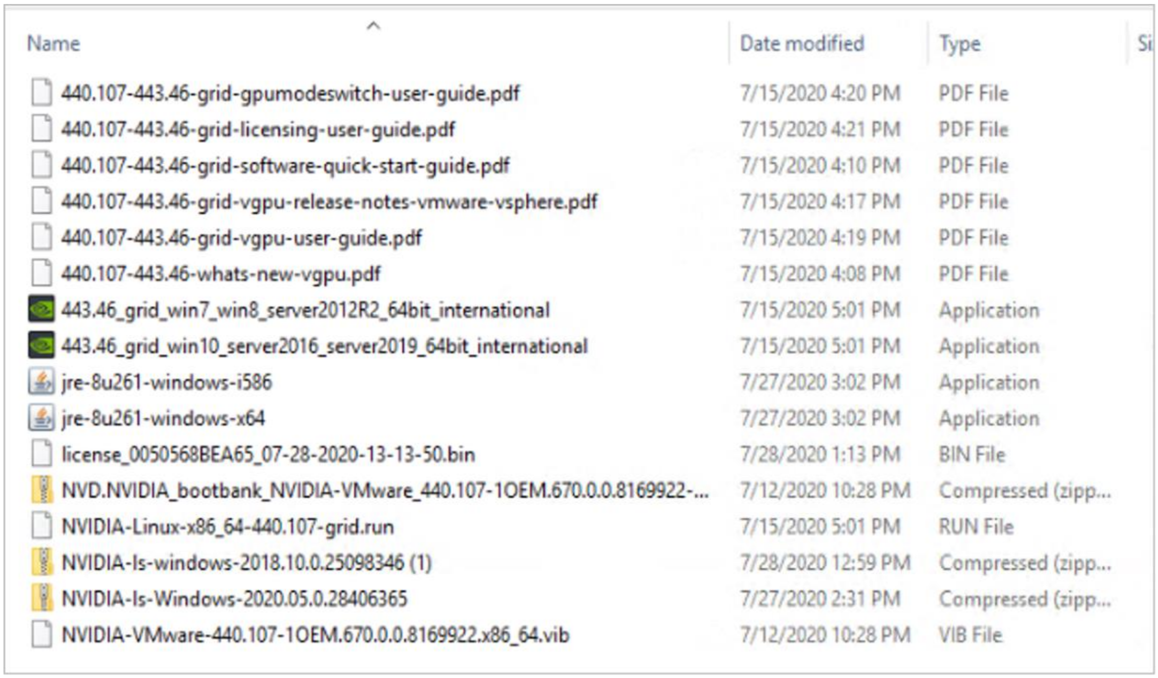
This section summarizes the installation and configuration process for the GRID license server.

The NVIDIA GRID vGPU is a licensed feature on RTX 6000 and 8000 cards. A software license is required to use the full vGPU features on a guest virtual machine. An NVIDIA license server with the appropriate licenses is required.

To get an evaluation license code and download the software, register at http://www.nvidia.com/object/grid-evaluation.html-utm_source=shorturl&utm_medium=referrer&utm_campaign=grideval.

Three packages are required for VMware ESXi host setup, as shown in Figure 16:

- The GRID license server installer
- The NVIDIA GRID Manager software, which is installed on VMware vSphere ESXi; the NVIDIA drivers and software that are installed in Microsoft Windows are also in this folder
- The GPU Mode Switch utility, which changes the cards from the default Compute mode to Graphics mode



Name	Date modified	Type	Si
440.107-443.46-grid-gpumodeswitch-user-guide.pdf	7/15/2020 4:20 PM	PDF File	
440.107-443.46-grid-licensing-user-guide.pdf	7/15/2020 4:21 PM	PDF File	
440.107-443.46-grid-software-quick-start-guide.pdf	7/15/2020 4:10 PM	PDF File	
440.107-443.46-grid-vgpu-release-notes-vmware-vsphere.pdf	7/15/2020 4:17 PM	PDF File	
440.107-443.46-grid-vgpu-user-guide.pdf	7/15/2020 4:19 PM	PDF File	
440.107-443.46-whats-new-vgpu.pdf	7/15/2020 4:08 PM	PDF File	
443.46_grid_win7_win8_server2012R2_64bit_international	7/15/2020 5:01 PM	Application	
443.46_grid_win10_server2016_server2019_64bit_international	7/15/2020 5:01 PM	Application	
jre-8u261-windows-i586	7/27/2020 3:02 PM	Application	
jre-8u261-windows-x64	7/27/2020 3:02 PM	Application	
license_0050568BEA65_07-28-2020-13-13-50.bin	7/28/2020 1:13 PM	BIN File	
NVD.NVIDIA_bootbank_NVIDIA-VMware_440.107-1OEM.670.0.0.8169922-...	7/12/2020 10:28 PM	Compressed (zipp...	
NVIDIA-Linux-x86_64-440.107-grid.run	7/15/2020 5:01 PM	RUN File	
NVIDIA-ls-windows-2018.10.0.25098346 (1)	7/28/2020 12:59 PM	Compressed (zipp...	
NVIDIA-ls-Windows-2020.05.0.28406365	7/27/2020 2:31 PM	Compressed (zipp...	
NVIDIA-VMware-440.107-1OEM.670.0.0.8169922.x86_64.vib	7/12/2020 10:28 PM	VIB File	

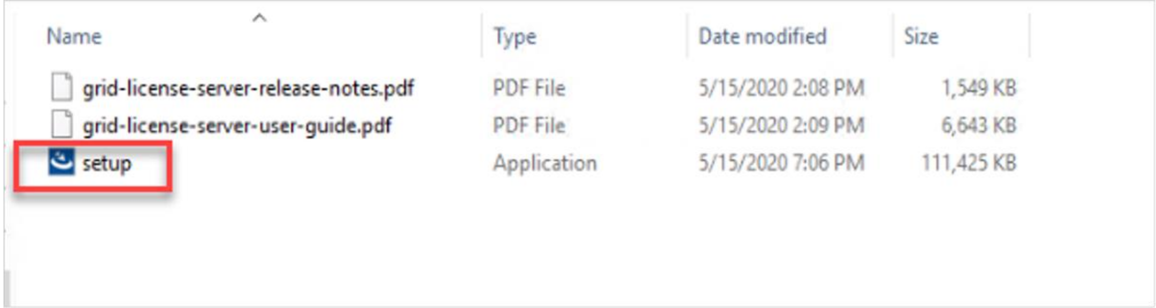
Figure 16.
Software required for NVIDIA GRID setup on the VMware ESXi host

Install the GRID 5.0 license server

The steps shown here use the Microsoft Windows version of the license server installed on Windows Server 2019. A Linux version of the license server is also available.

The GRID license server requires Java Version 7 or later. Go to Java.com and install the latest version.

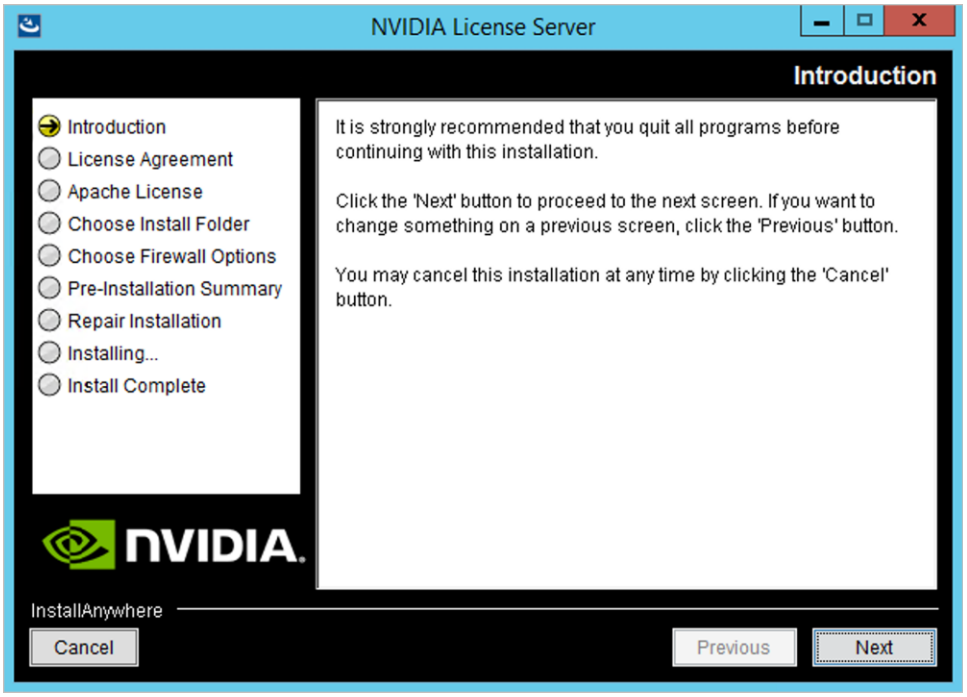
1. Extract and open the NVIDIA-ls-windows-2020.05.0.28406365 folder. Run setup.exe (Figure 17).



Name	Type	Date modified	Size
grid-license-server-release-notes.pdf	PDF File	5/15/2020 2:08 PM	1,549 KB
grid-license-server-user-guide.pdf	PDF File	5/15/2020 2:09 PM	6,643 KB
setup	Application	5/15/2020 7:06 PM	111,425 KB

Figure 17.
Run setup.exe

2. Click Next (Figure 18).



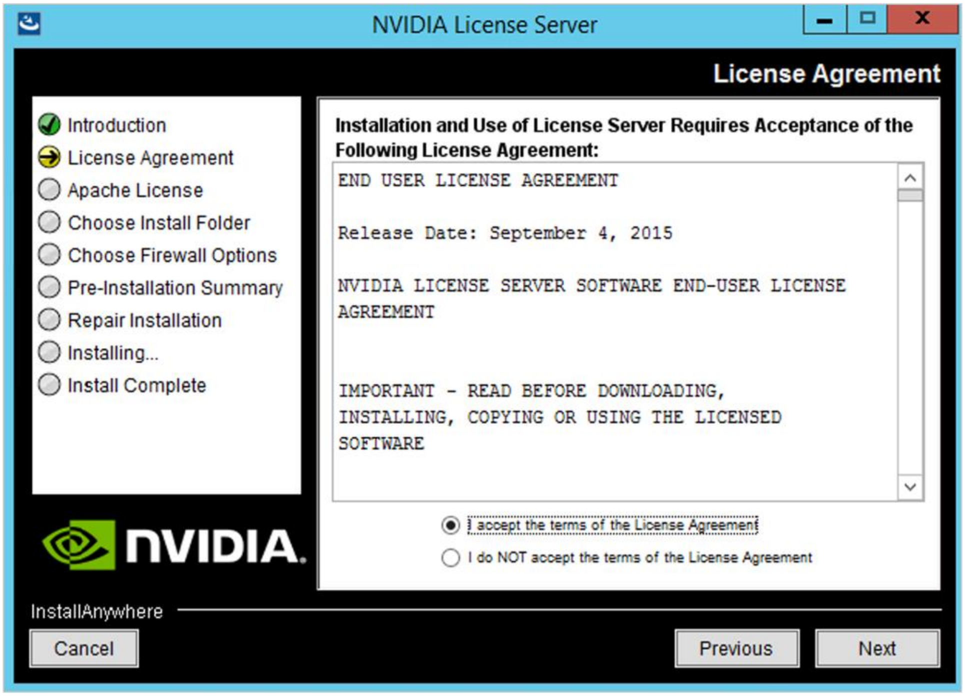
It is strongly recommended that you quit all programs before continuing with this installation.

Click the 'Next' button to proceed to the next screen. If you want to change something on a previous screen, click the 'Previous' button.

You may cancel this installation at any time by clicking the 'Cancel' button.

Figure 18.
NVIDIA License Server screen

3. Accept the license agreement and click Next Figure 19).



NVIDIA License Server

License Agreement

Installation and Use of License Server Requires Acceptance of the Following License Agreement:

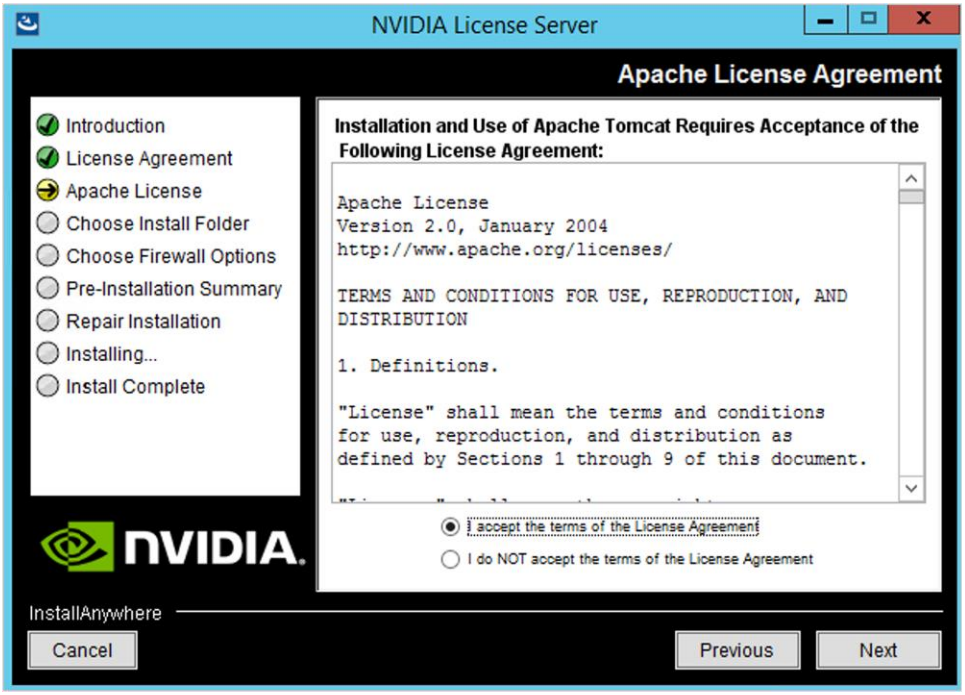
```
END USER LICENSE AGREEMENT

Release Date: September 4, 2015

NVIDIA LICENSE SERVER SOFTWARE END-USER LICENSE
AGREEMENT


IMPORTANT - READ BEFORE DOWNLOADING,
INSTALLING, COPYING OR USING THE LICENSED
SOFTWARE
```

● I accept the terms of the License Agreement

○ I do NOT accept the terms of the License Agreement

Figure 19.
NVIDIA License Agreement screen

4. Accept the Apache license agreement and click Next (Figure 20).



NVIDIA License Server

Apache License Agreement

Installation and Use of Apache Tomcat Requires Acceptance of the Following License Agreement:

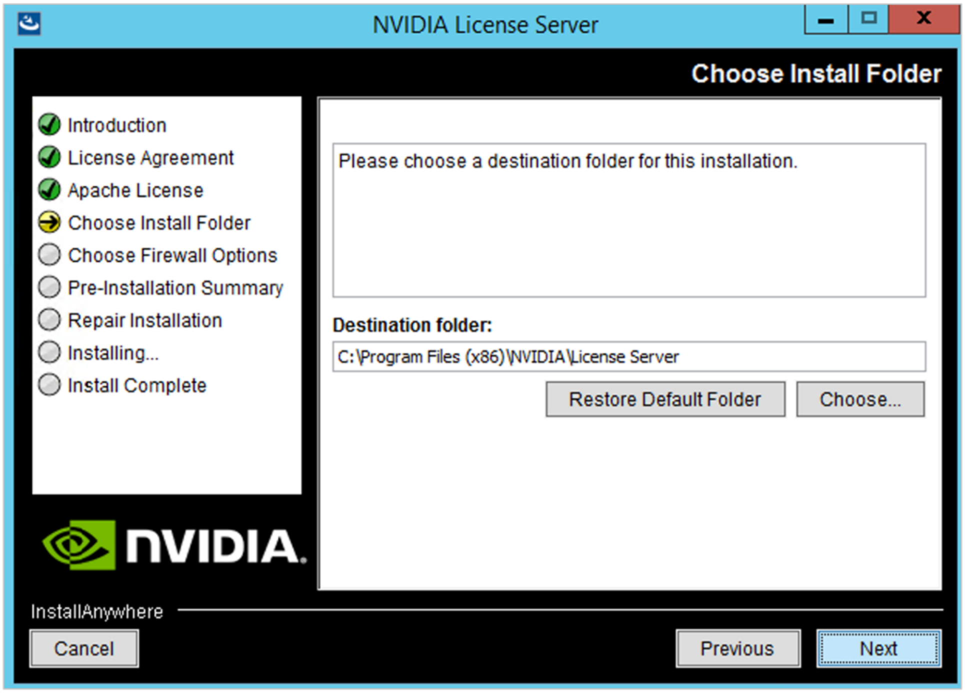
```
Apache License
Version 2.0, January 2004
http://www.apache.org/licenses/

TERMS AND CONDITIONS FOR USE, REPRODUCTION, AND
DISTRIBUTION

1. Definitions.

"License" shall mean the terms and conditions
for use, reproduction, and distribution as
defined by Sections 1 through 9 of this document.
```

● I accept the terms of the License Agreement

○ I do NOT accept the terms of the License Agreement

Figure 20.
NVIDIA License Agreement screen

5. Choose the desired installation folder and click Next (Figure 21).

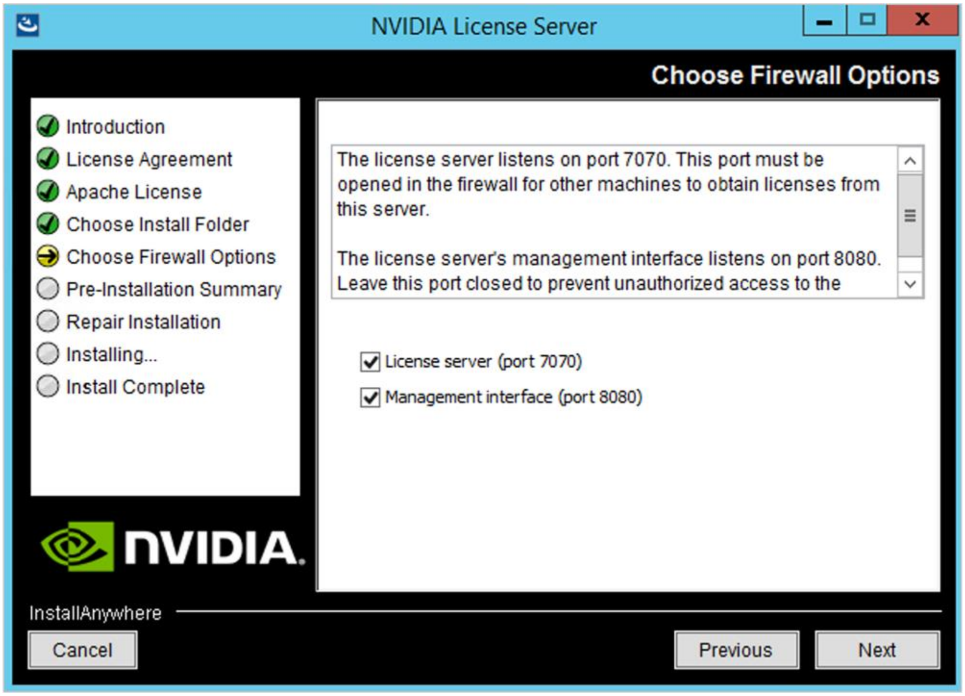


Please choose a destination folder for this installation.

Destination folder: C:\Program Files (x86)\NVIDIA\License Server

Figure 21.
Choosing a destination folder

6. The license server listens on port 7070. This port must be opened in the firewall for other machines to obtain licenses from this server. Select the "License server (port 7070)" option.
7. The license server's management interface listens on port 8080. If you want the administration page accessible from other machines, you will need to open port 8080. Select the "Management interface (port 8080)" option.
8. Click Next (Figure 22).



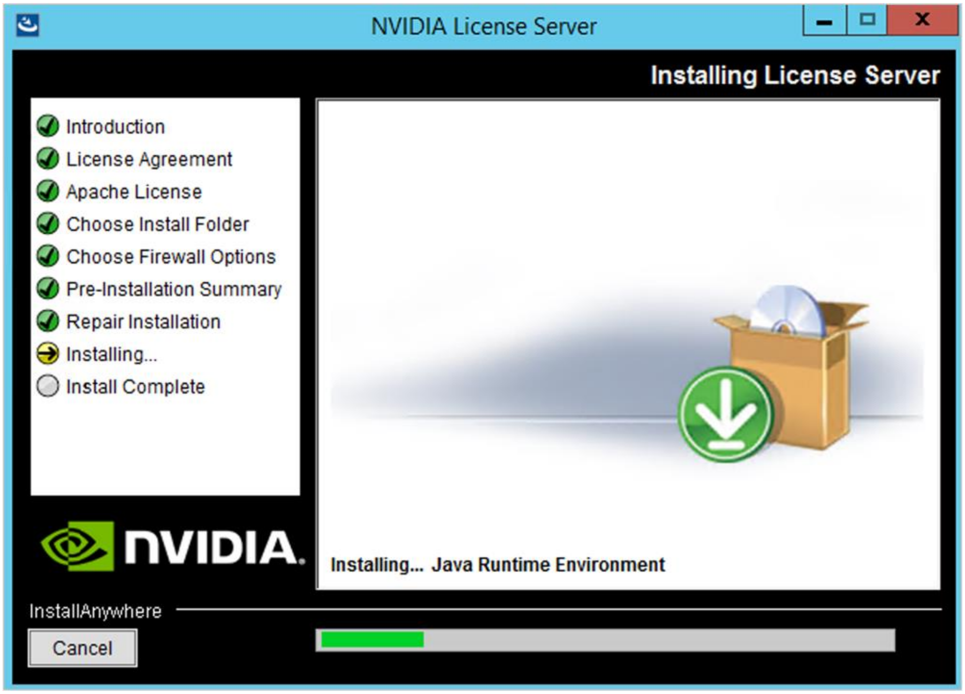
The license server listens on port 7070. This port must be opened in the firewall for other machines to obtain licenses from this server.

The license server's management interface listens on port 8080. Leave this port closed to prevent unauthorized access to the

☑ License server (port 7070)
☑ Management interface (port 8080)

Figure 22.
Setting firewall options

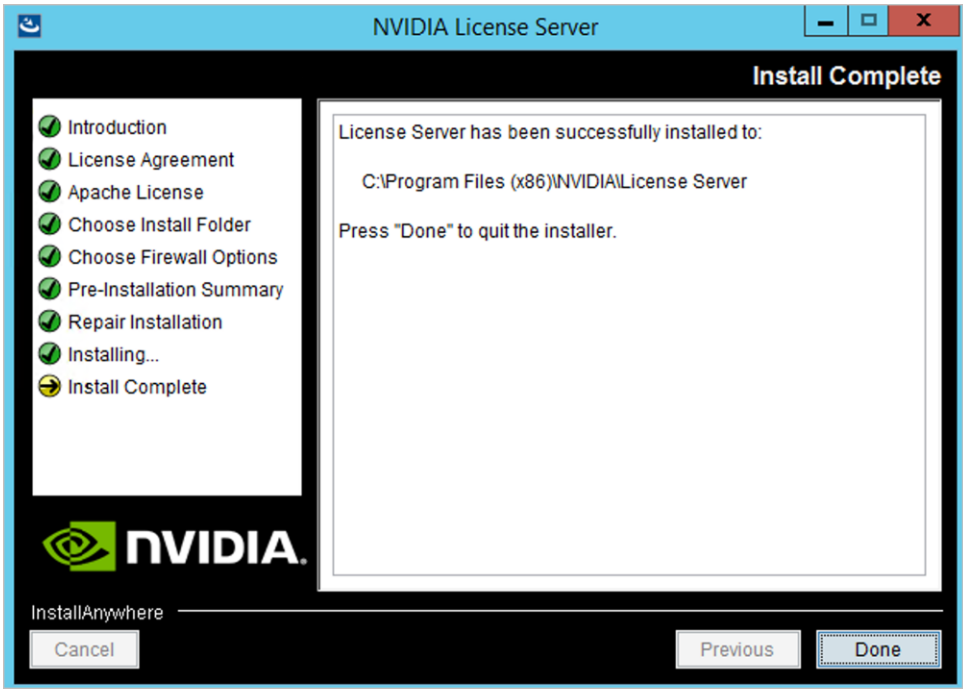
9. The Pre-installation Summary and Repair Installation options automatically progresses without user input (Figure 23).



Installing... Java Runtime Environment

Figure 23.
Installing the license server

10. When the installation process is complete, click Done (Figure 24).



License Server has been successfully installed to:

C:\Program Files (x86)\NVIDIA\License Server

Press "Done" to quit the installer.

Figure 24.
Installation complete

Configure the NVIDIA GRID license server

Now configure the NVIDIA Grid license server.

1. Log in to the license server site with the credentials set up during the registration process at nvidia.com/grideval. A license file is generated from https://nvid.nvidia.com/.
2. After you are logged in, click NVIDIA Licensing Portal.
3. Click Create Server (Figure 25).
4. Specify the fields as shown in Figure 26. In the License Server ID field, enter the MAC address of your local license server's NIC, a description, the failover server name, and the MAC address.
5. Specify the product and quantity using the drop-down menu and click Add (Figure 27).
6. Click Create License Server.

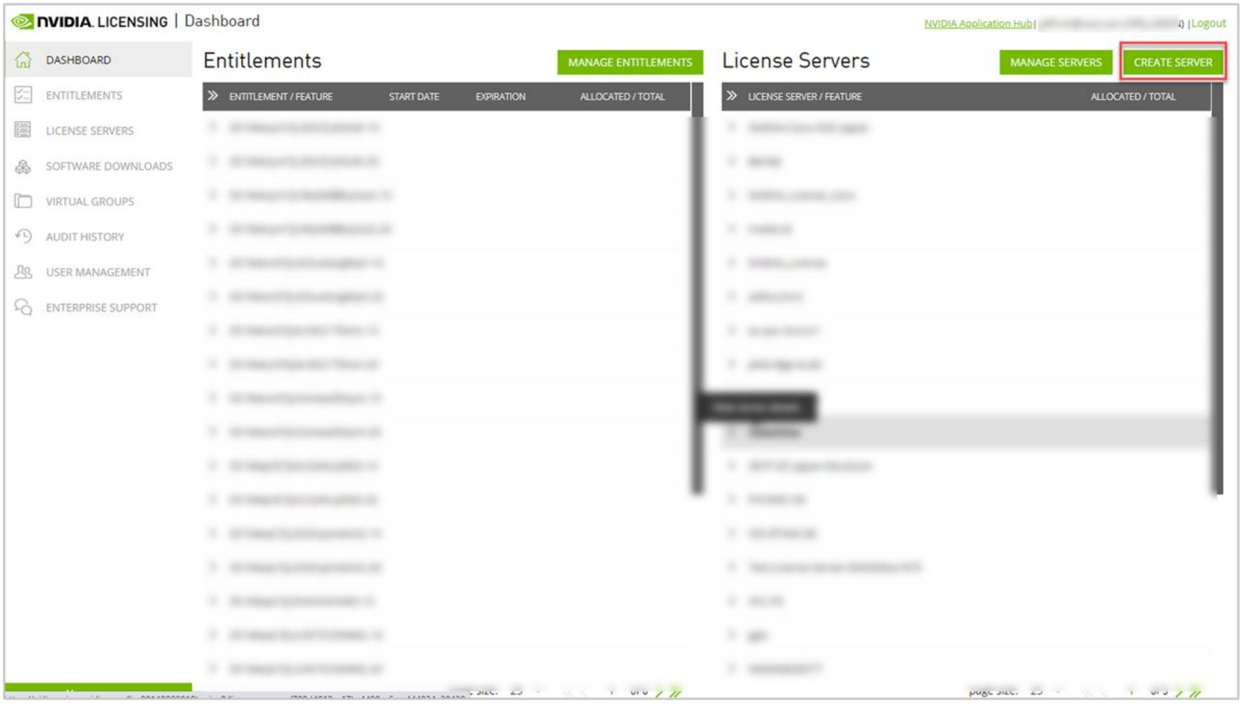
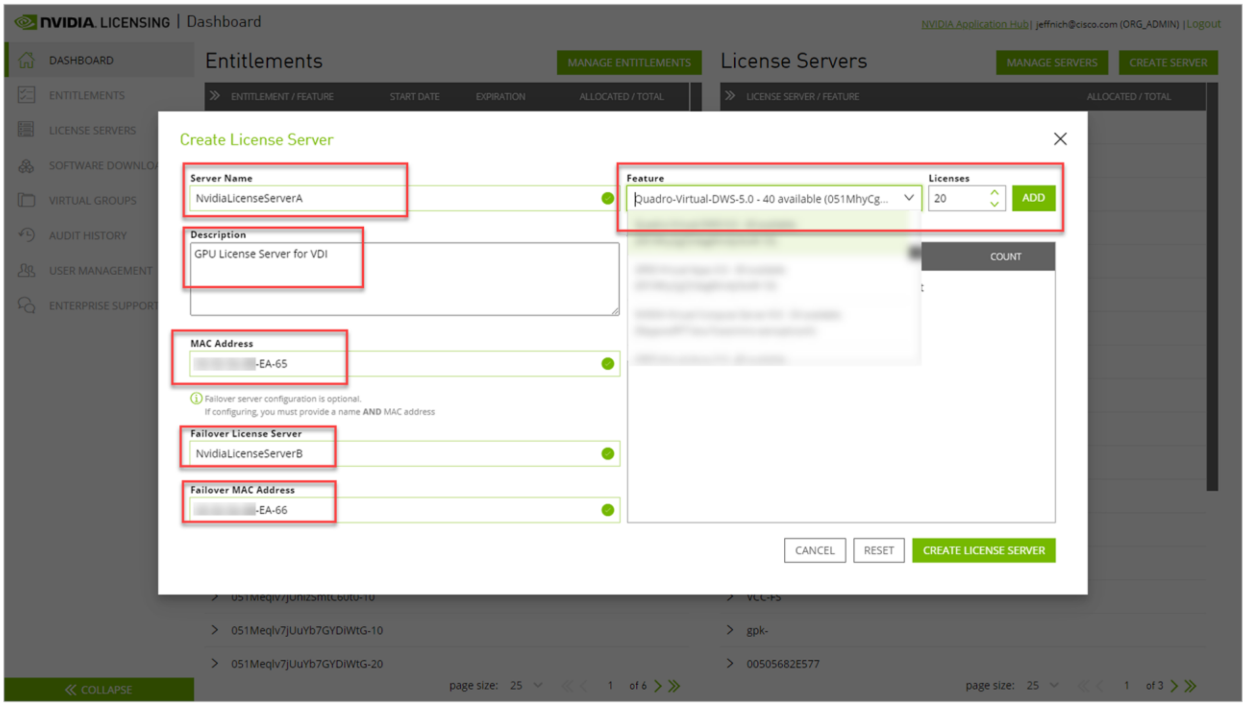


Figure 25.
Creating the license server



Create License Server

Server Name: NvidiaLicenseServerA

Description: GPU License Server for VDI

MAC Address: [redacted]-EA-65

Failover server configuration is optional.
If configuring, you must provide a name **AND** MAC address

Failover License Server: NvidiaLicenseServerB

Failover MAC Address: [redacted]-EA-66

Feature: Quadro-Virtual-DWS-5.0 - 40 available (051MhyCg...

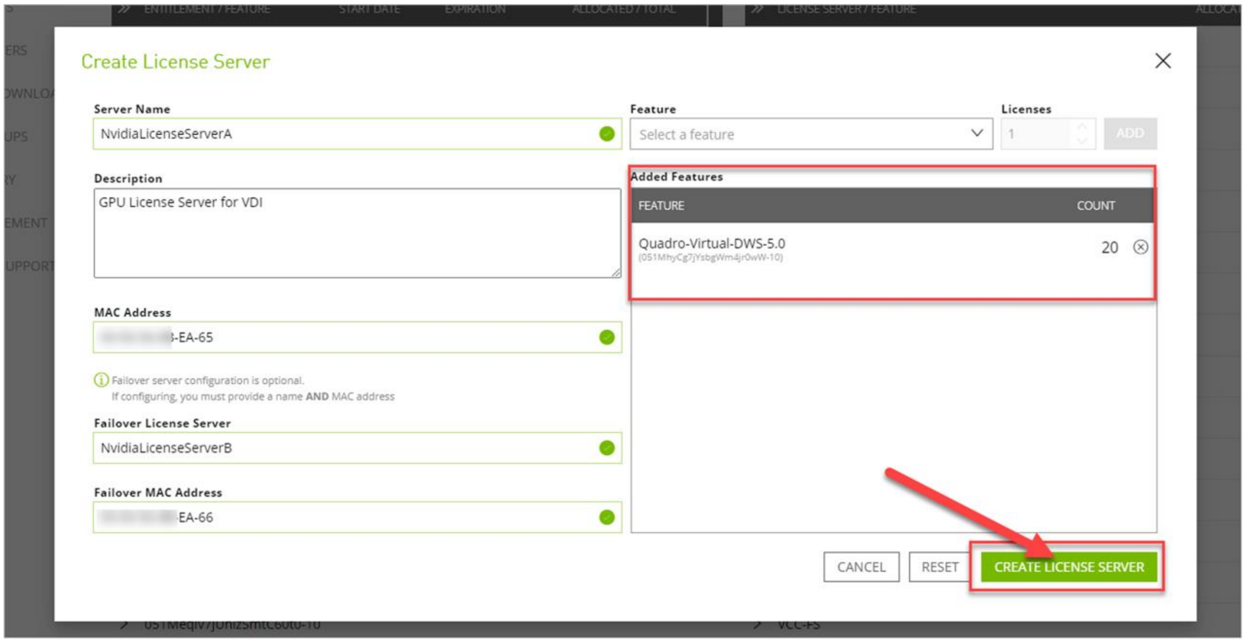
Licenses: 20

051Meqlv7jUuYb7GYDiWtG-10
051Meqlv7jUuYb7GYDiWtG-20
gpk-
00505682E577

page size: 25 1 of 6

page size: 25 1 of 3

Figure 26.
Configuring the license server



Create License Server

Server Name: NvidiaLicenseServerA

Description: GPU License Server for VDI

MAC Address: [redacted]-EA-65

Failover server configuration is optional.
If configuring, you must provide a name **AND** MAC address

Failover License Server: NvidiaLicenseServerB

Failover MAC Address: [redacted]-EA-66

Feature: Select a feature

Licenses: 1

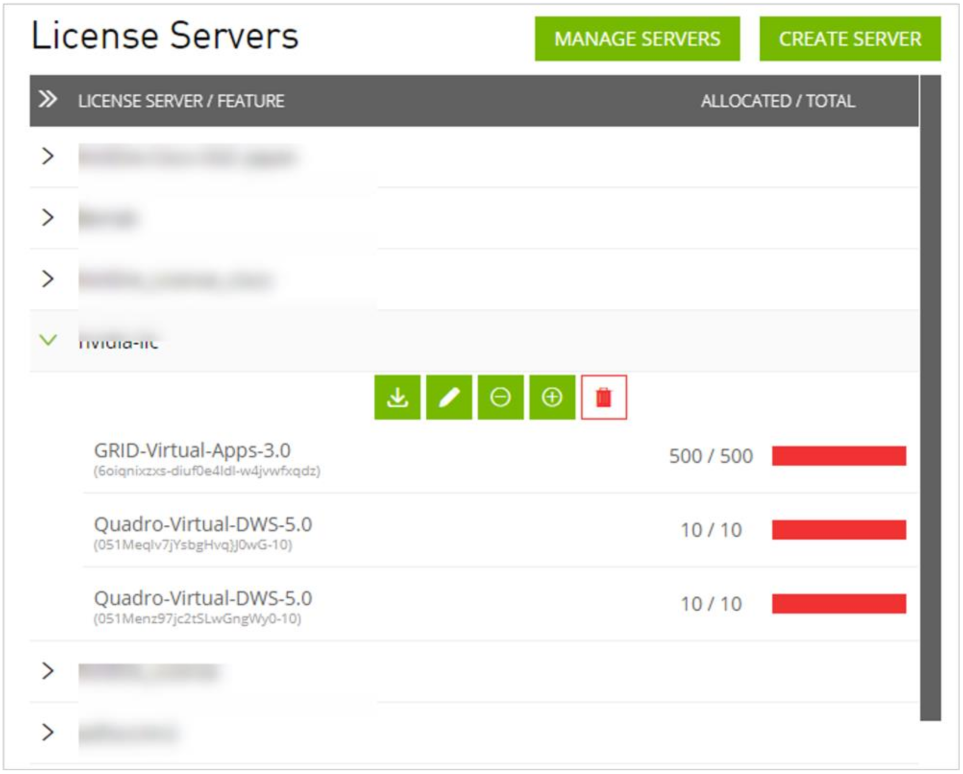
Added Features

FEATURE	COUNT
Quadro-Virtual-DWS-5.0 (051MhyCg7jYsbgWm4jr0wW-10)	20

Figure 27.
Adding the licenses and creating the server

7. Find your license server under License Servers.

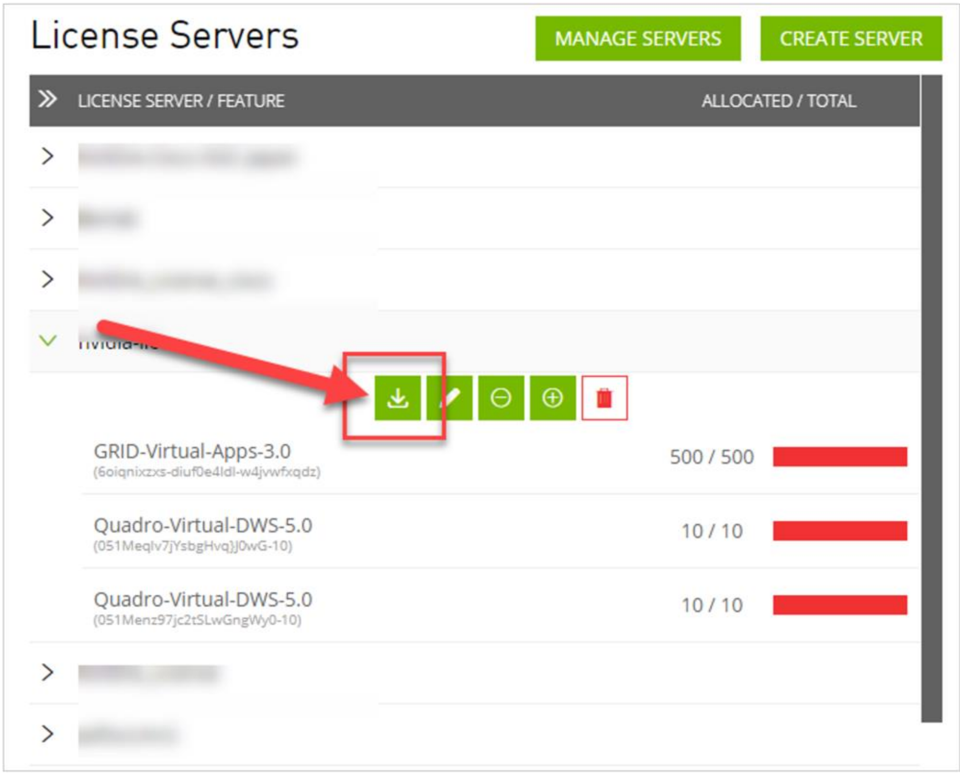
8. Click your license server ID (Figure 28).



LICENSE SERVER / FEATURE	ALLOCATED / TOTAL
[redacted]	
[redacted]	
[redacted]	
nvidia-lic	
GRID-Virtual-Apps-3.0 (6oiqnixzxs-diuf0e4ldl-w4jvwfxqdz)	500 / 500
Quadro-Virtual-DWS-5.0 (051Meqlv7jYsbgHvqJJ0wG-10)	10 / 10
Quadro-Virtual-DWS-5.0 (051Menz97jc2tSLwGngWy0-10)	10 / 10
[redacted]	
[redacted]	

Figure 28.
Selecting the license server ID

9. Click Download License File and save the .bin file to your license server (Figure 29).



LICENSE SERVER / FEATURE	ALLOCATED / TOTAL
[redacted]	
[redacted]	
[redacted]	
nvidia-lic	
GRID-Virtual-Apps-3.0 (6oiqnixzxs-diuf0e4ldl-w4jvwfxqdz)	500 / 500
Quadro-Virtual-DWS-5.0 (051Meqlv7jYsbgHvqJJ0wG-10)	10 / 10
Quadro-Virtual-DWS-5.0 (051Menz97jc2tSLwGngWy0-10)	10 / 10
[redacted]	
[redacted]	

Figure 29.
Saving the .bin file

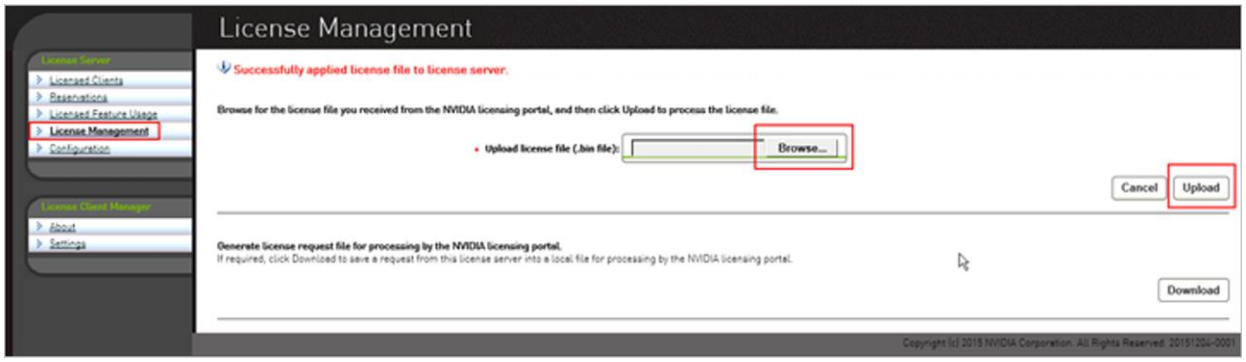
Note: The .bin file must be uploaded to your local license server within 24 hours of its generation. Otherwise, you will need to generate a new .bin file.

10. On the local license server, browse to https://<FQDN>:8080/licserver to display the License Server Configuration page.

11. Click License Management in the left pane.

    Click Browse to locate your recently downloaded .bin license file. Select the .bin file and click OK.

    Click Upload. The message "Successfully applied license file to license server" should appear on the screen (Figure 30). The features are available (Figure 31).



Successfully applied license file to license server.

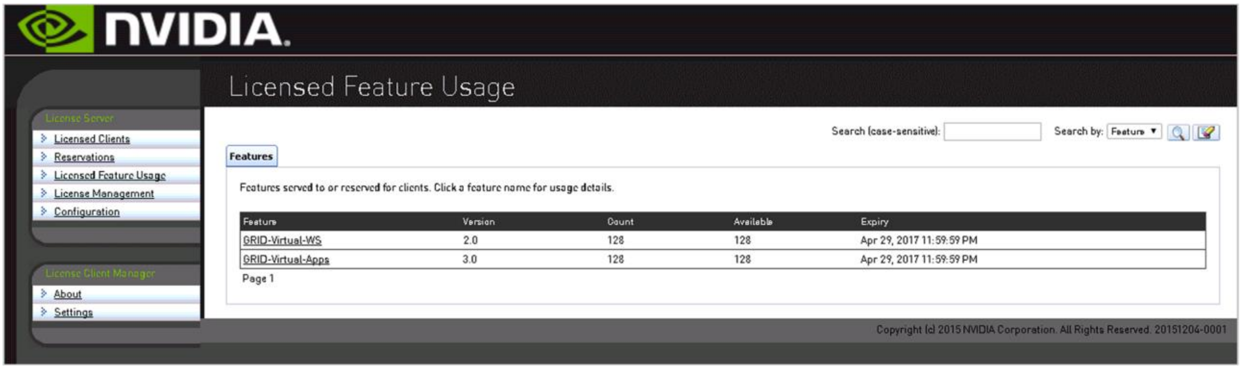
Browse for the license file you received from the NVIDIA licensing portal, and then click Upload to process the license file.

• Upload license file (.bin file):

Generate license request file for processing by the NVIDIA licensing portal.
If required, click Download to save a request from this license server into a local file for processing by the NVIDIA licensing portal.

Copyright (c) 2015 NVIDIA Corporation. All Rights Reserved. 20151204-0001

Figure 30.
License file successfully applied



Features served to or reserved for clients. Click a feature name for usage details.

Feature	Version	Count	Available	Expiry
GRID-Virtual-WS	2.0	128	128	Apr 29, 2017 11:59:59 PM
GRID-Virtual-Apps	3.0	128	128	Apr 29, 2017 11:59:59 PM

Page 1

Copyright (c) 2015 NVIDIA Corporation. All Rights Reserved. 20151204-0001

Figure 31.
NVIDIA license server available for use

Install NVIDIA GRID Software on the VMware ESX Host and Microsoft Windows Virtual Machine

This section summarizes the installation process for configuring an ESXi host and virtual machine for vGPU support. Figure 32 shows the components used for vGPU support.

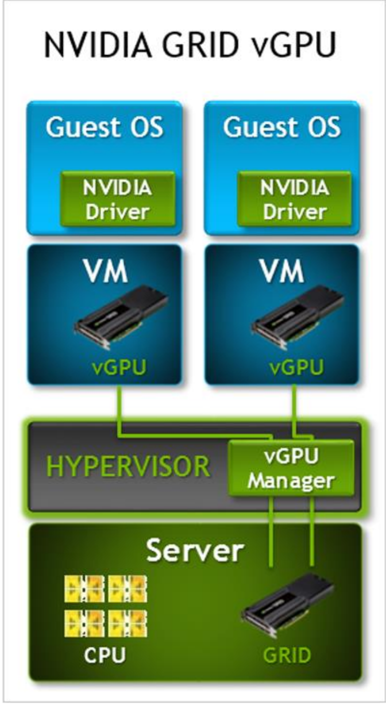
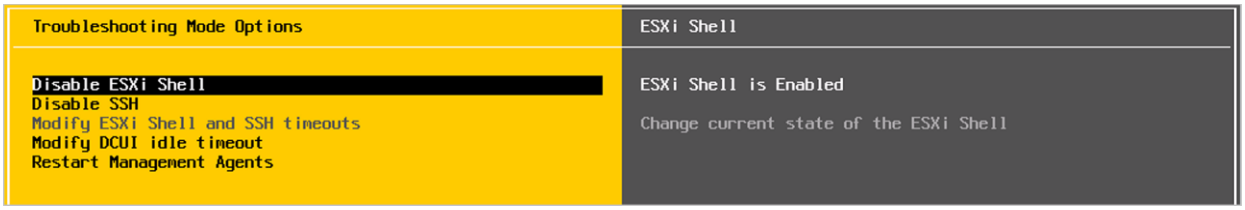


Figure 32.
NVIDIA GRID vGPU components

1. Download the NVIDIA GRID GPU driver pack for VMware vSphere ESXi 6.7.
2. Enable the ESXi shell and the Secure Shell (SSH) protocol on the vSphere host from the Troubleshooting Mode Options menu of the vSphere Configuration Console (Figure 33).



```
Troubleshooting Mode Options                              ESXi Shell

Disable ESXi Shell                                        ESXi Shell is Enabled
Disable SSH
Modify ESXi Shell and SSH timeouts                        Change current state of the ESXi Shell
Modify DCUI idle timeout
Restart Management Agents
```

Figure 33.
VMware ESXi configuration console

3. Upload the NVIDIA driver (VIB file) to the /tmp directory on the ESXi host using a tool such as WinSCP. (Shared storage is preferred if you are installing drivers on multiple servers or using the VMware Update Manager.)
4. Log in as **root** to the vSphere console through SSH using a tool such as Putty.

Note: The ESXi host must be in maintenance mode for you to install the VIB module. To place the host in maintenance mode, use the command esxcli system maintenanceMode set -enable true.

5. Enter the following command to install the NVIDIA vGPU drivers:

```
esxcli software vib install --no-sig-check -v /<path>/<filename>.VIB
```

The command should return output similar to that shown here:

```
[root@C240M5-GPU:~] esxcli software vib install -v /tmp/NVIDIA-VMware-440.107-
1OEM.670.0.0.8169922.x86_64.vib --no-sig-check
```

```
Installation Result
   Message: Operation finished successfully.
   Reboot Required: false
   VIBs Installed: NVIDIA_bootbank_NVIDIA-VMware_ESXi_6.7_Host_Driver_440.107-
1OEM.670.0.0.8169922
   VIBs Removed:
   VIBs Skipped:
```

Note: Although the display shows "Reboot Required: false," a reboot is necessary for the VIB file to load and for xorg to start.

6. Exit the ESXi host from maintenance mode and reboot the host by using the vSphere Web Client or by entering the following commands:

```
#esxcli system maintenanceMode set -e false
#reboot
```

7. After the host reboots successfully, verify that the kernel module has loaded successfully using the following command:

```
esxcli software vib list | grep -i nvidia
```

The command should return output similar to that shown here:

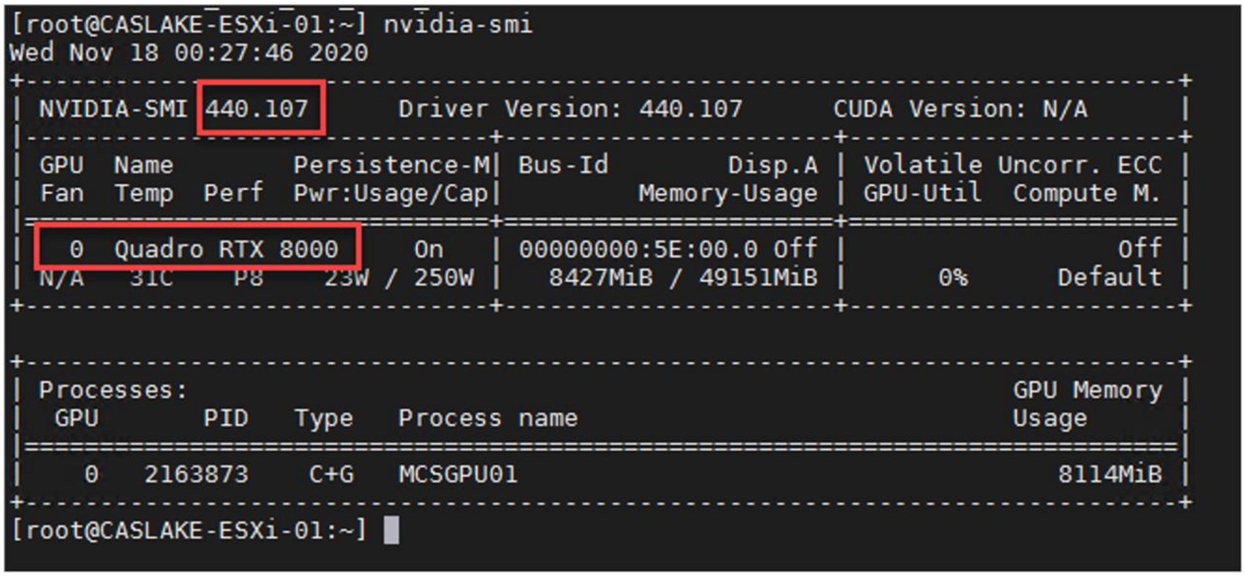
```
root@C240M5-GPU:~] esxcli software vib list | grep -i NVidia
NVIDIA-VMware_ESXi_6.7_Host_Driver  440.107-1OEM.670.0.0.8169922
NVIDIA              VMwareAccepted    2020-07-27
```

Note: See the VMware knowledge base article for information about removing any existing NVIDIA drivers before installing new drivers: https://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2033434.

8. Confirm GRID GPU detection on the ESXi host. To determine the status of the GPU card's CPU, the card's memory, and the amount of disk space remaining on the card, enter the following command:

```
nvidia-smi
```

You should see results as shown in Figure 34.



```
[root@CASLAKE-ESXi-01:~] nvidia-smi
Wed Nov 18 00:27:46 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 440.107      Driver Version: 440.107      CUDA Version: N/A      |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Quadro RTX 8000     On   | 00000000:5E:00.0 Off |                  Off |
| N/A   31C    P8    23W / 250W |   8427MiB / 49151MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|    0   2163873    C+G   MCSGPU01                                    8114MiB |
+-----------------------------------------------------------------------------+
[root@CASLAKE-ESXi-01:~]
```

Figure 34.
VMware ESX SSH console report for GPU RTX 8000 and 6000 card detection on Cisco UCS C240 M5 server

Note: The NVIDIA System Management Interface (SMI) also allows GPU monitoring using the following command (this command adds a loop, automatically refreshing the display): nvidia-smi –l.

NVIDIA RTX 6000 and 8000 profile specifications

The Tesla P6 and P40 cards have a single physical GPU, and the Tesla M10 card has multiple physical GPUs. Each physical GPU can support several different types of vGPU. Each type of vGPU has a fixed amount of frame buffer space, a fixed number of supported display heads, and a fixed maximum resolution, and each is targeted at a different class of workload. Table 5 lists the vGPU types supported by GRID GPUs.

For more information, see https://www.nvidia.com/object/grid-enterprise-resources.html.

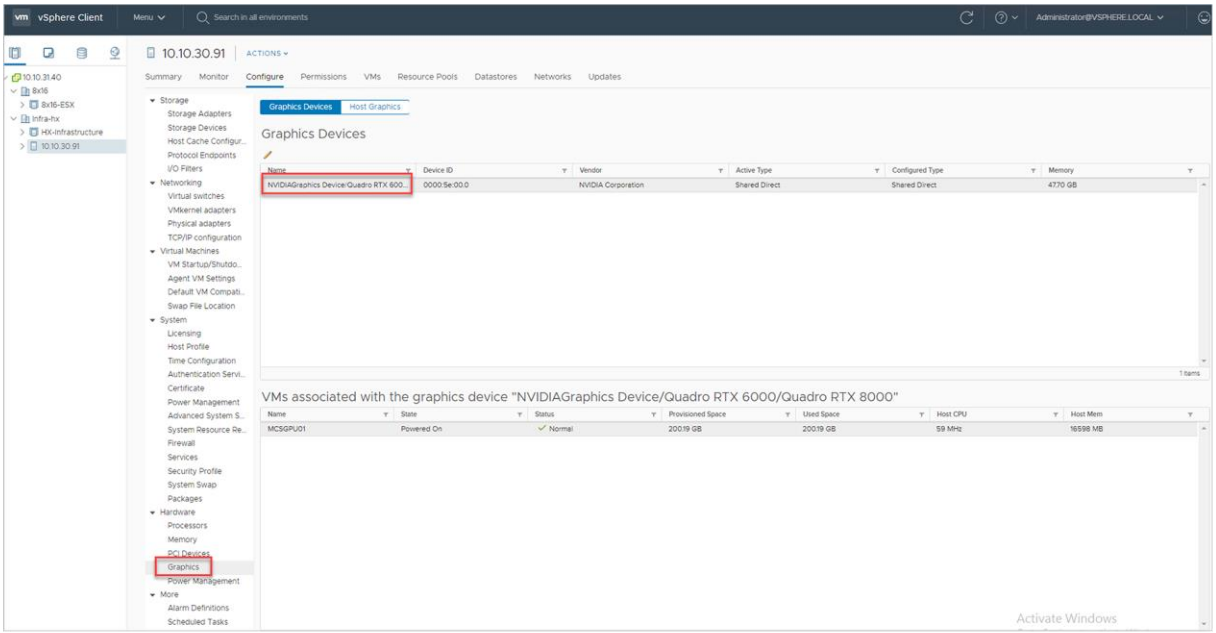
Table 5. User profile specifications for NVIDIA RTX 8000 and 6000 cards

Specs	T4	RTX 6000	RTX 8000
GPUs /Board (Architecture)	1 (Turing)	1 (Turing)	1 (Turing)
CUDA Cores	2,560	4,608	4,608
Tensor Cores	320	576	576
RT Cores	40	72	72
Memory Size	16 GB GDDR6	24 GB GDDR6	48 GB GDDR6
VGPU Profiles	1 GB, 2 GB, 4 GB, 8 GB, 16 GB	1 GB, 2 GB, 3 GB, 4 GB, 6 GB, 8 GB, 12 GB, 24 GB	1 GB, 2 GB, 3 GB, 4 GB, 6 GB, 8 GB, 12 GB,16GB, 24 GB, 48 GB
Form Factor	PCIe 3.0 Single Slot	PCIe 3.0 Dual Slot	PCIe 3.0 Dual Slot
Power	70W	250W	250W
Thermal	Passive	Passive	Passive

Prepare a virtual machine for vGPU support

Use the following procedure to create the virtual machine that will later be used as the VDI base image.

1. In the vSphere Web Client, for the host, go to Settings > PCI Devices and make GPUs available for passthrough (Figure 35).



Graphics Devices

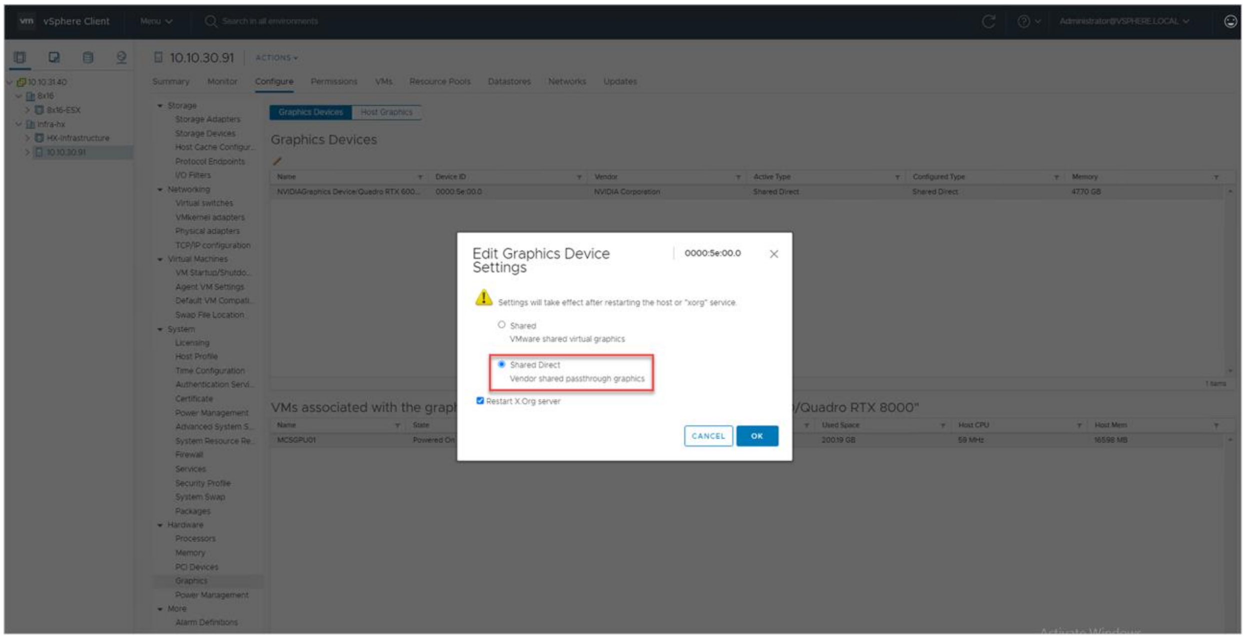
Name	Device ID	Vendor	Active Type	Configured Type	Memory
NVIDIAGraphics Device/Quadro RTX 600...	0000:5e:00.0	NVIDIA Corporation	Shared Direct	Shared Direct	47.70 GB

VMs associated with the graphics device "NVIDIAGraphics Device/Quadro RTX 6000/Quadro RTX 8000"

Name	State	Status	Provisioned Space	Used Space	Host CPU	Host Mem
MCSGPU01	Powered On	✓ Normal	200.19 GB	200.19 GB	59 MHz	16598 MB

Figure 35.
Make GPUs available for passthrough

2. Select the ESXi host and click the Configure tab. From list of options on the left, select Graphics.
3. Click Edit Host Graphics Settings.
4. Select “Shared direct (Vendor Shared passthrough graphics)” (Figure 36). A reboot is required for the changes to take effect.



Graphics Devices

Name	Device ID	Vendor	Active Type	Configured Type	Memory
NVIDIAGraphics Device/Quadro RTX 600...	0000:5e:00.0	NVIDIA Corporation	Shared Direct	Shared Direct	47.70 GB

Edit Graphics Device Settings | 0000:5e:00.0

Settings will take effect after restarting the host or "xorg" service.

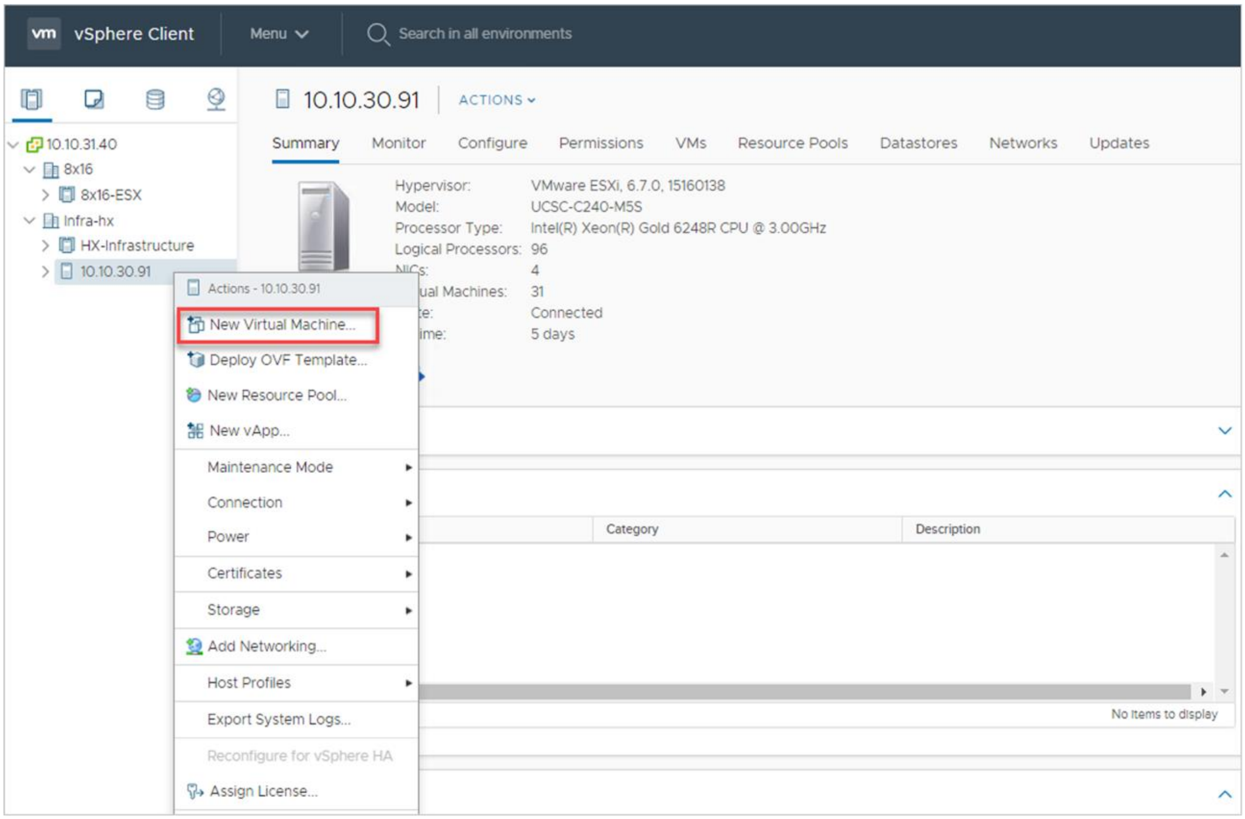
○ Shared
VMware shared virtual graphics

● Shared Direct
Vendor shared passthrough graphics

☑ Restart X.Org server

Figure 36.
Edit host graphics settings

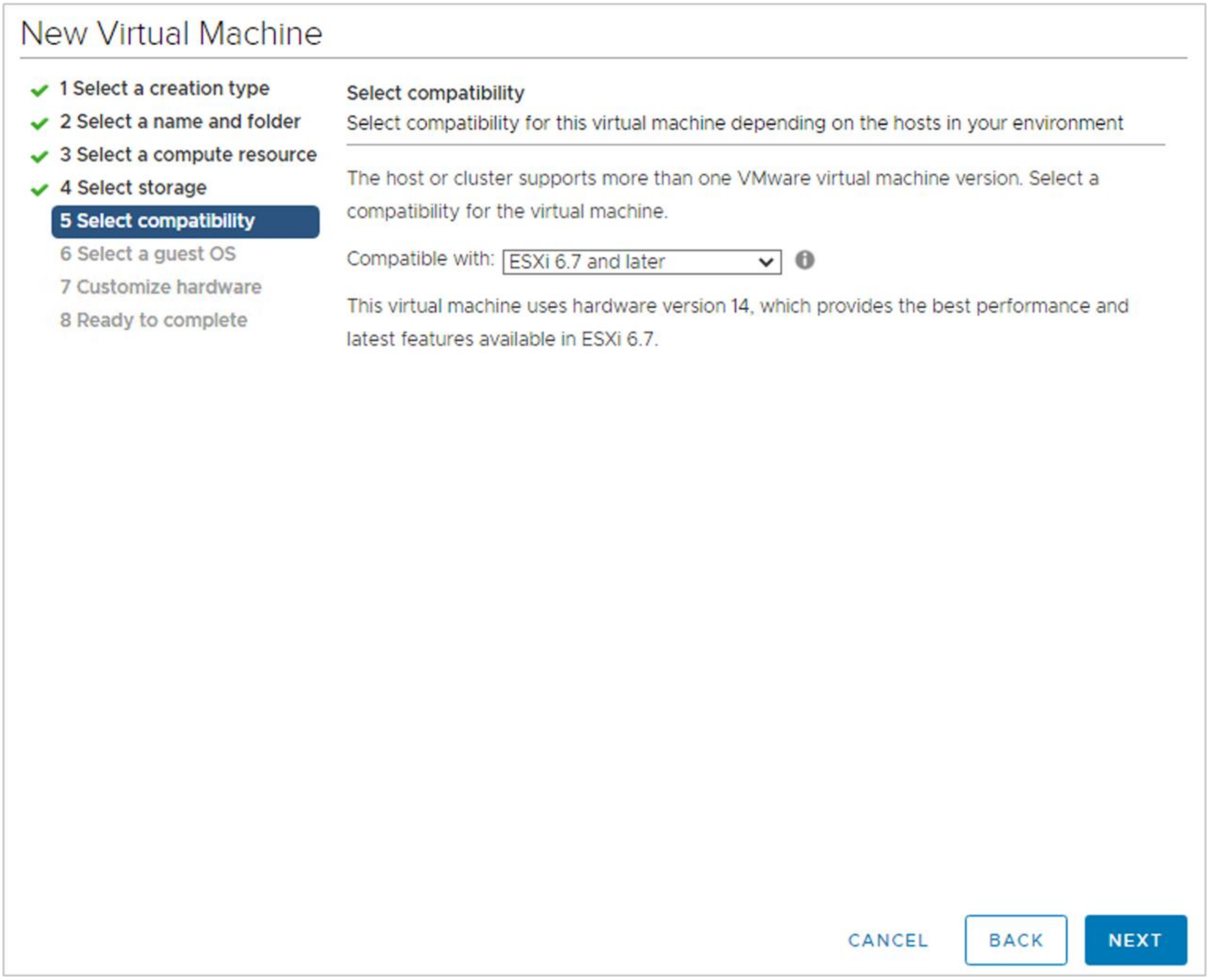
5. Using the vSphere Web Client, create a new virtual machine. To do this, right-click a host or cluster and choose New Virtual Machine. Follow the New Virtual Machine wizard. Unless another configuration is specified, select the configuration settings appropriate for your environment (Figure 37).



Hypervisor: VMware ESXi, 6.7.0, 15160138
Model: UCSC-C240-M5S
Processor Type: Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz
Logical Processors: 96
NICs: 4
ual Machines: 31
te: Connected
ime: 5 days

Figure 37.
Creating a new virtual machine in the VMware vSphere Web Client

6. Choose “ESXi 6.0 and later” from the “Compatible with” drop-down menu to use the latest features, including the mapping of shared PCI devices, which is required for the vGPU feature (Figure 38). “ESXi 6.5 and later” is selected for this document. This selection provides the latest features available in ESXi 6.5 and virtual machine hardware Version 13.



New Virtual Machine

- ✓ 1 Select a creation type
- ✓ 2 Select a name and folder
- ✓ 3 Select a compute resource
- ✓ 4 Select storage
- 5 Select compatibility
- 6 Select a guest OS
- 7 Customize hardware
- 8 Ready to complete

Select compatibility

Select compatibility for this virtual machine depending on the hosts in your environment

The host or cluster supports more than one VMware virtual machine version. Select a compatibility for the virtual machine.

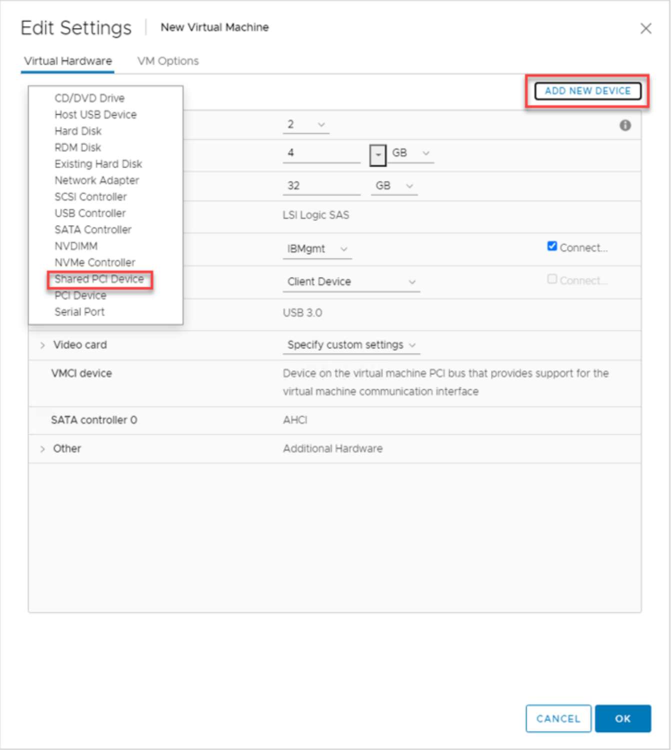
Compatible with: ESXi 6.7 and later

This virtual machine uses hardware version 14, which provides the best performance and latest features available in ESXi 6.7.

Figure 38.
Selecting virtual machine hardware Version 11 or later

7. In customizing the hardware of the new virtual machine, add a new shared PCI device, select the appropriate GPU profile, and reserve all virtual machine memory (Figures 39 and 40).

Note: If you are creating a new virtual machine and using the vSphere Web Client's virtual machine console functions, the mouse will not be usable in the virtual machine until after both the operating system and VMware Tools have been installed. If you cannot use the traditional vSphere Client to connect to the virtual machine, do not enable the NVIDIA GRID vGPU at this time.



2

4 GB

32 GB

LSI Logic SAS

IBMgmt ☑ Connect...

Client Device □ Connect...

USB 3.0

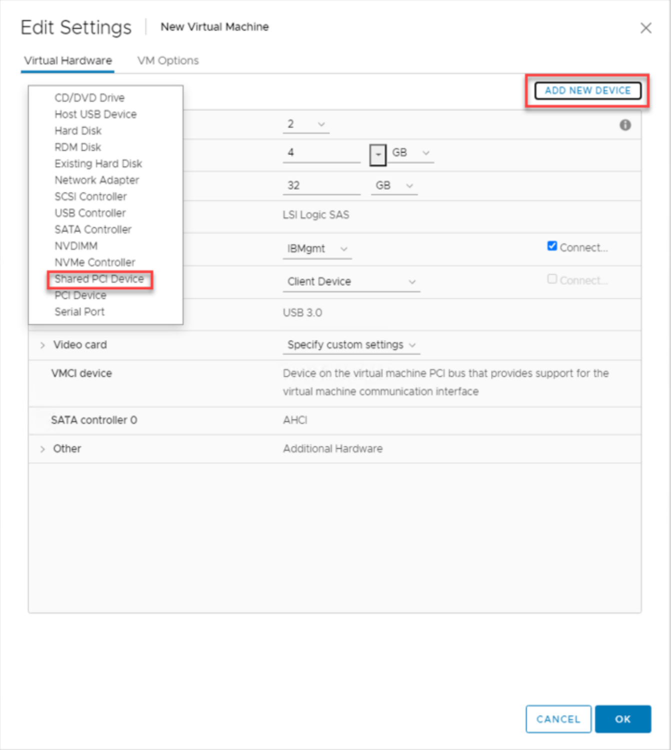
Video card: Specify custom settings

VMCI device: Device on the virtual machine PCI bus that provides support for the virtual machine communication interface

SATA controller 0: AHCI

Other: Additional Hardware

Figure 39.
Adding a shared PCI device to the virtual machine to attach the GPU profile



2

4 GB

32 GB

LSI Logic SAS

IBMgmt ☑ Connect...

Client Device □ Connect...

USB 3.0

Video card: Specify custom settings

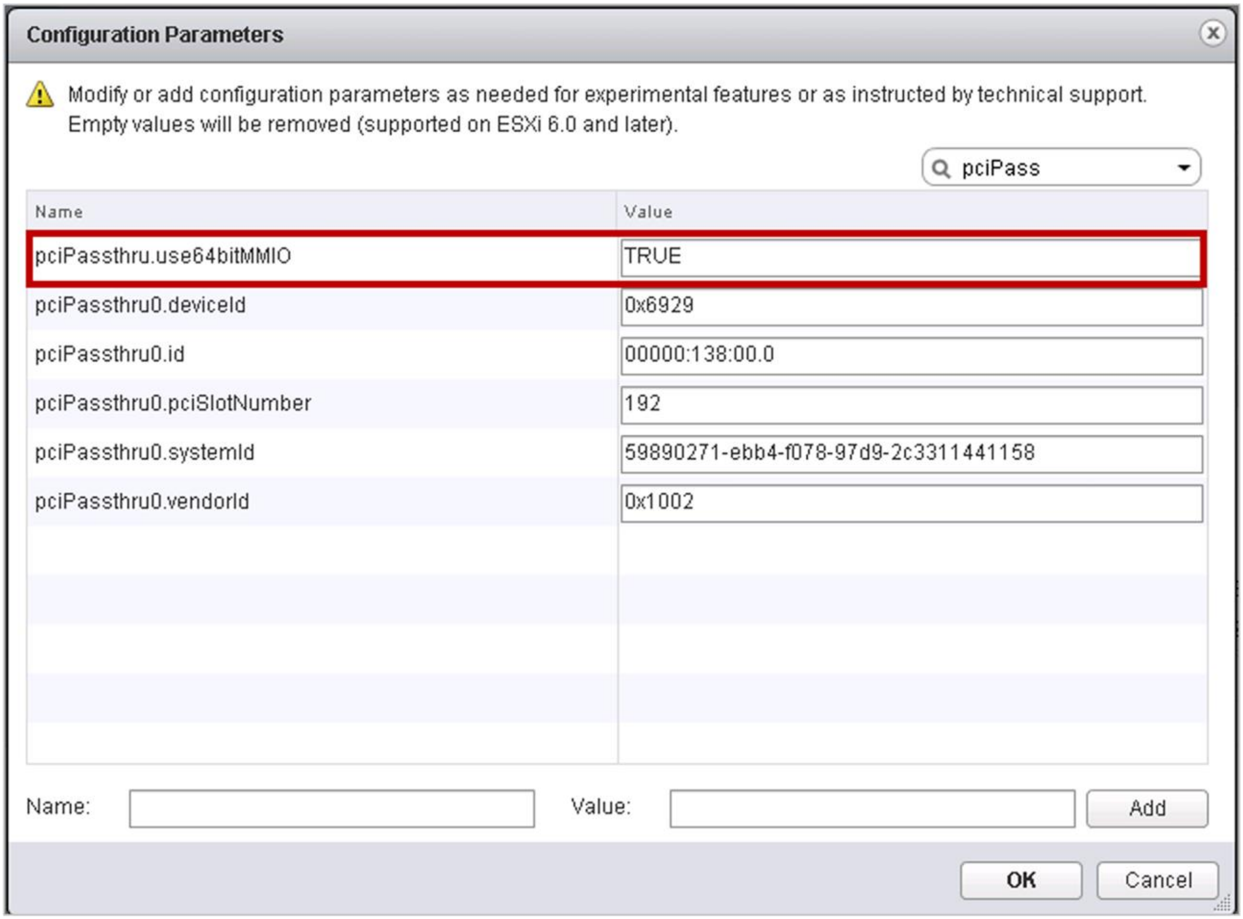
VMCI device: Device on the virtual machine PCI bus that provides support for the virtual machine communication interface

SATA controller 0: AHCI

Other: Additional Hardware

Figure 40.
Attaching the GPU profile to a shared PCI device

8. Go to the VM Options tab. From the drop-down menu under Advanced > Configuration Parameters, choose Edit Configuration. Add the name **pciPassthru.use64bitMIMO** and set the value to TRUE (Figure 41).



Configuration Parameters

Modify or add configuration parameters as needed for experimental features or as instructed by technical support. Empty values will be removed (supported on ESXi 6.0 and later).

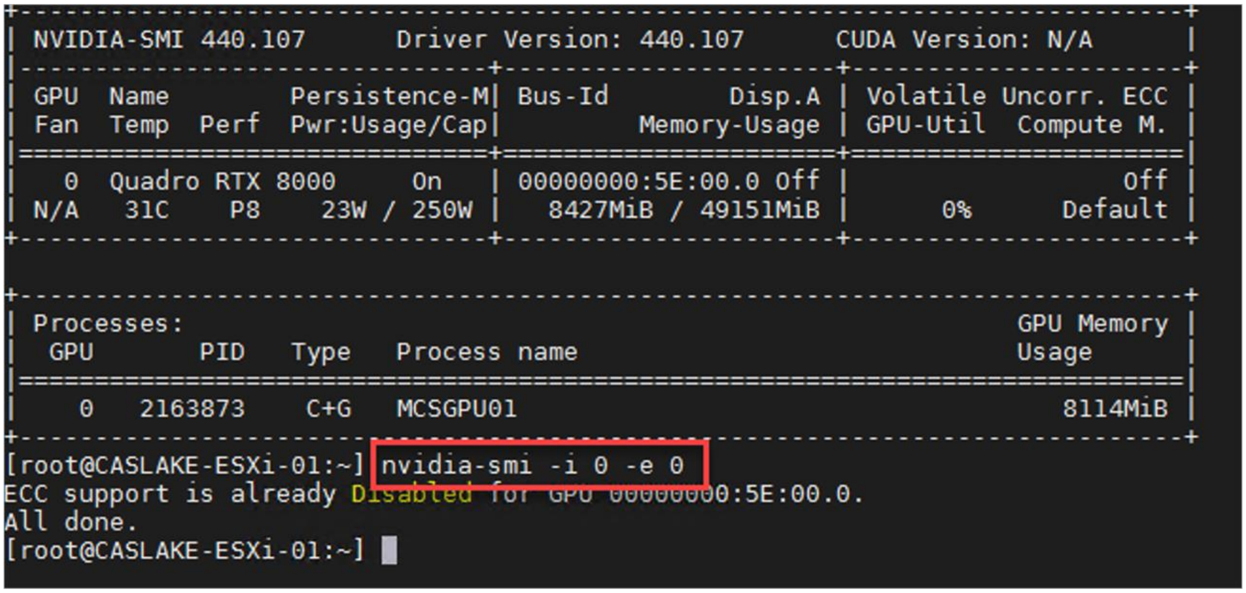
Name	Value
pciPassthru.use64bitMMIO	TRUE
pciPassthru0.deviceId	0x6929
pciPassthru0.id	00000:138:00.0
pciPassthru0.pciSlotNumber	192
pciPassthru0.systemId	59890271-ebb4-f078-97d9-2c3311441158
pciPassthru0.vendorId	0x1002

Figure 41.
Add pciPassthru.use64bitMIMO = TRUE

Note: A virtual machine with vGPU assigned will fail to start if ECC is enabled. As a workaround, disable ECC by using the process shown here (Figure 42):

```
#nvidia-smi -i 0 -e 0
#nvidia-smi -i 1 -e 0 (if more than one GPU is installed)
```

Note: Use -i to target a specific GPU. If two card are installed in a server, run the command twice, where 0 and 1 are used to distinguish the two GPU cards.



```
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 440.107      Driver Version: 440.107      CUDA Version: N/A      |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Quadro RTX 8000     On   | 00000000:5E:00.0 Off |                  Off |
| N/A   31C    P8    23W / 250W |   8427MiB / 49151MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|    0   2163873    C+G   MCSGPU01                                    8114MiB |
+-----------------------------------------------------------------------------+
[root@CASLAKE-ESXi-01:~] nvidia-smi -i 0 -e 0
ECC support is already Disabled for GPU 00000000:5E:00.0.
All done.
[root@CASLAKE-ESXi-01:~]
```

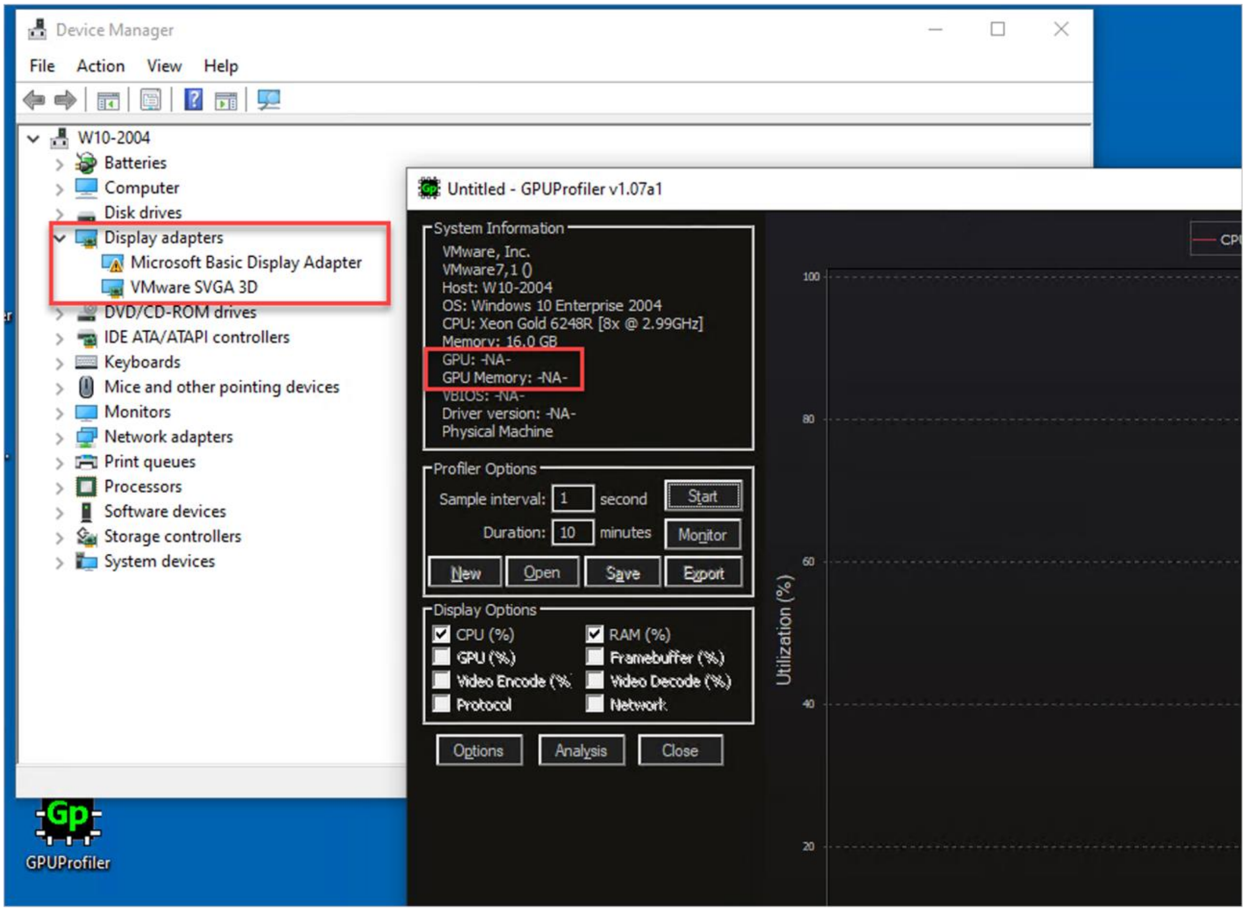
Figure 42.
Disable ECC

9. Install and configure Microsoft Windows on the virtual machine:
    - For higher-end vGPU desktops such as those used with the RTX 6000 and 8000 cards, configure the number of vCPUs and amount of RAM according to the GPU profile selected (the configuration described in this document uses eight vCPU and 16 GB of RAM).
    - Install VMware Tools.
    - Join the virtual machine to the Microsoft Active Directory domain.
    - Install or upgrade VMware Horizon Agent (in this document, Version 7.12 is used).

Install the NVIDIA vGPU software driver

Use the following procedure to install the NVIDIA GRID vGPU drivers on the desktop virtual machine. To fully enable vGPU operation, the NVIDIA driver must be installed.

Before the NVIDIA driver is installed on the guest virtual machine, the Device Manager shows the standard VGA graphics adapter (Figure 43).



```
W10-2004
  Batteries
  Computer
  Disk drives
  Display adapters
    Microsoft Basic Display Adapter
    VMware SVGA 3D
  DVD/CD-ROM drives
  IDE ATA/ATAPI controllers
  Keyboards
  Mice and other pointing devices
  Monitors
  Network adapters
  Print queues
  Processors
  Software devices
  Storage controllers
  System devices
```

```
System Information
VMware, Inc.
VMware7,1 ()
Host: W10-2004
OS: Windows 10 Enterprise 2004
CPU: Xeon Gold 6248R [8x @ 2.99GHz]
Memory: 16.0 GB
GPU: -NA-
GPU Memory: -NA-
VBIOS: -NA-
Driver version: -NA-
Physical Machine
```

Profiler Options

Sample interval: 1 second

Duration: 10 minutes

Display Options

☑ CPU (%) ☑ RAM (%)

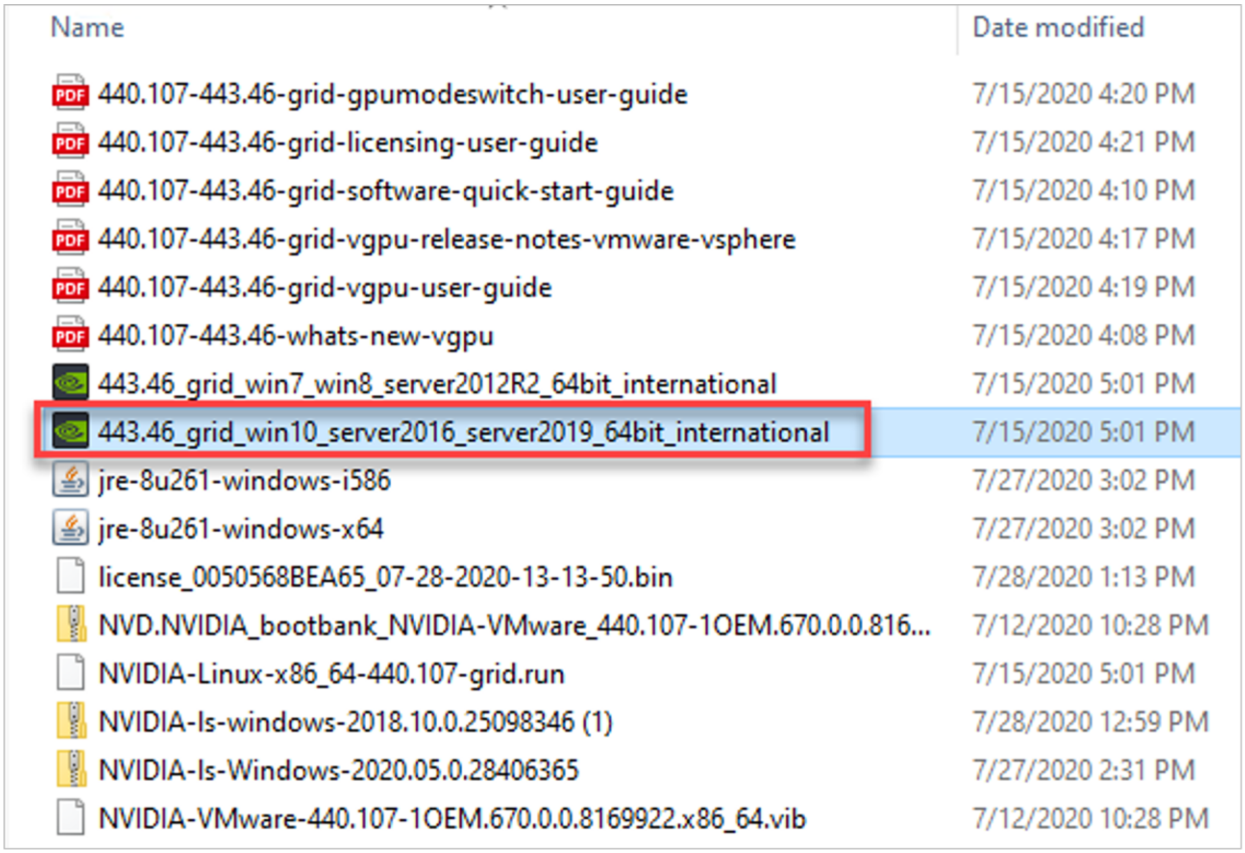
☐ GPU (%) ☐ Framebuffer (%)

☐ Video Encode (% ☐ Video Decode (%)

☐ Protocol ☐ Network

Figure 43.
Device Manager before the NVIDIA driver is installed

1. Copy the Windows drivers from the NVIDIA GRID vGPU driver pack, downloaded earlier, to the master virtual machine.
2. Copy the 32- or 64-bit NVIDIA Windows driver from the vGPU driver pack to the desktop virtual machine and run setup.exe (Figure 44).



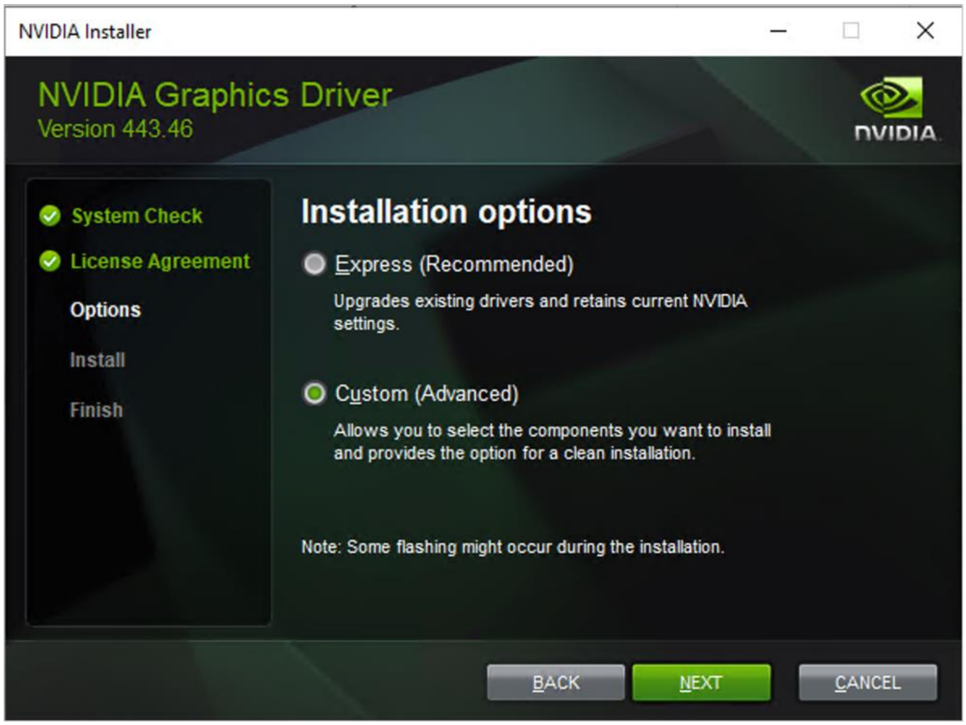
Name	Date modified
440.107-443.46-grid-gpumodeswitch-user-guide	7/15/2020 4:20 PM
440.107-443.46-grid-licensing-user-guide	7/15/2020 4:21 PM
440.107-443.46-grid-software-quick-start-guide	7/15/2020 4:10 PM
440.107-443.46-grid-vgpu-release-notes-vmware-vsphere	7/15/2020 4:17 PM
440.107-443.46-grid-vgpu-user-guide	7/15/2020 4:19 PM
440.107-443.46-whats-new-vgpu	7/15/2020 4:08 PM
443.46_grid_win7_win8_server2012R2_64bit_international	7/15/2020 5:01 PM
443.46_grid_win10_server2016_server2019_64bit_international	7/15/2020 5:01 PM
jre-8u261-windows-i586	7/27/2020 3:02 PM
jre-8u261-windows-x64	7/27/2020 3:02 PM
license_0050568BEA65_07-28-2020-13-13-50.bin	7/28/2020 1:13 PM
NVD.NVIDIA_bootbank_NVIDIA-VMware_440.107-1OEM.670.0.0.816...	7/12/2020 10:28 PM
NVIDIA-Linux-x86_64-440.107-grid.run	7/15/2020 5:01 PM
NVIDIA-ls-windows-2018.10.0.25098346 (1)	7/28/2020 12:59 PM
NVIDIA-ls-Windows-2020.05.0.28406365	7/27/2020 2:31 PM
NVIDIA-VMware-440.107-1OEM.670.0.0.8169922.x86_64.vib	7/12/2020 10:28 PM

Figure 44.
NVIDIA driver pack

Note: The vGPU host driver and guest driver versions need to match. Do not attempt to use a newer guest driver with an older vGPU host driver or an older guest driver with a newer vGPU host driver. In addition, the vGPU driver from NVIDIA is a different driver than the GPU passthrough driver.

3. Install the graphics drivers using the Express or Custom option (Figure 45). After the installation has been completed successfully (Figure 46), restart the virtual machine.

Note: Be sure that remote desktop connections have been enabled. After this step, console access to the virtual machine may not be usable when connecting from a vSphere Client.



Installation options

○ Express (Recommended)

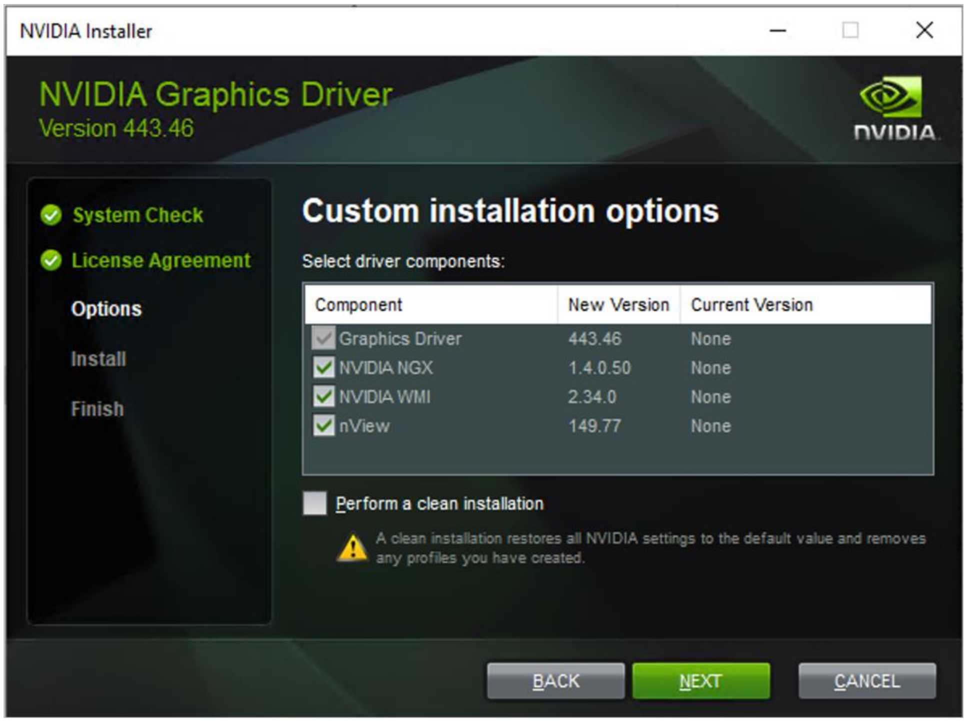
Upgrades existing drivers and retains current NVIDIA settings.

● Custom (Advanced)

Allows you to select the components you want to install and provides the option for a clean installation.

Note: Some flashing might occur during the installation.

Figure 45.
Select the Express or Custom installation option



Custom installation options

Select driver components:

Component	New Version	Current Version
☑ Graphics Driver	443.46	None
☑ NVIDIA NGX	1.4.0.50	None
☑ NVIDIA WMI	2.34.0	None
☑ nView	149.77	None

□ Perform a clean installation

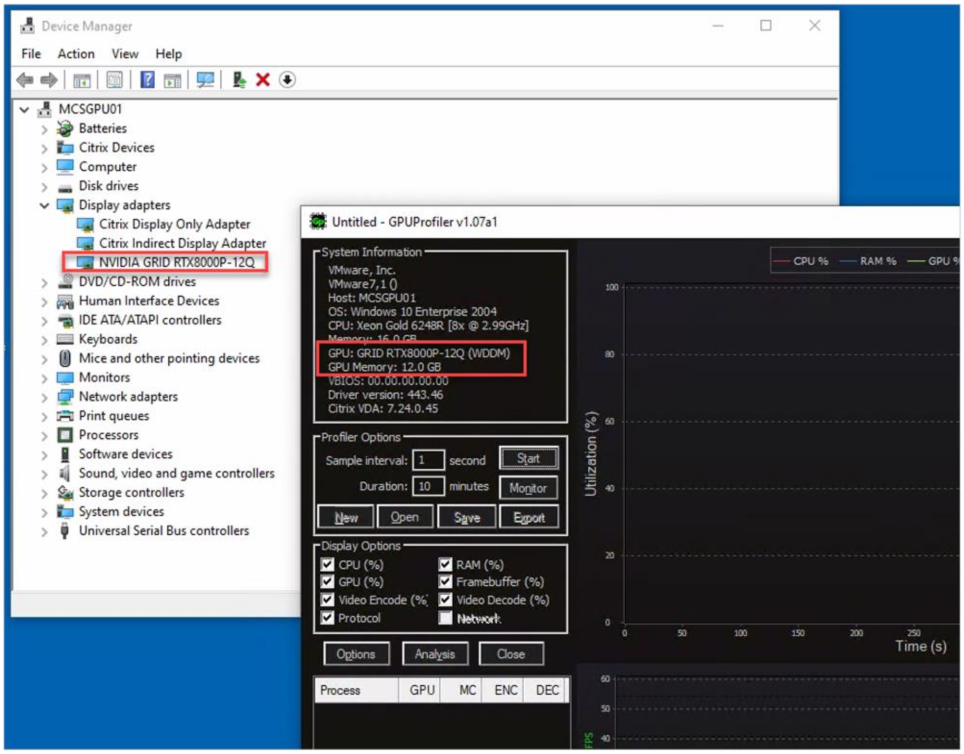
⚠ A clean installation restores all NVIDIA settings to the default value and removes any profiles you have created.

Figure 46.
Components to be installed during NVIDIA graphics driver installation process

Verify that applications are ready to support vGPU

Validate the successful installation of the graphics drivers and the vGPU device.

Open Windows Device Manager and expand the Display Adapter section. The device will reflect chosen profile (Figure 47).



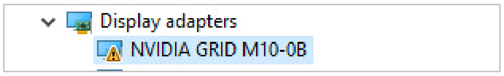
System Information
VMware, Inc.
VMware7,1 ()
Host: MCSGPU01
OS: Windows 10 Enterprise 2004
CPU: Xeon Gold 6248R [8x @ 2.99GHz]
Memory: 16.0 GB
GPU: GRID RTX8000P-12Q (WDDM)
GPU Memory: 12.0 GB
VBIOS: 00.00.00.00.00
Driver version: 443.46
Citrix VDA: 7.24.0.45

Profiler Options
Sample interval: 1 second
Duration: 10 minutes

Display Options
☑ CPU (%) ☑ RAM (%)
☑ GPU (%) ☑ Framebuffer (%)
☑ Video Encode (% ☑ Video Decode (%)
☑ Protocol □ Network

Figure 47.
Validating the driver installation

Note: If you see an exclamation point as shown here, a problem has occurred.


Display adapters
NVIDIA GRID M10-0B

The following are the most likely reasons:

- The GPU driver service is not running.
- The GPU driver is incompatible.

Configure the virtual machine for an NVIDIA GRID vGPU license

You need to point the master image to the license server so the virtual machines with vGPUs can obtain a license.

Note: The license settings persist across reboots. These settings can also be preloaded through register keys.

1. In the Microsoft Windows Control Panel, double-click NVIDIA Control Panel (Figure 48).

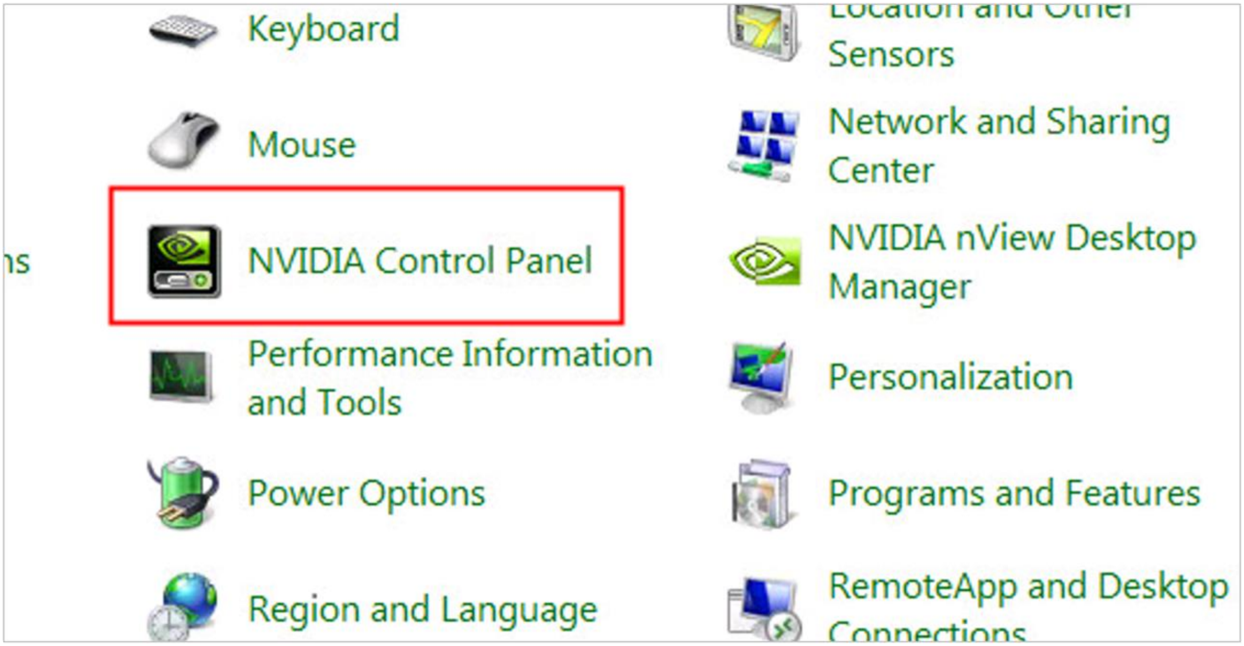
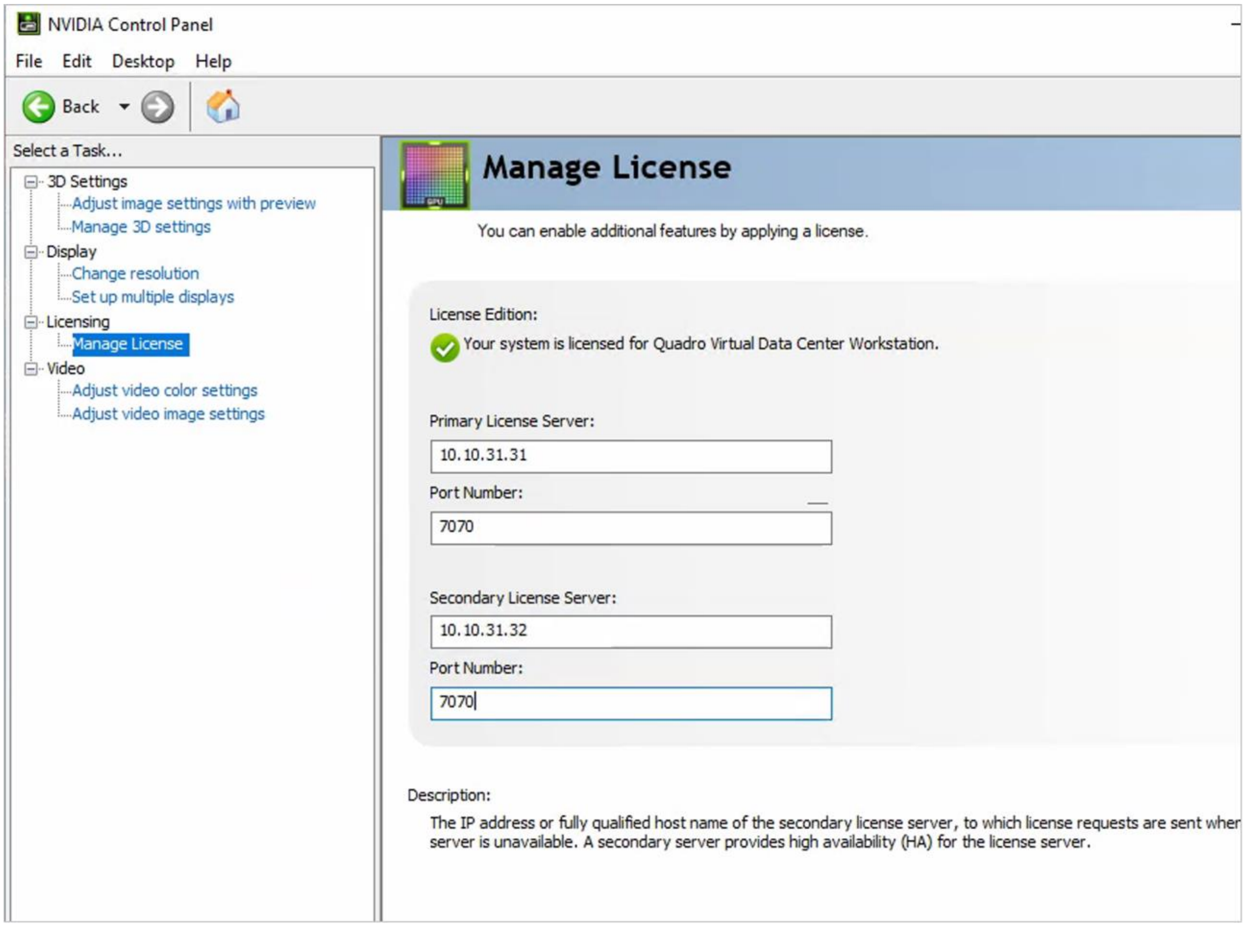

Keyboard
Location and Other Sensors
Mouse
Network and Sharing Center
ns
NVIDIA Control Panel
NVIDIA nView Desktop Manager
Performance Information and Tools
Personalization
Power Options
Programs and Features
Region and Language
RemoteApp and Desktop Connections

Figure 48.
Choosing NVIDIA Control Panel

2. Select Manage License from the left pane and enter your license server address and port (Figure 49).



Manage License

You can enable additional features by applying a license.

License Edition:
✓ Your system is licensed for Quadro Virtual Data Center Workstation.

Primary License Server: 10.10.31.31
Port Number: 7070

Secondary License Server: 10.10.31.32
Port Number: 7070

Description:
The IP address or fully qualified host name of the secondary license server, to which license requests are sent wher
server is unavailable. A secondary server provides high availability (HA) for the license server.

Figure 49.
Managing your license

3. Select Apply.

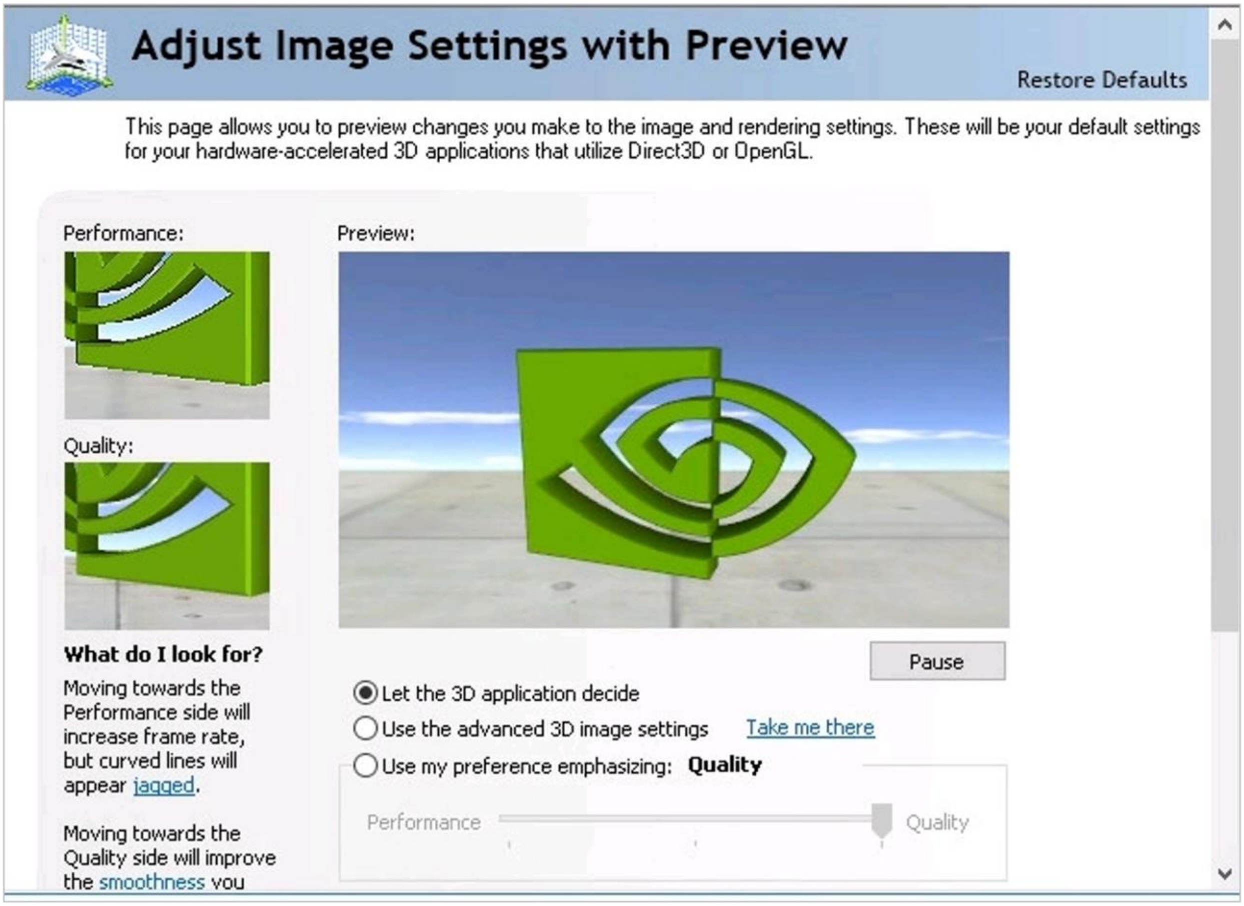
Verify vGPU deployment

After the desktops are provisioned, use the following steps to verify vGPU deployment in the VMware Horizon environment.

Verify that the NVIDIA driver is running on the desktop

Follow these steps to verify that the NVIDIA driver is running on the desktop:

1. Right-click the desktop. In the menu, choose NVIDIA Control Panel to open the control panel.
2. In the control panel, select System Information to see the vGPU that the virtual machine is using, the vGPU's capabilities, and the NVIDIA driver version that is loaded (Figure 50).



Adjust Image Settings with Preview

This page allows you to preview changes you make to the image and rendering settings. These will be your default settings for your hardware-accelerated 3D applications that utilize Direct3D or OpenGL.

What do I look for?

Moving towards the Performance side will increase frame rate, but curved lines will appear jagged.

Moving towards the Quality side will improve the smoothness you

● Let the 3D application decide

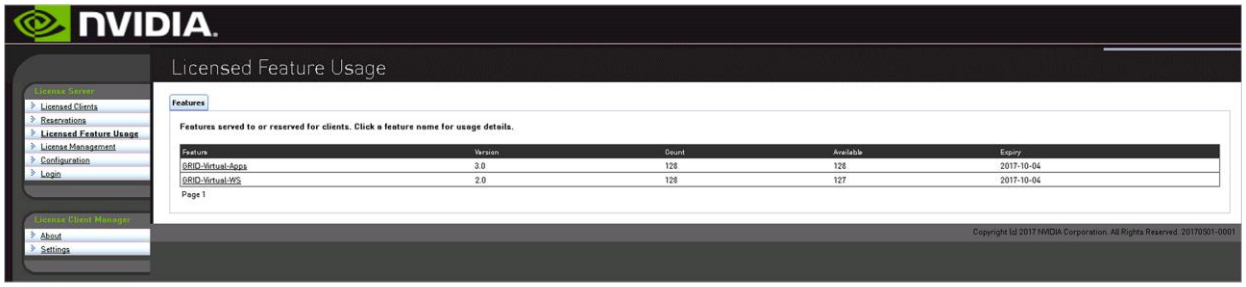
○ Use the advanced 3D image settings Take me there

○ Use my preference emphasizing: **Quality**

Figure 50.
NVIDIA Control Panel

Verify NVDIA license acquisition by desktops

A license is obtained before the user logs on to the virtual machine after the virtual machine is fully booted (Figure 51).



Licensed Feature Usage

Features served to or reserved for clients. Click a feature name for usage details.

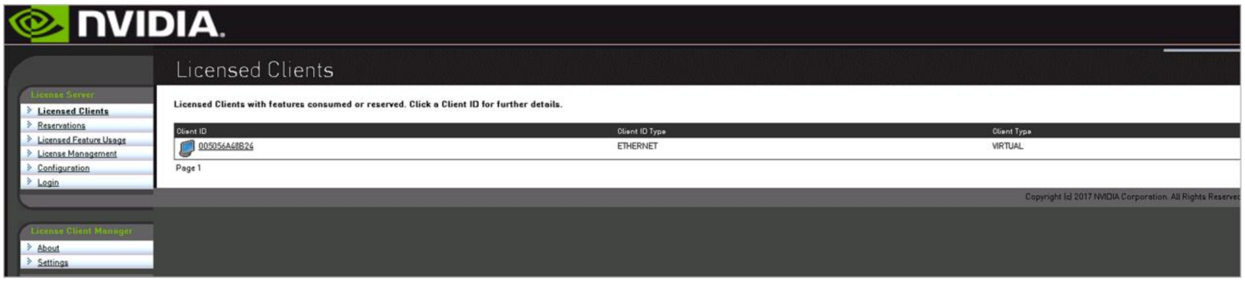
Feature	Version	Count	Available	Expiry
GRID-Virtual-Apps	3.0	128	128	2017-10-04
GRID-Virtual-WS	2.0	128	127	2017-10-04

Page 1

Copyright (c) 2017 NVIDIA Corporation. All Rights Reserved. 20170501-0001

Figure 51.
NVIDIA License Server: Licensed Feature Usage

To view the details, select Licensed Clients in the left pane (Figure 52).



Licensed Clients

Licensed Clients with features consumed or reserved. Click a Client ID for further details.

Client ID	Client ID Type	Client Type
005056A48B24	ETHERNET	VIRTUAL

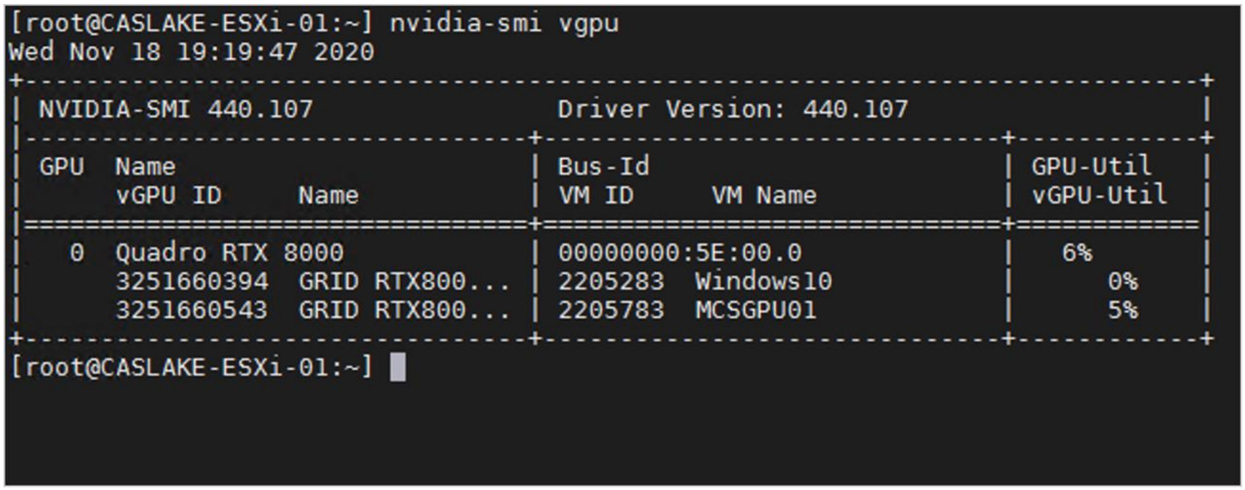
Page 1

Copyright (c) 2017 NVIDIA Corporation. All Rights Reserved

Figure 52.
NVIDIA License Server: Licensed Clients

Verify the NVDIA configuration on the host

To obtain a hostwide overview of the NVIDIA GPUs, enter the nvidia-smi command without any arguments (Figure 53).



```
[root@CASLAKE-ESXi-01:~] nvidia-smi vgpu
Wed Nov 18 19:19:47 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 440.107                 Driver Version: 440.107                  |
|-------------------------------+------------------------------+------------+
| GPU  Name                     | Bus-Id                       | GPU-Util   |
|      vGPU ID     Name         | VM ID      VM Name           | vGPU-Util  |
|===============================+==============================+============|
|   0  Quadro RTX 8000          | 00000000:5E:00.0             |   6%       |
|      3251660394  GRID RTX800... | 2205283  Windows10         |      0%    |
|      3251660543  GRID RTX800... | 2205783  MCSGPU01          |      5%    |
+-------------------------------+------------------------------+------------+
[root@CASLAKE-ESXi-01:~] █
```

Figure 53.
The nvdia-smi command output from the host with two NVIDIA P40 cards and 48 Microsoft Windows 10 desktops with P40-1B vGPU profile

Additional configurations

This section presents additional configuration options.

Install and upgrade NVIDIA drivers

The NVIDIA GRID API provides direct access to the frame buffer of the GPU, providing the fastest possible frame rate for a smooth and interactive user experience.

Use GPU acceleration for Microsoft Windows Server DirectX, Direct3D, and WPF rendering

DirectX, Direct3D, and WPF rendering are available only on servers with a GPU that supports Display Driver Interface (DDI) Version 9ex, 10, or 11.

Use the OpenGL Software Accelerator

The OpenGL Software Accelerator is a software rasterizer for OpenGL applications such as ArcGIS, Google Earth, Nehe, Maya, Blender, Voxler, CAD, and CAM. In some cases, the OpenGL Software Accelerator can eliminate the need to use graphics cards to deliver a good user experience with OpenGL applications.

Note: The OpenGL Software Accelerator is provided as is and must be tested with all applications. It may not work with some applications and is intended as a solution to try if the Windows OpenGL rasterizer does not provide adequate performance. If the OpenGL Software Accelerator works with your applications, you can use it to avoid the cost of GPU hardware.

The OpenGL Software Accelerator is provided in the Support folder on the installation media, and it is supported on all valid VDA platforms.

Try the OpenGL Software Accelerator in the following cases:

- If the performance of OpenGL applications running in virtual machines is a concern, try using the OpenGL accelerator. For some applications, the accelerator outperforms the Microsoft OpenGL software rasterizer that is included with Windows because the OpenGL accelerator uses SSE4.1 and AVX. The OpenGL accelerator also supports applications using OpenGL versions up to Version 2.1.
- For applications running on a workstation, first try the default version of OpenGL support provided by the workstation's graphics adapter. If the graphics card is the latest version, in most cases it will deliver the best performance. If the graphics card is an earlier version or does not deliver satisfactory performance, then try the OpenGL Software Accelerator.
- 3D OpenGL applications that are not adequately delivered using CPU-based software rasterization may benefit from OpenGL GPU hardware acceleration. This feature can be used on bare-metal devices and virtual machines.

Performance testing with SPECviewperf 13

After desktops were provisioned, high-end graphics solutions were tested with SPECviewperf 13 to determine the best practices for sizing high-end graphics workstations using RTX 6000 and 8000 cards. Figures 54, 55, 56, and 57 show the results.

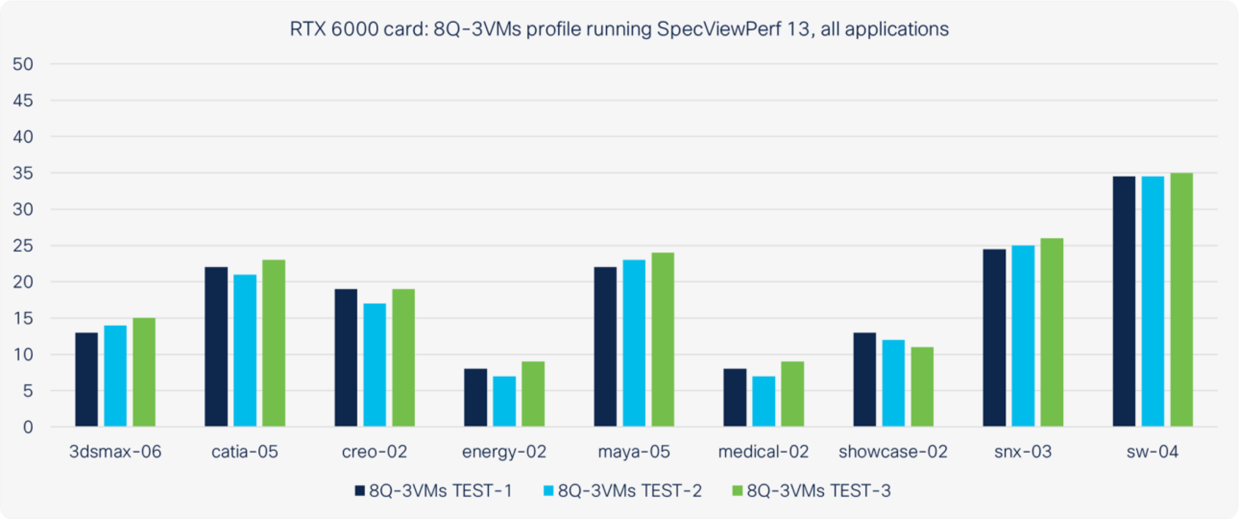


Figure 54.
RTX 6000 and 8000 with 8Q DWS profile

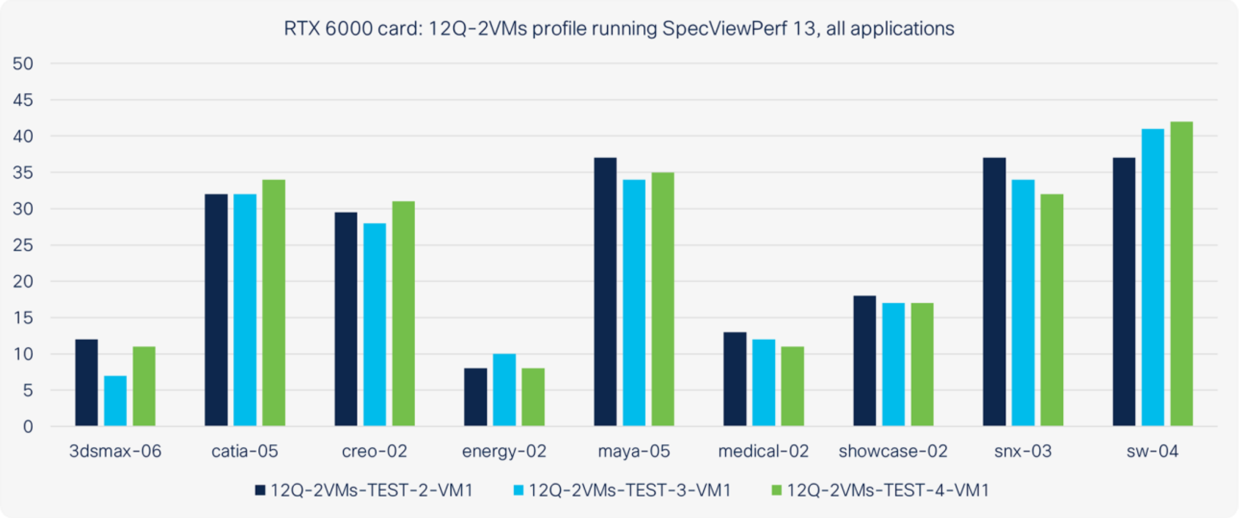


Figure 55.
RTX 6000 and 8000 with 12Q DWS profile

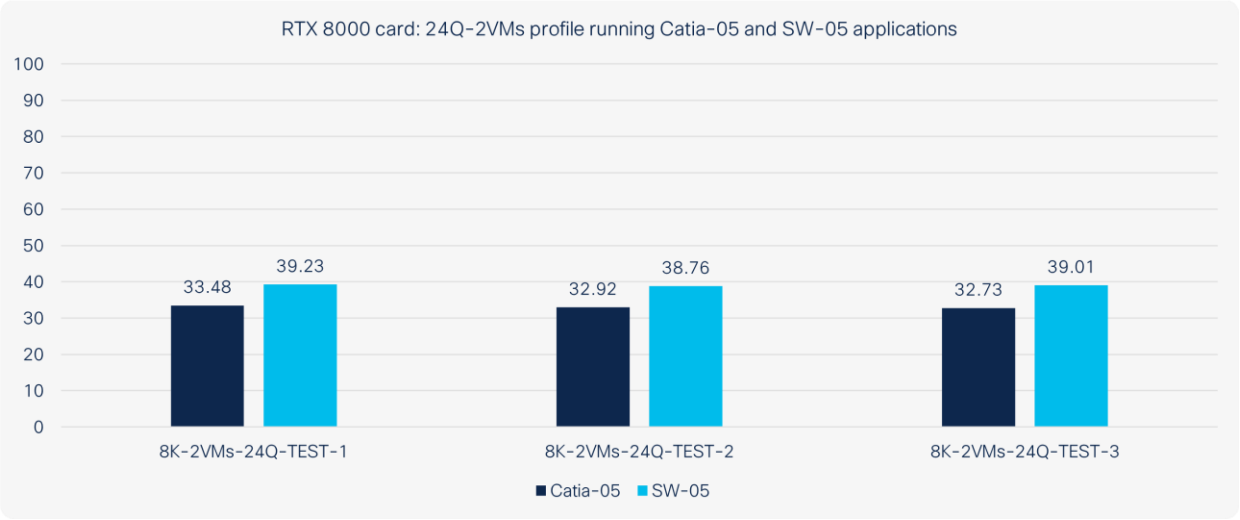


Figure 56.
RTX 6000 and 8000 with 24Q DWS profile

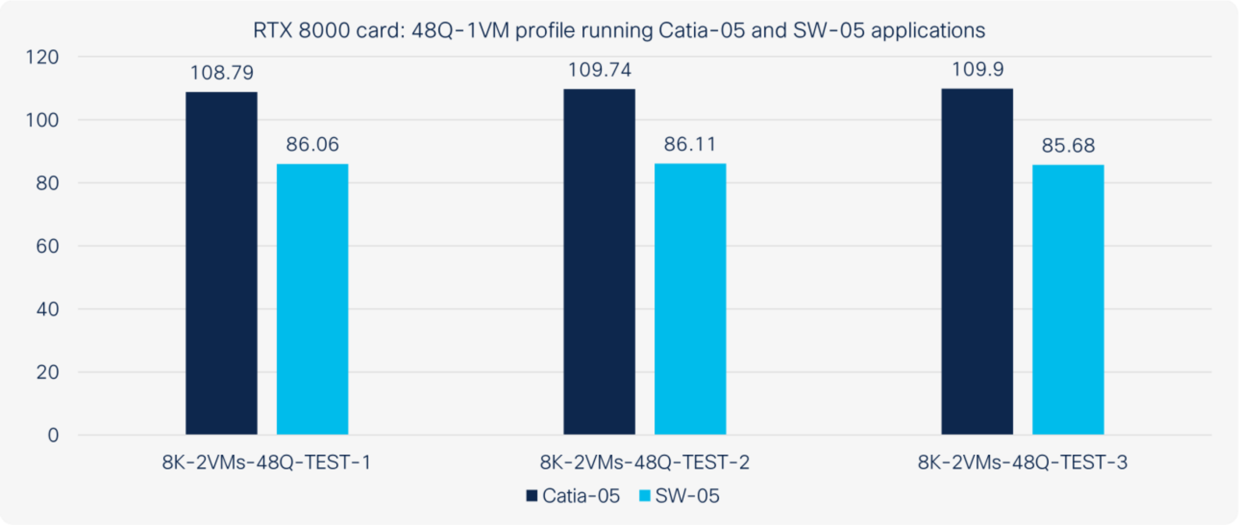


Figure 57.
RTX 6000/8000 with 48Q DWS Profile

Conclusion

The combination of Cisco UCS Manager, Cisco UCS C240 M5 Rack Servers, and NVIDIA RTX cards running on VMware vSphere 6.7 and VMware RDS server sessions and desktops provides a high-performance platform for virtualizing graphics-intensive applications.

By following the guidance in this document, our customers and partners can be assured that they are ready to host the growing list of graphics applications that are supported by our partners.

For more information

For additional information, see the following resources:

- Cisco UCS C-Series Rack Servers and B-Series Blade Servers:
    - https://www.cisco.com/en/US/products/ps10265/
- NVIDIA:
    - https://www.nvidia.com/object/grid-technology.html
    - https://www.nvidia.com/content/dam/en-zz/Solutions/design-visualization/solutions/resources/documents1/Deployment-Guide-vGPU-Software-VMware-vSphere.pdf
    - https://docs.nvidia.com/grid/latest/grid-software-quick-start-guide/index.html
- VMware Horizon desktops and RDS host server sessions:
    - https://techzone.vmware.com/resource/deploying-hardware-accelerated-graphics-vmware-horizon-7
    - https://docs.vmware.com/en/VMware-Horizon-6/6.2/com.vmware.horizon-view.desktops.doc/GUID-87B0EF3F-DE82-45B8-AD8C-E247084BD2CC.html
    - https://docs.vmware.com/en/VMware-vSphere/6.7/com.vmware.vsphere.vm_admin.doc/GUID-C597DC2A-FE28-4243-8F40-9F8061C7A663.html
    - https://techzone.vmware.com/resource/deploying-hardware-accelerated-graphics-vmware-horizon-7
- Microsoft Windows and VMware optimization guides for virtual desktops:
    - https://labs.vmware.com/flings/vmware-os-optimization-tool
- VMware vSphere ESXi and vCenter Server 6.7:
    - https://docs.vmware.com/en/VMware-vSphere/index.html

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **https://www.cisco.com/go/offices**.

Printed in USA C11-2852654-00 03/22