

Deploy Cisco UCS M5 Servers with NVIDIA GRID 8.5 on VMware vSphere 6.7 and Citrix Virtual Apps and Desktops 1912 LTSR

Contents

What you will learn	3
GRID vGPU profiles	4
Cisco Unified Computing System	5
NVIDIA	11
VMware vSphere 6.7	14
Solution configuration	19
Additional configurations	47
Performance testing with SPECviewperf 13	49
Conclusion	49
For more information	55

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 Citrix Virtual Apps 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 Citrix 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).

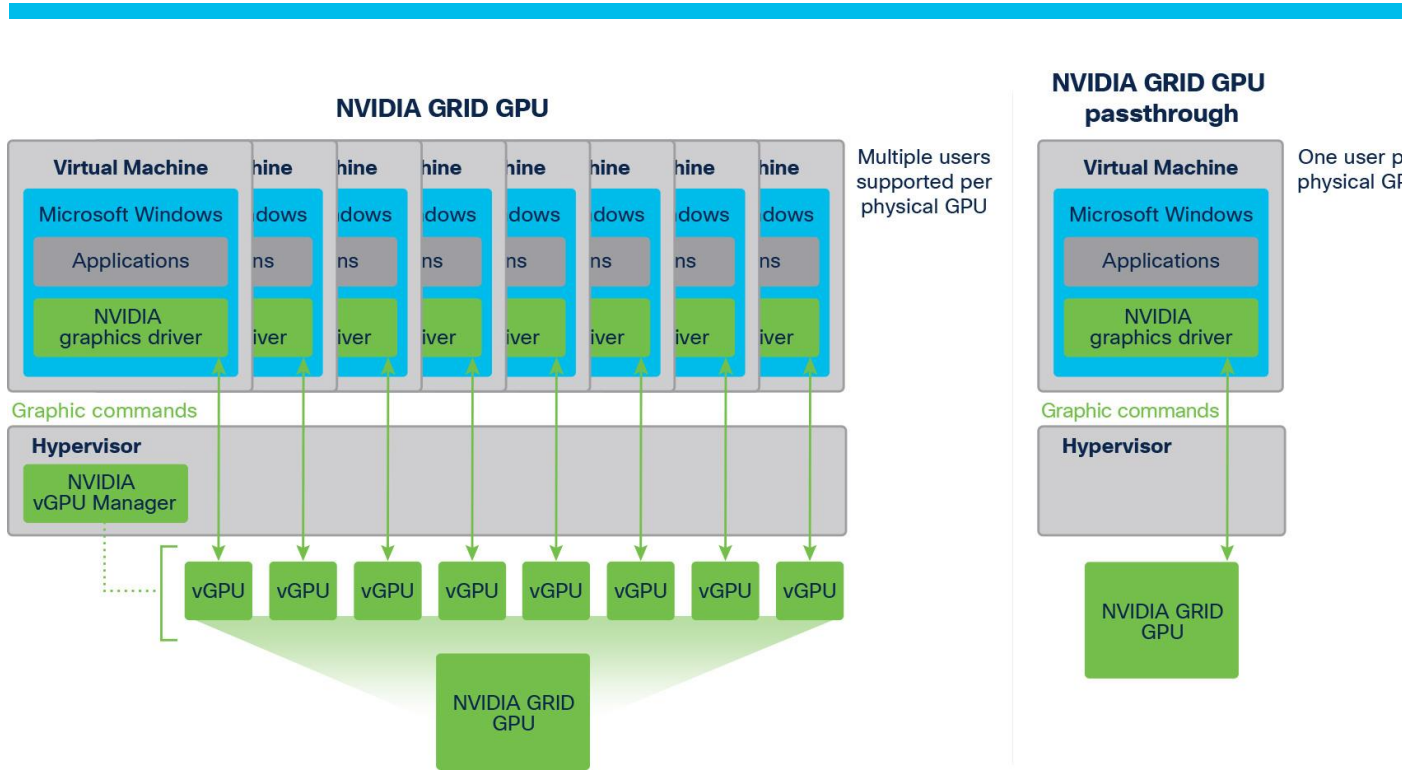


Figure 1.
NVIDIA GRID vGPU GPU system architecture

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).

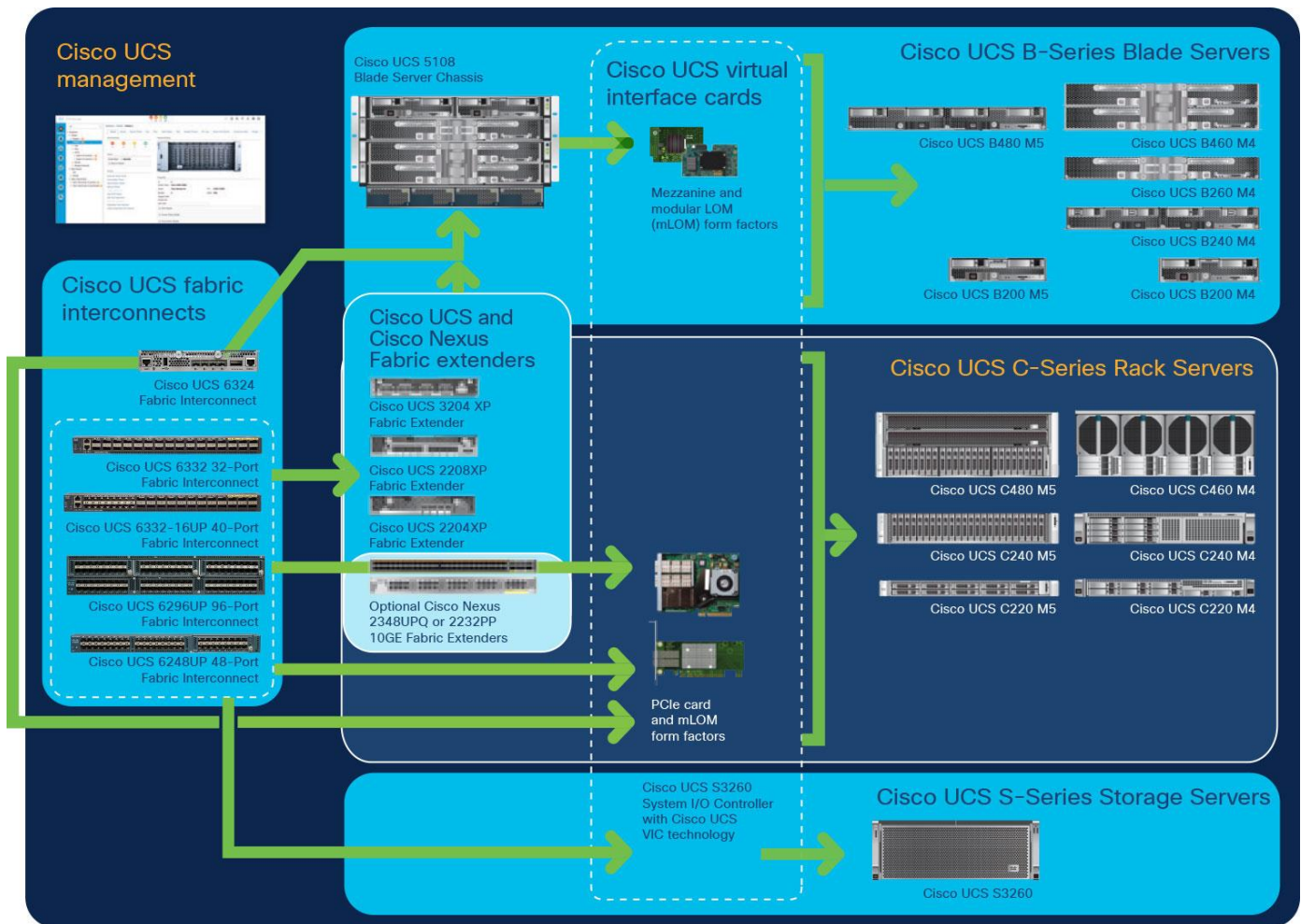


Figure 2.
Cisco UCS components

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>.

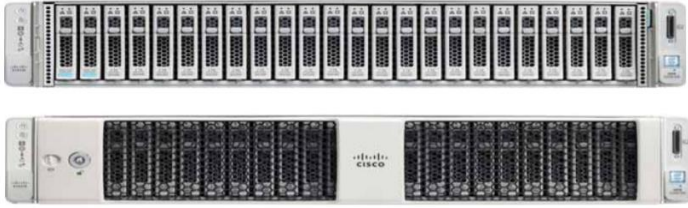


Figure 4.
Cisco UCS C240 M5 Rack Server

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

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 6) 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 6.
NVIDIA Quadro RTX 6000 GPU

NVIDIA Quadro RTX 8000 GPU

The NVIDIA Quadro RTX 8000 GPU (Figure 7) 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 7.
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	<ul style="list-style-type: none">• Read-only for review and documentation• Project management• Small parts or subassembly
Medium	<ul style="list-style-type: none">• Read-only and full application• Medium assemblies
Heavy	<ul style="list-style-type: none">• 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 <http://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 Citrix Virtual Apps and Desktops

Citrix HDX 3D Pro enables you to deliver the desktops and applications that perform best with a GPU for hardware acceleration, including 3D professional graphics applications based on OpenGL and DirectX. (The standard Citrix Virtual Delivery Agent [VDA] supports GPU acceleration of DirectX only.)

Examples of 3D professional applications include the following:

- Computer-aided design (CAD), manufacturing (CAM), and engineering (CAE) applications
- Geographical information system (GIS) software
- Picture archiving and communication system (PACS) for medical imaging
- Applications using the latest OpenGL, DirectX, NVIDIA Compute Unified Device Architecture (CUDA), and OpenCL versions
- Computationally intensive nongraphical applications that use CUDA GPUs for parallel computing

HDX 3D Pro provides an outstanding user experience over any bandwidth:

- On WAN connections: Delivers an interactive user experience over WAN connections with bandwidth as low as 1.5 Mbps
- On LAN connections: Delivers a user experience equivalent to that of a local desktop on LAN connections with bandwidth of 100 Mbps

You can replace complex and expensive workstations with simpler user devices by moving graphics processing into the data center for centralized management.

HDX 3D Pro provides GPU acceleration for Microsoft Windows desktops and Microsoft Windows Server. When used with VMware vSphere 6 and NVIDIA GRID GPUs, HDX 3D Pro provides vGPU acceleration for Windows desktops. For more information, see [Citrix Virtual GPU Solution](#).

GPU acceleration for Microsoft Windows desktops

With Citrix HDX 3D Pro, you can deliver graphics-intensive applications as part of hosted desktops or applications on desktop OS machines. HDX 3D Pro supports physical host computers (including desktop, blade, and rack workstations) and GPU pass-through and GPU virtualization technologies offered by VMware vSphere Hypervisor.

Using GPU pass-through, you can create virtual machines with exclusive access to dedicated graphics processing hardware. You can install multiple GPUs on the hypervisor and assign virtual machines to each of these GPUs on a one-to-one basis.

Using GPU virtualization, multiple virtual machines can directly access the graphics processing power of a single physical GPU. The true hardware GPU sharing provides desktops suitable for users with complex and demanding design requirements. GPU virtualization for NVIDIA GRID cards uses the same NVIDIA graphics drivers as are deployed on nonvirtualized operating systems.

HDX 3D Pro offers the following features:

- Adaptive H.264-based deep compression for optimal WAN and wireless performance: HDX 3D Pro uses CPU-based full-screen H.264 compression as the default compression technique for encoding. Hardware encoding is used with NVIDIA cards that support NVIDIA NVENC.
- Lossless compression option for specialized use cases: HDX 3D Pro offers a CPU-based lossless codec to support applications that require pixel-perfect graphics, such as medical imaging. True lossless compression is recommended only for specialized use cases because it consumes significantly more network and processing resources.
 - When you use lossless compression:

The lossless indicator, a system tray icon, shows the user whether the screen displayed is a lossy frame or a lossless frame. This information is helpful when the Visual Quality policy setting specifies a lossless build. The lossless indicator turns green when the frames sent are lossless.

The lossless switch enables the user to change to Always Lossless mode at any time in the session. To select or deselect Always Lossless at any time in a session, right-click the icon or use the shortcut Alt+Shift+1.
 - For lossless compression, HDX 3D Pro uses the lossless codec for compression regardless of the codec selected through policy.
 - For lossy compression, HDX 3D Pro uses the original codec: either the default or the one selected through policy.
 - Lossless switch settings are not retained for subsequent sessions. To use the lossless codec for every connection, select Always Lossless for the Visual Quality policy setting.
- Multiple and high-resolution monitor support: For Microsoft Windows 7 and 8 desktops, HDX 3D Pro supports user devices with up to four monitors. Users can arrange their monitors in any configuration and can mix monitors with different resolutions and orientations. The number of monitors is limited by the capabilities of the host computer GPU, the user device, and the available bandwidth. HDX 3D Pro supports all monitor resolutions and is limited only by the capabilities of the GPU on the host computer.
- Dynamic resolution: You can resize the virtual desktop or application window to any resolution.
- Support for NVIDIA Kepler architecture: HDX 3D Pro supports NVIDIA GRID K1 and K2 cards for GPU pass-through and GPU sharing. The GRID vGPU enables multiple virtual machines to have simultaneous, direct access to a single physical GPU, using the same NVIDIA graphics drivers as are deployed on nonvirtualized operating systems.
- Support for VMware vSphere and ESX using vDGA: You can use HDX 3D Pro with vDGA for both remote desktop service (RDS) and virtual desktop infrastructure (VDI) workloads. When you use HDX 3D Pro with vSGA, support is limited to one monitor. Use of vSGA with large 3D models can result in performance problems because of its use of API-intercept technology. For more information, see VMware vSphere 5.1: Citrix Known Issues.

As shown in Figure 8:

- The host computer must reside in the same Microsoft Active Directory domain as the delivery controller.
- When a user logs on to Citrix Receiver and accesses the virtual application or desktop, the controller authenticates the user and contacts the VDA for HDX 3D Pro to broker a connection to the computer hosting the graphical application.
- The VDA for HDX 3D Pro uses the appropriate hardware on the host to compress views of the complete desktop or of just the graphical application.
- The desktop or application views and the user interactions with them are transmitted between the host computer and the user device through a direct HDX connection between Citrix Receiver and the VDA for HDX 3D Pro.

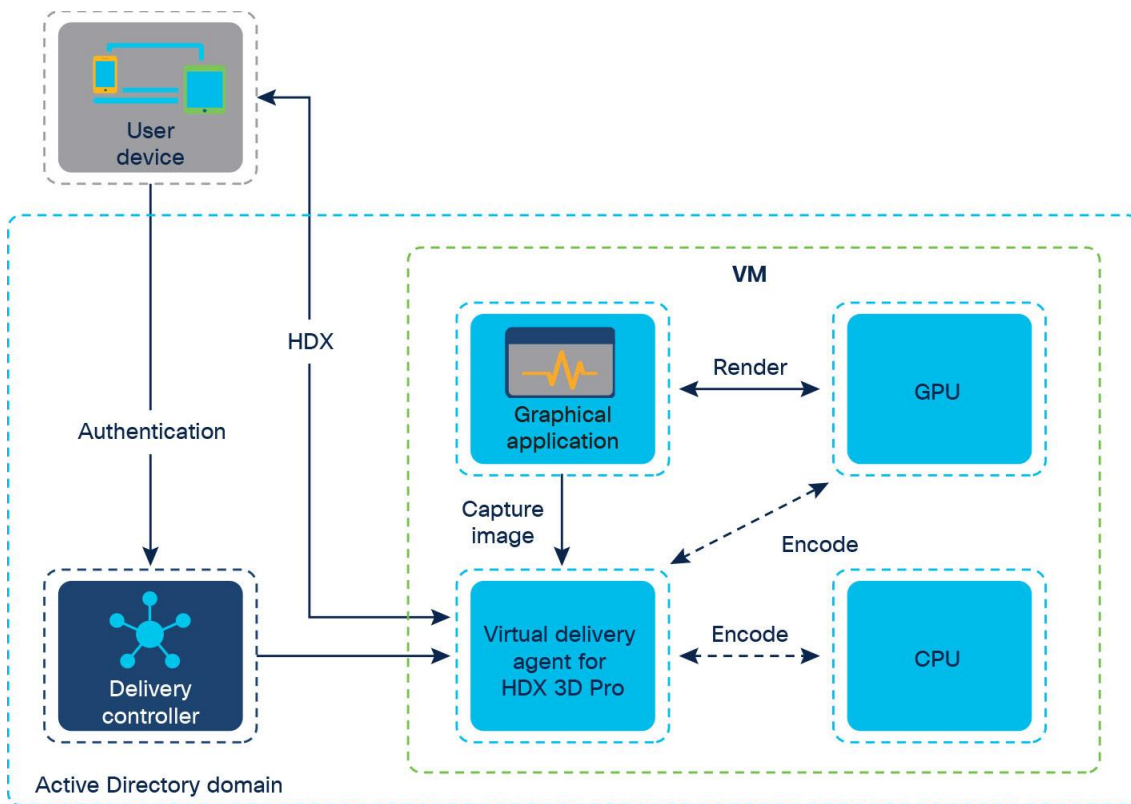


Figure 8.
Citrix HDX 3D Pro process flow

GPU acceleration for Microsoft Windows Server

Citrix HDX 3D Pro allows graphics-intensive applications running in Microsoft Windows Server sessions to render on the server's GPU. By moving OpenGL, DirectX, Direct3D, and Windows Presentation Foundation (WPF) rendering to the server's GPU, the server's CPU is not slowed by graphics rendering. Additionally, the server can process more graphics because the workload is split between the CPU and the GPU.

GPU sharing for Citrix RDS workloads

RDS GPU sharing enables GPU hardware rendering of OpenGL and Microsoft DirectX applications in remote desktop sessions.

- Sharing can be used on bare-metal devices or virtual machines to increase application scalability and performance.
- Sharing enables multiple concurrent sessions to share GPU resources (most users do not require the rendering performance of a dedicated GPU).
- Sharing requires no special settings.

For DirectX applications, only one GPU is used by default. That GPU is shared by multiple users. The allocation of sessions across multiple GPUs with DirectX is experimental and requires registry changes. Contact Citrix Support for more information.

You can install multiple GPUs on a hypervisor and assign virtual machines to each of these GPUs on a one-to-one basis: either install a graphics card with more than one GPU, or install multiple graphics cards with one or more GPUs each. Mixing heterogeneous graphics cards on a server is not recommended.

Virtual machines require direct pass-through access to a GPU, which is available with VMware vSphere 6. When Citrix HDX 3D Pro is used with GPU pass-through, each GPU in the server supports one multiuser virtual machine.

Scalability using RDS GPU sharing depends on several factors:

- The applications being run
- The amount of video RAM that the applications consume
- The graphics card's processing power

Some applications handle video RAM shortages better than others. If the hardware becomes extremely overloaded, the system may become unstable, or the graphics card driver may fail. Limit the number of concurrent users to avoid such problems.

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

GPU acceleration for Microsoft single-session OS

With HDX 3D Pro, you can deliver graphically intensive applications as part of hosted desktops or applications on single-session OS machines. HDX 3D Pro supports physical host computers (including desktop, blade, and rack workstations) and GPU pass-through and GPU virtualization technologies offered by Citrix XenServer, VMware vSphere, and Microsoft Hyper-V (pass-through only) hypervisors.

Using GPU pass-through, you can create virtual machines with exclusive access to dedicated graphics processing hardware. You can install multiple GPUs on the hypervisor and assign virtual machines to each of these GPUs on a one-to-one basis.

Citrix HDX 3D Pro requirements

The physical or virtual machine hosting the application can use GPU pass-through or vGPU:

- GPU pass-through is available with Citrix Hypervisor; VMware vSphere and ESX, where it is referred to as virtual direct graphics acceleration (vDGA); and Microsoft Hyper-V in Microsoft Windows Server 2016 and 2019, where it is referred to as discrete device assignment (DDA).
- vGPU is available with Citrix Hypervisor and VMware vSphere.
- Citrix recommends that the host computer have at least 4 GB of RAM and four virtual CPUs with a clock speed of 2.3 GHz or higher.

The requirements for the GPU are as follows:

- For CPU-based compression (including lossless compression), Citrix HDX 3D Pro supports any display adapter on the host computer that is compatible with the application being delivered.
- For virtualized graphics acceleration using the NVIDIA GRID API, HDX 3D Pro can be used with supported GRID cards (see NVIDIA GRID). GRID delivers a high frame rate, resulting in a highly interactive user experience.
- Virtualized graphics acceleration is supported on the Intel Xeon processor E3 family data center graphics platform. For more information, see <http://www.citrix.com/intel> and <http://www.intel.com/content/www/us/en/servers/data-center-graphics.html>.

The requirements for the user device are as follows:

- HDX 3D Pro supports all monitor resolutions that are supported by the GPU on the host computer. However, for optimal performance with the minimum recommended user device and GPU specifications, Citrix recommends a maximum monitor resolution for user devices of 1920 x 1200 pixels for LAN connections, and 1280 x 1024 pixels for WAN connections.
- Citrix recommends that user devices have at least 1 GB of RAM and a CPU with a clock speed of 1.6 GHz or higher. Use of the default deep compression codec, which is required on low-bandwidth connections, requires a more powerful CPU unless the decoding is performed in hardware. For optimum performance, Citrix recommends that user devices have at least 2 GB of RAM and a dual-core CPU with a clock speed of 3 GHz or higher.
- For multiple-monitor access, Citrix recommends user devices with quad-core CPUs.
- User devices do not need a GPU to access desktops or applications delivered with HDX 3D Pro.
- Citrix Receiver must be installed.

For more information, see the Citrix HDX 3D Pro documentation at <https://docs.citrix.com/en-us/citrix-virtual-apps-desktops/1912-ltsr/graphics/hdx-3d-pro.html>.

Solution configuration

Figure 9 provides an overview of the solution configuration.

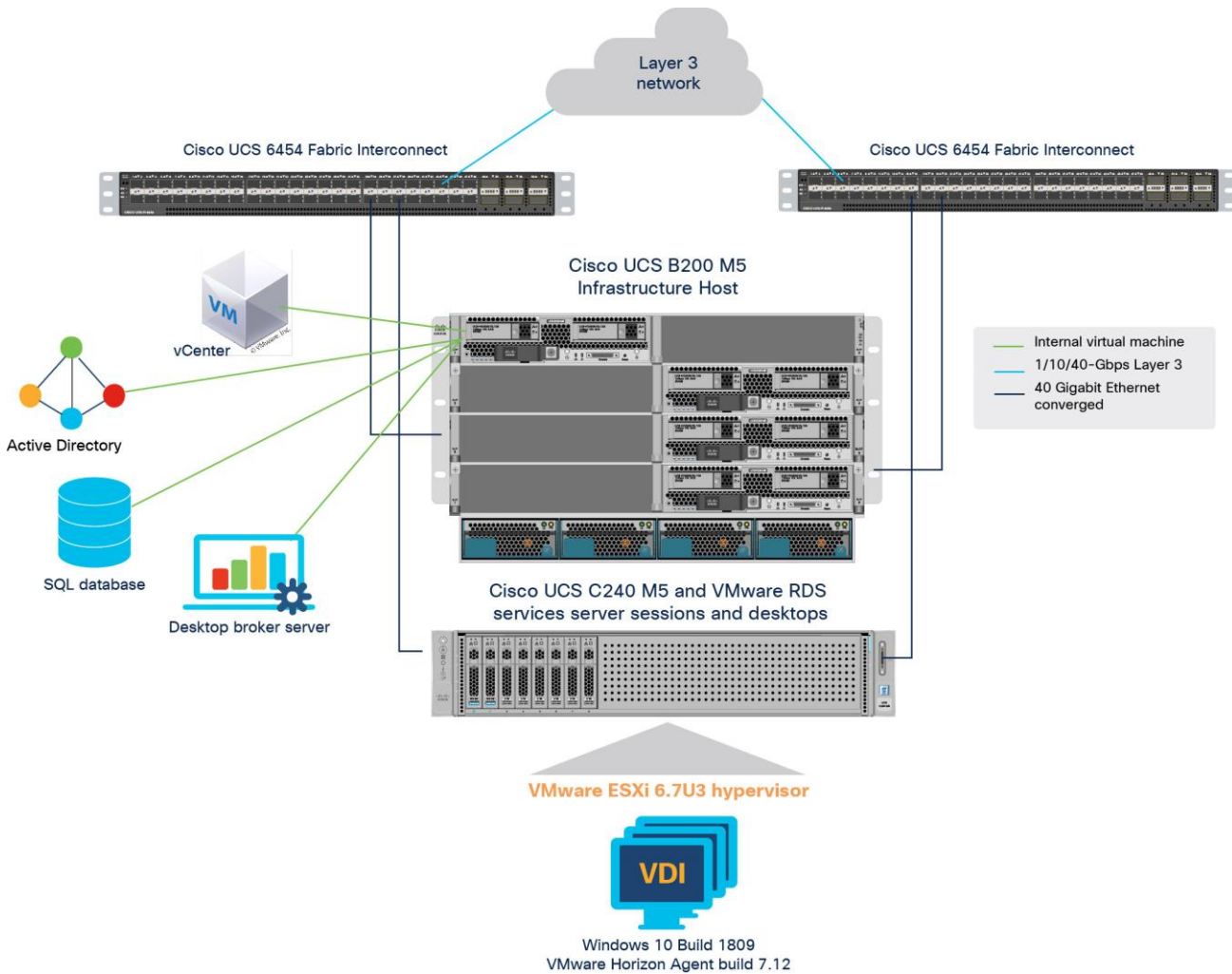


Figure 9.
Reference architecture

The hardware components in the solution are listed here:

- Cisco UCS C240 M5 Rack Server (two Intel Xeon Scalable Platinum 6154 CPUs at 2.10 GHz) with 1.5 TB of memory (64 GB x 24 DIMMs at 2933MHz)
- Cisco UCS VIC 1457 mLOM (Cisco UCS C240 M5)
- Two Cisco UCS 6454 third-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
- Citrix Virtual Apps and Desktops 1912 Long-Term Service Release (LTSR)
- Microsoft Windows 10 64-bit
- Microsoft Server 2019
- NVIDIA GRID software and licenses:
 - NVIDIA-VMware-440.107-1OEM.670.0.0.8169922.x86_64.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 4.0(4g)
NVIDIA RTX 6000	Release 4.0(4g)

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 10 shows a one-GPU installation, and Figure 11 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

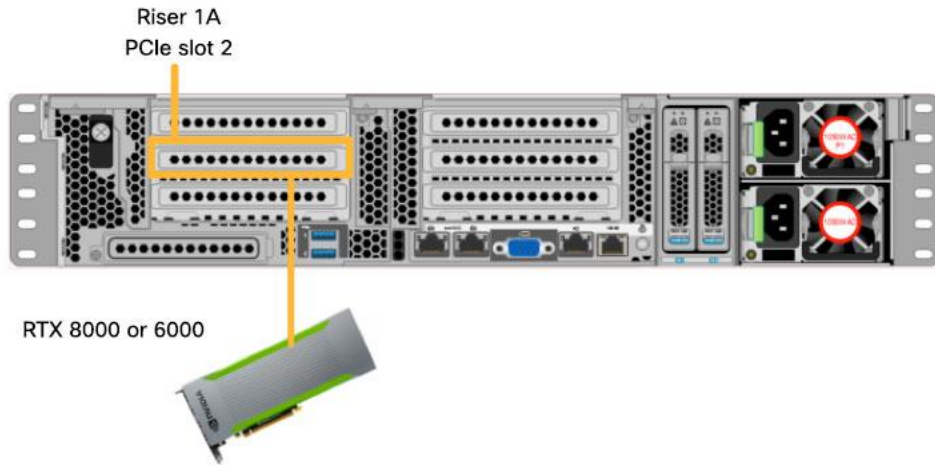


Figure 10.
One-GPU scenario

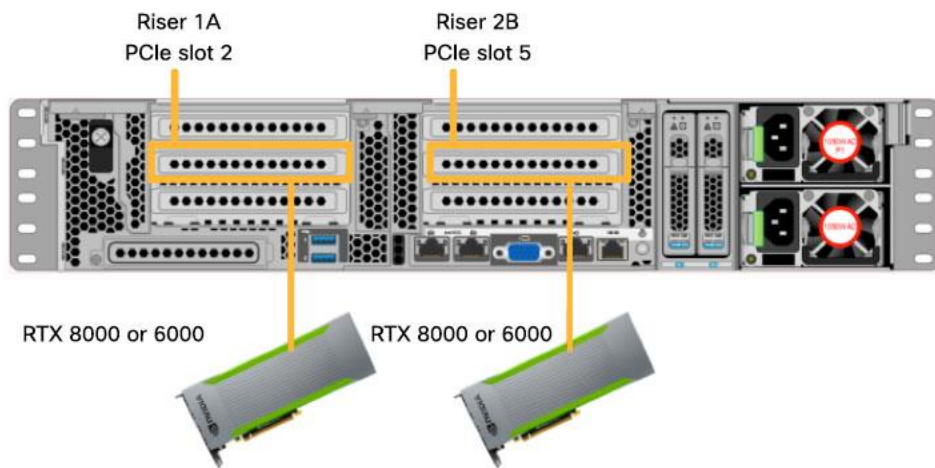


Figure 11.
Two-GPU scenario

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 IMC, select the server and choose Inventory > GPUs. As shown in Figure 12, PCIe slots 2 and 5 are used with two RTX 6000 and 8000 cards.

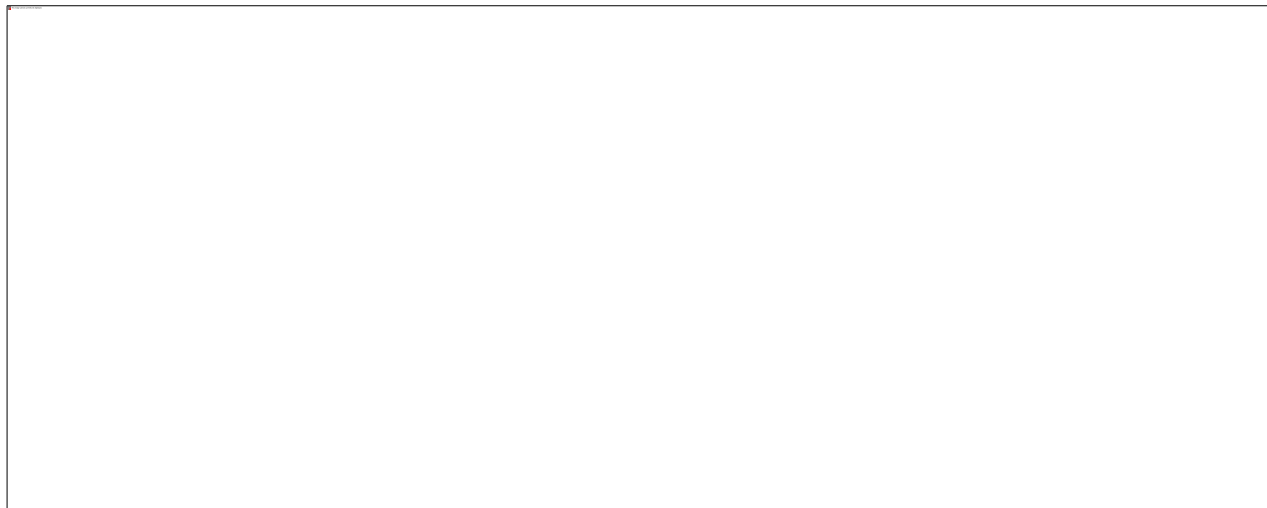


Figure 12.

NVIDIA GRID RTX card inventory displayed in Cisco IMC

2. If necessary, you can use the Cisco IMC to 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](http://www.nvidia.com/object/grid-evaluation.html).

Three packages are required for VMware ESXi host setup, as shown in Figure 13:

- 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-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_440.107-10EM.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-Is-windows-2018.10.0.25098346 (1)	7/28/2020 12:59 PM	Compressed (zipp...	
NVIDIA-Is-Windows-2020.05.0.28406365	7/27/2020 2:31 PM	Compressed (zipp...	
NVIDIA-VMware-440.107-10EM.670.0.0.8169922.x86_64.vib	7/12/2020 10:28 PM	VIB File	

Figure 13.
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-Is-windows-2020.05.0.28406365 folder. Run setup.exe (Figure 14).

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 14.
Run setup.exe

2. Click Next (Figure 15).

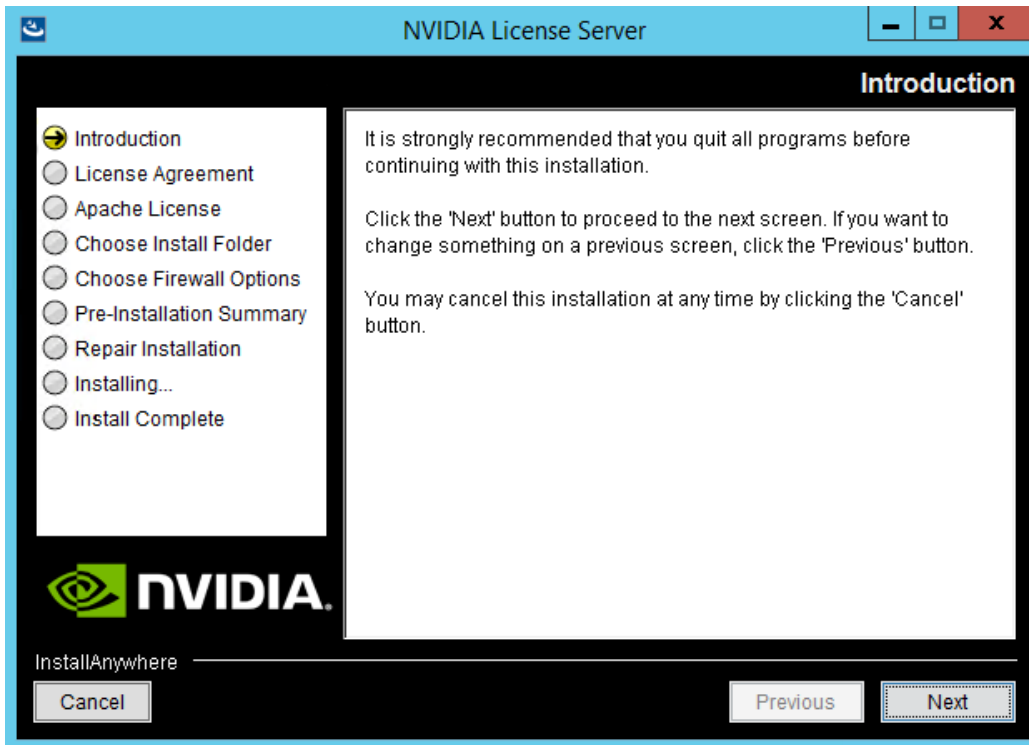


Figure 15.
NVIDIA License Server screen

3. Accept the license agreement and click Next (Figure 16).

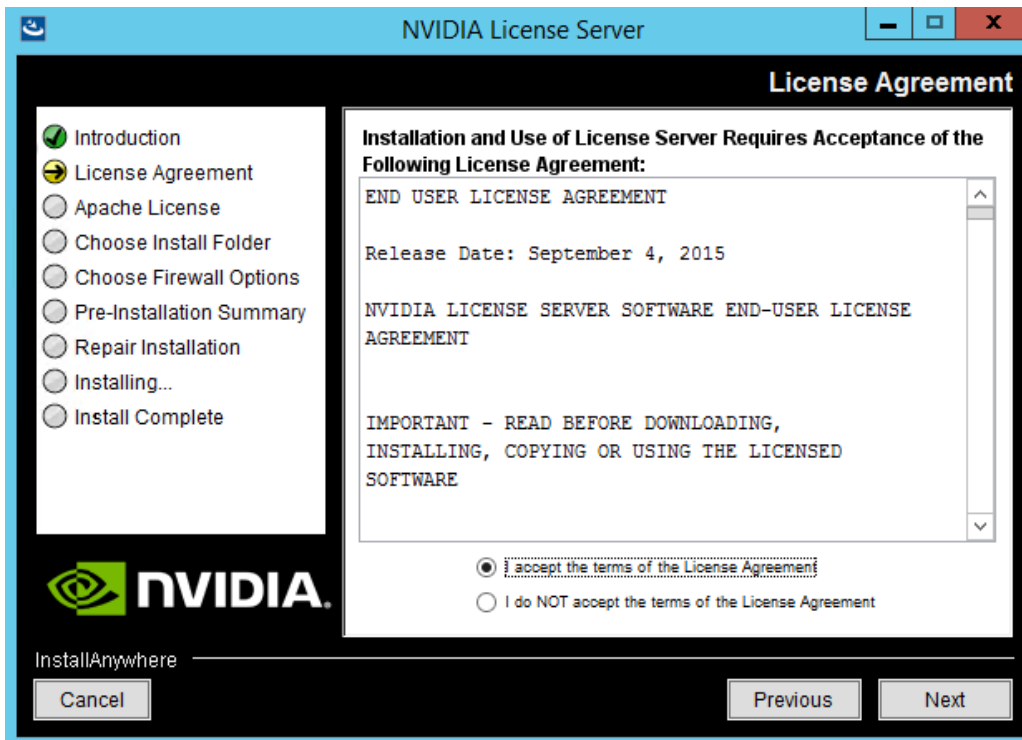


Figure 16.
NVIDIA License Agreement screen

4. Accept the Apache license agreement and click Next (Figure 17).

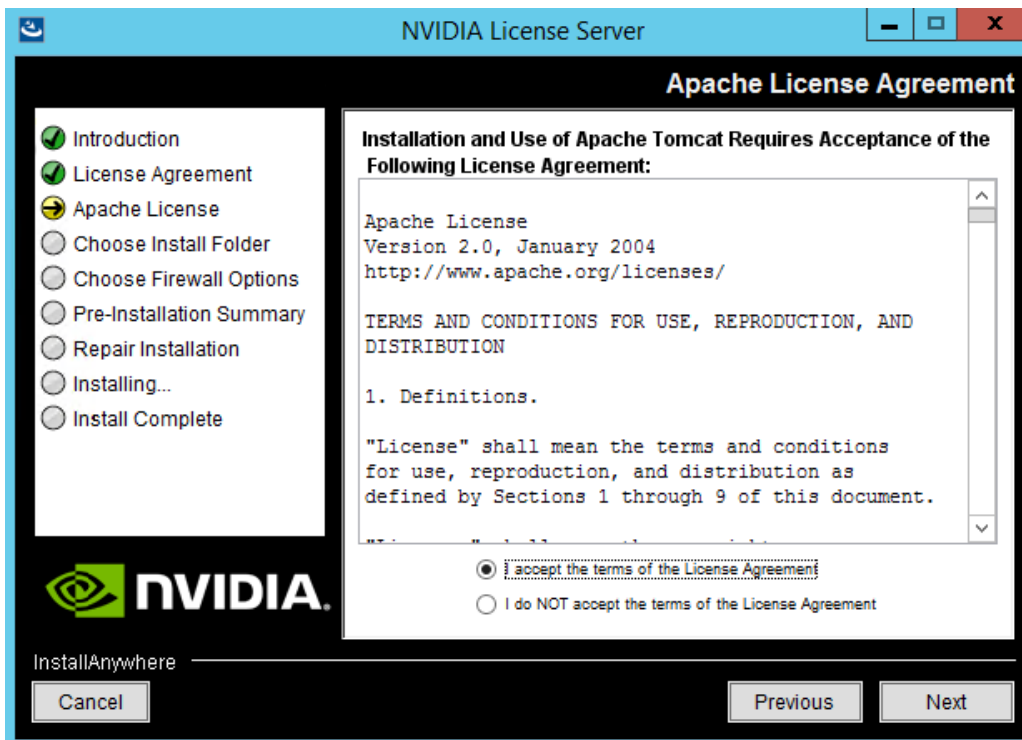


Figure 17. NVIDIA License Agreement screen

5. Choose the desired installation folder and click Next (Figure 18).

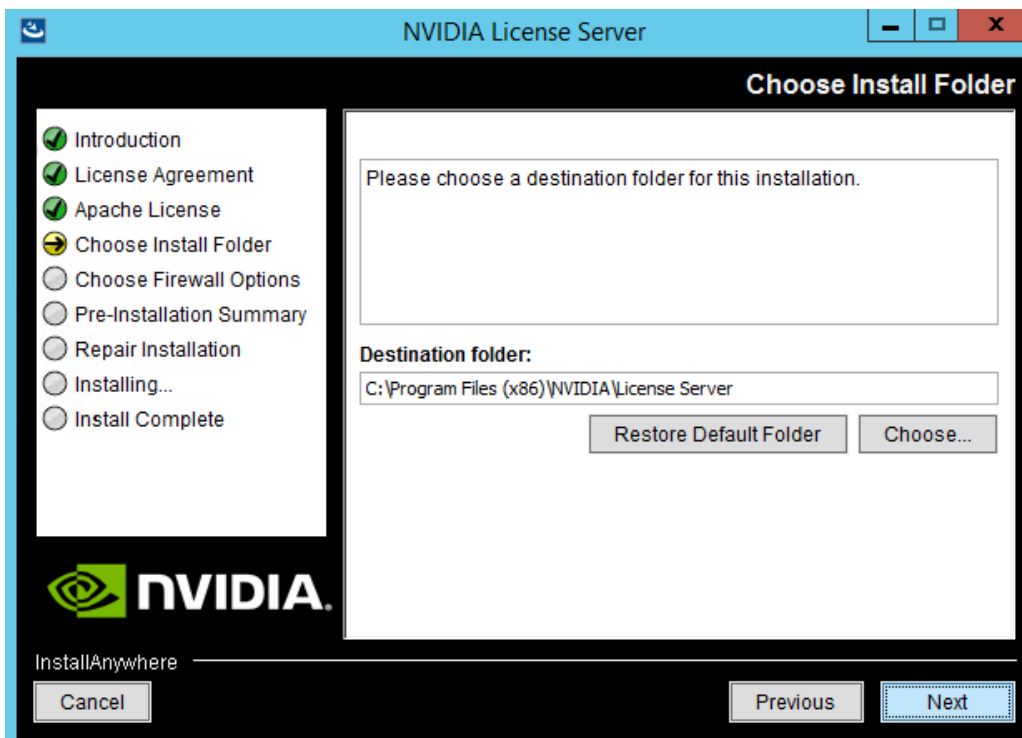


Figure 18. Choosing a destination folder

- 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.
- 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.
- Click Next (Figure 19).

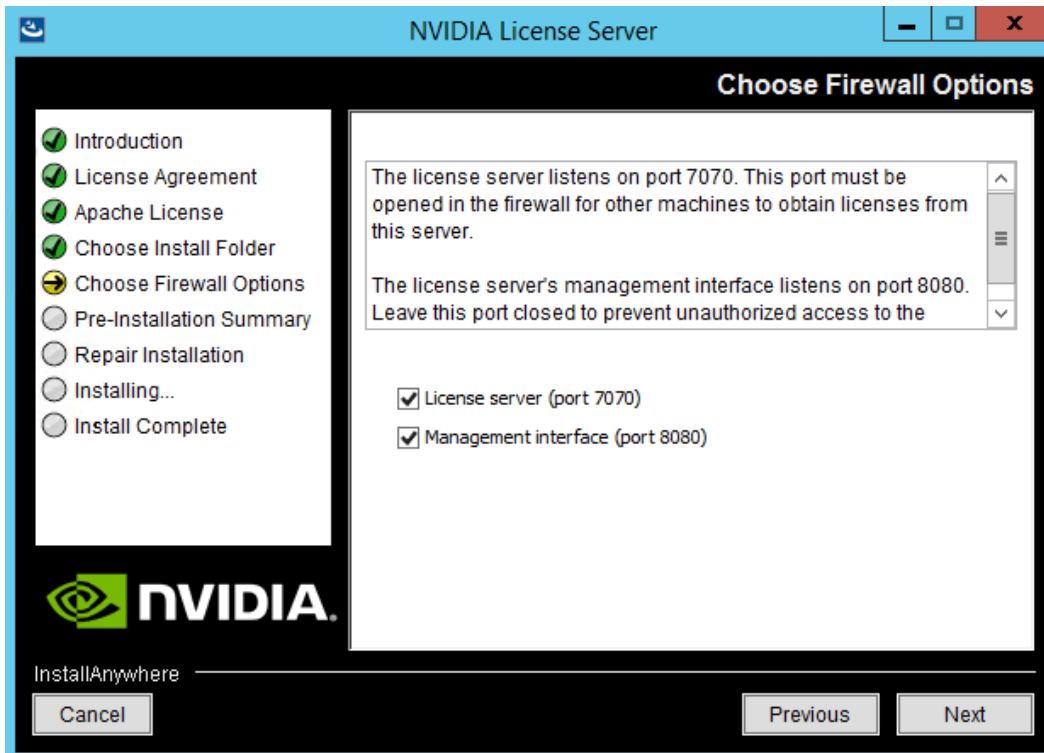


Figure 19.
Setting firewall options

- The Pre-installation Summary and Repair Installation options automatically progresses without user input (Figure 20).

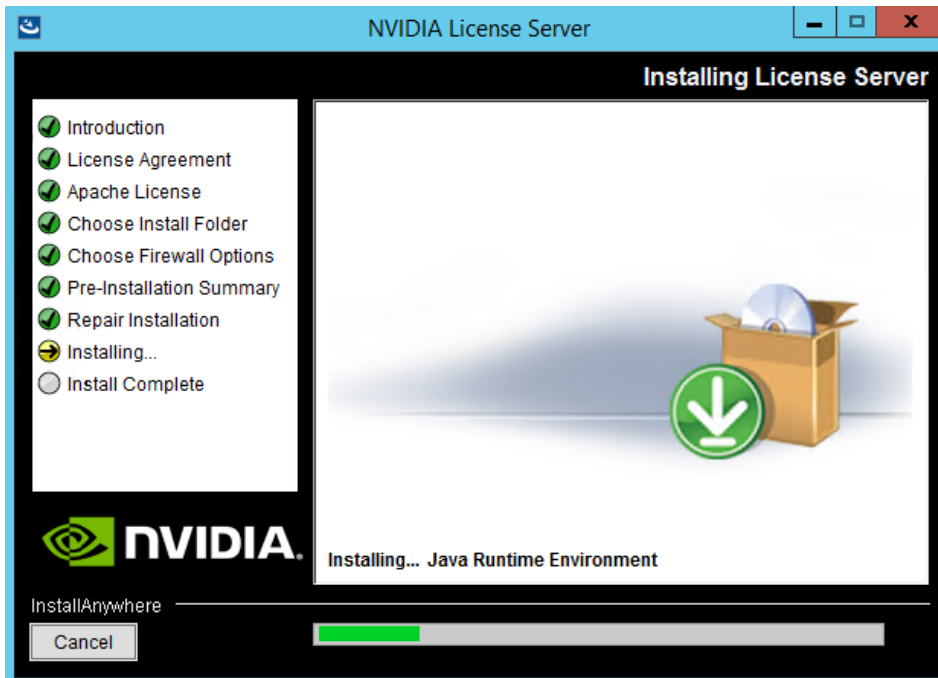


Figure 20.
Installing the license server

10. When the installation process is complete, click Done (Figure 21).

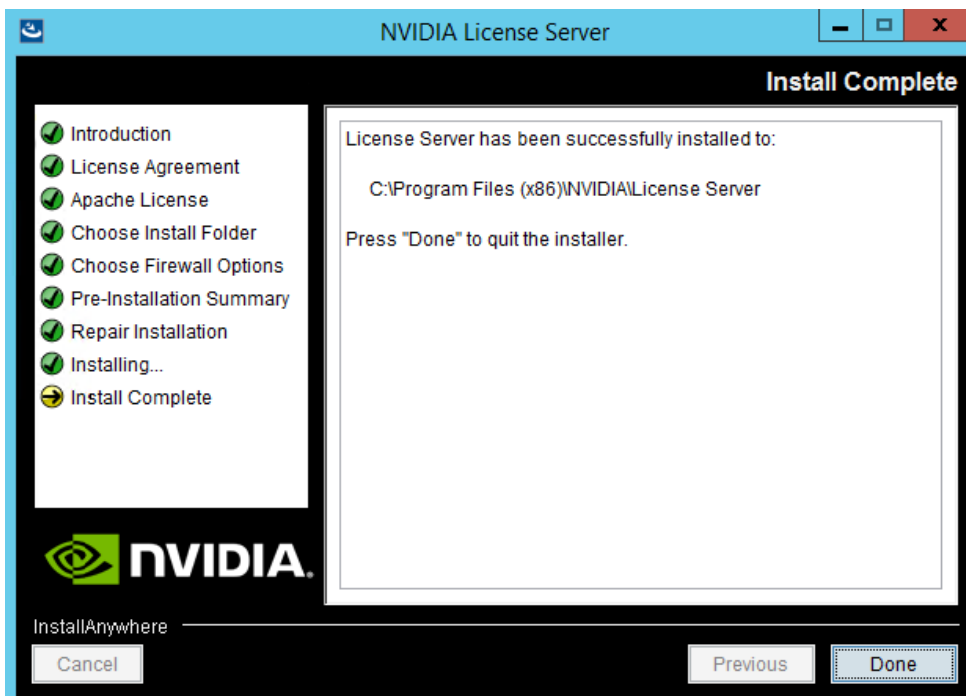


Figure 21.
Installation complete

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 22).
4. Specify the fields as shown in Figure 23. 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 24).
6. Click Create License Server.

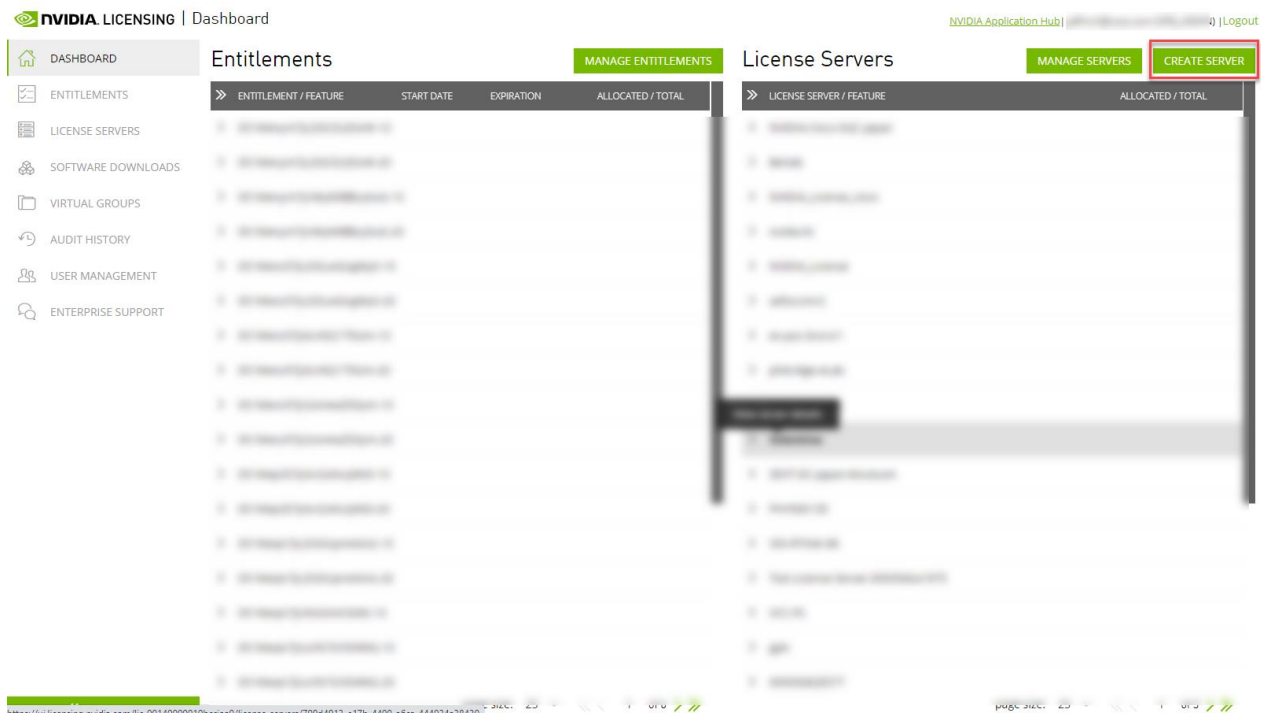


Figure 22.
Creating the license server

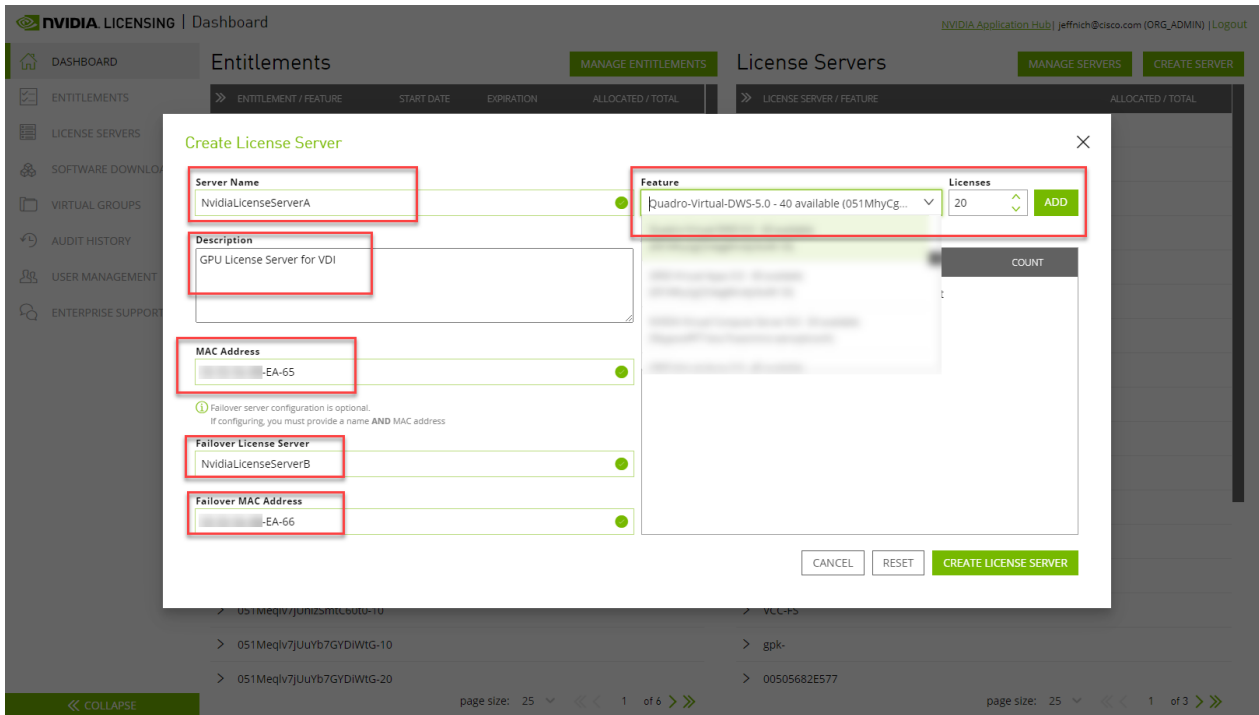


Figure 23.
Configuring the license server

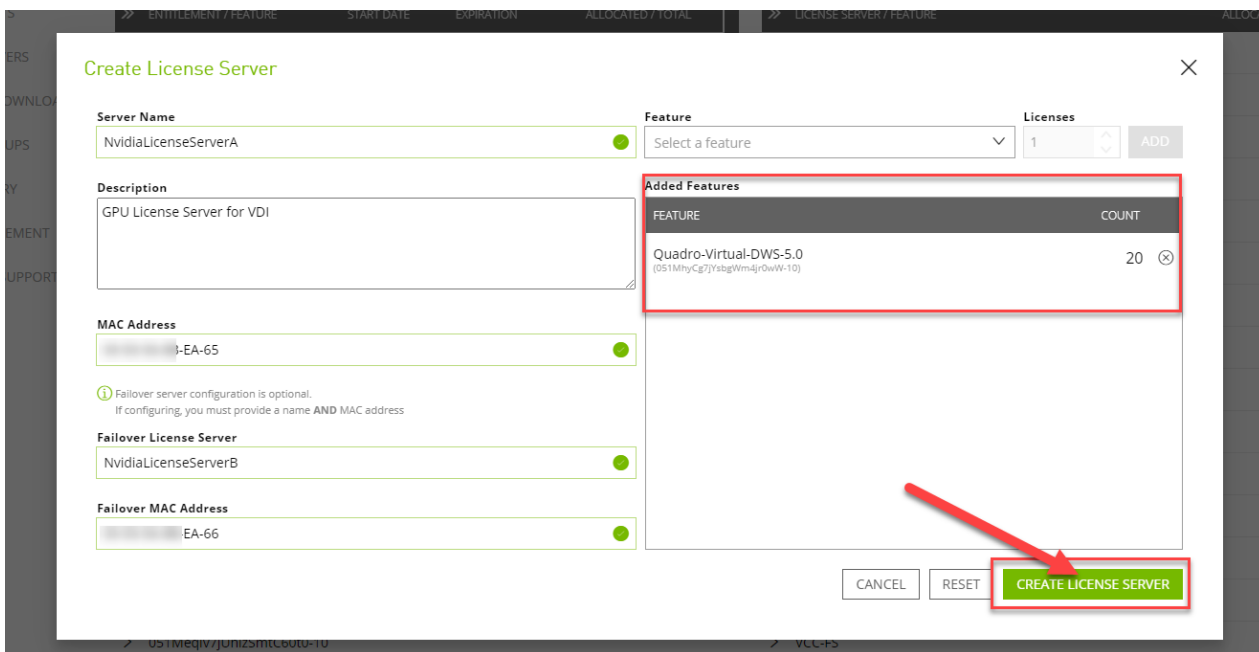


Figure 24.
Adding the licenses and creating the server

7. Find your license server under License Servers.
8. Click your license server ID (Figure 25).



License Servers

[MANAGE SERVERS](#)[CREATE SERVER](#)









LICENSE SERVER / FEATURE	ALLOCATED / TOTAL
> [blurred]	
> [blurred]	
> [blurred]	
✔ nvidia-lic	
    	
GRID-Virtual-Apps-3.0 (6oiqnixzs-diuF0e4ldl-w4jvwfxqdz)	500 / 500 
Quadro-Virtual-DWS-5.0 (051Meqlv7jYsbgHvqjJ0wG-10)	10 / 10 
Quadro-Virtual-DWS-5.0 (051Menz97jC2t5LwGngWy0-10)	10 / 10 
> [blurred]	
> [blurred]	

Figure 25.
Selecting the license server ID

9. Click Download License File and save the .bin file to your license server (Figure 26).

License Servers

[MANAGE SERVERS](#)[CREATE SERVER](#)









LICENSE SERVER / FEATURE	ALLOCATED / TOTAL
> [blurred]	
> [blurred]	
> [blurred]	
✔ nvidia-lic	
    	
GRID-Virtual-Apps-3.0 (6oiqnixzs-diuF0e4ldl-w4jvwfxqdz)	500 / 500 
Quadro-Virtual-DWS-5.0 (051Meqlv7jYsbgHvqjJ0wG-10)	10 / 10 
Quadro-Virtual-DWS-5.0 (051Menz97jC2t5LwGngWy0-10)	10 / 10 
> [blurred]	
> [blurred]	

Figure 26.
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 <http://<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 27). The features are available (Figure 28).

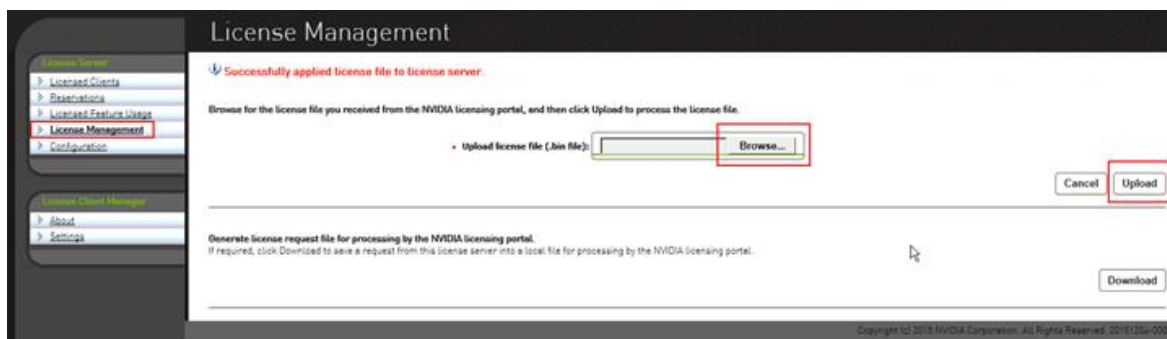


Figure 27.
License file successfully applied

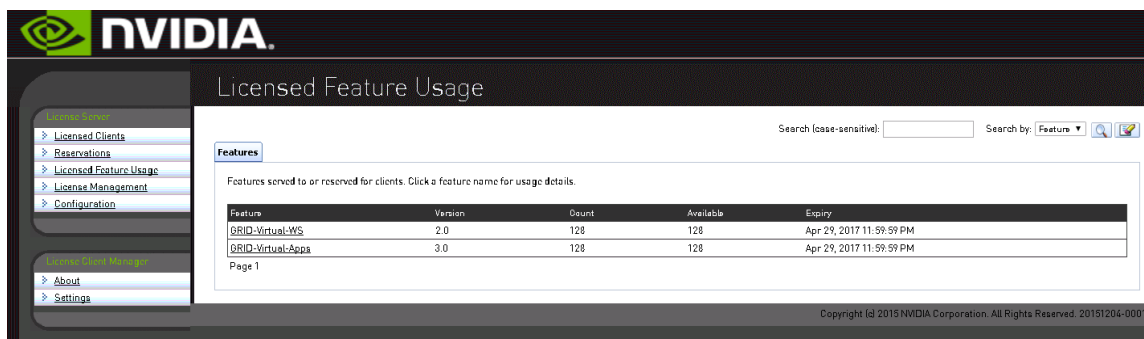


Figure 28.
NVIDIA license server available for use

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 29 shows the components used for vGPU support.

NVIDIA GRID vGPU



Figure 29.
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 30).

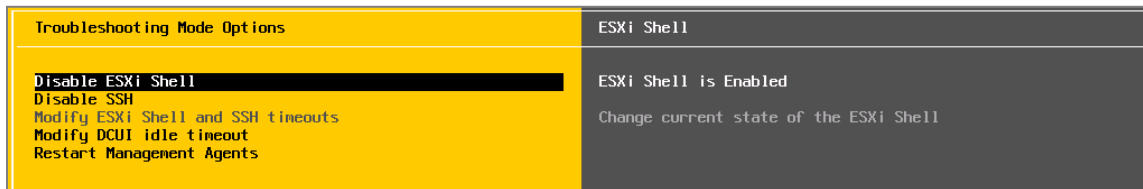


Figure 30.
VMware ESXi configuration console

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-440.107-10EM.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_440.107-10EM.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  440.107-10EM.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:
http://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 31.

```

[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 31. 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 <http://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, 16 GB, 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 32).

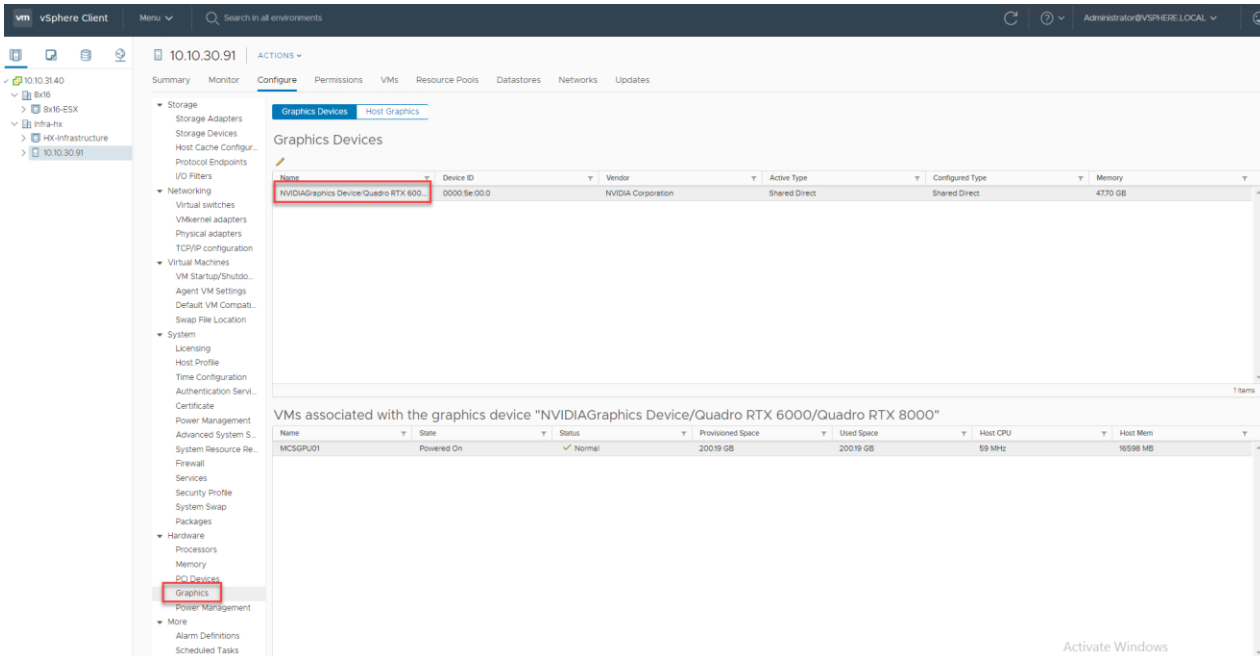


Figure 32. Make GPUs available for passthrough

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.

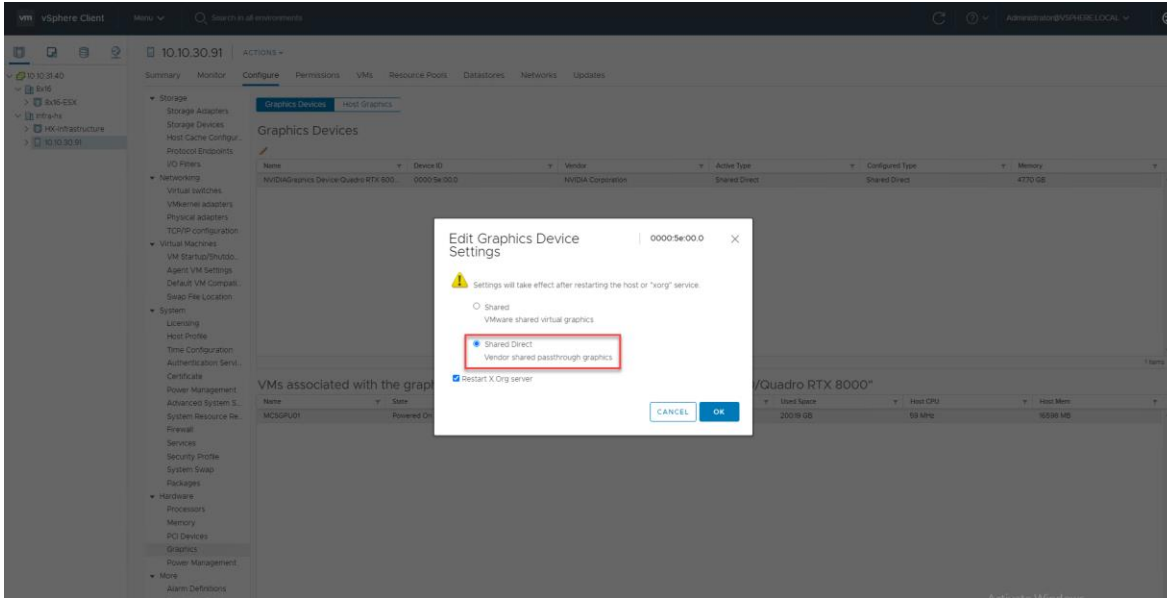


Figure 33.
Edit host graphics settings

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 34).

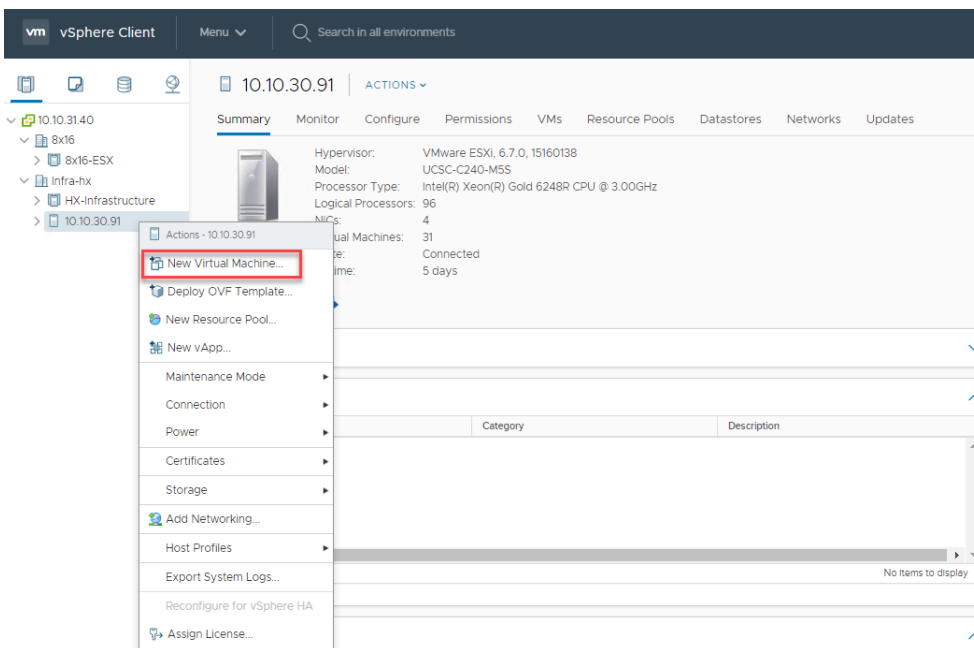


Figure 34.
Creating a new virtual machine in the VMware vSphere Web Client

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 35). “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.

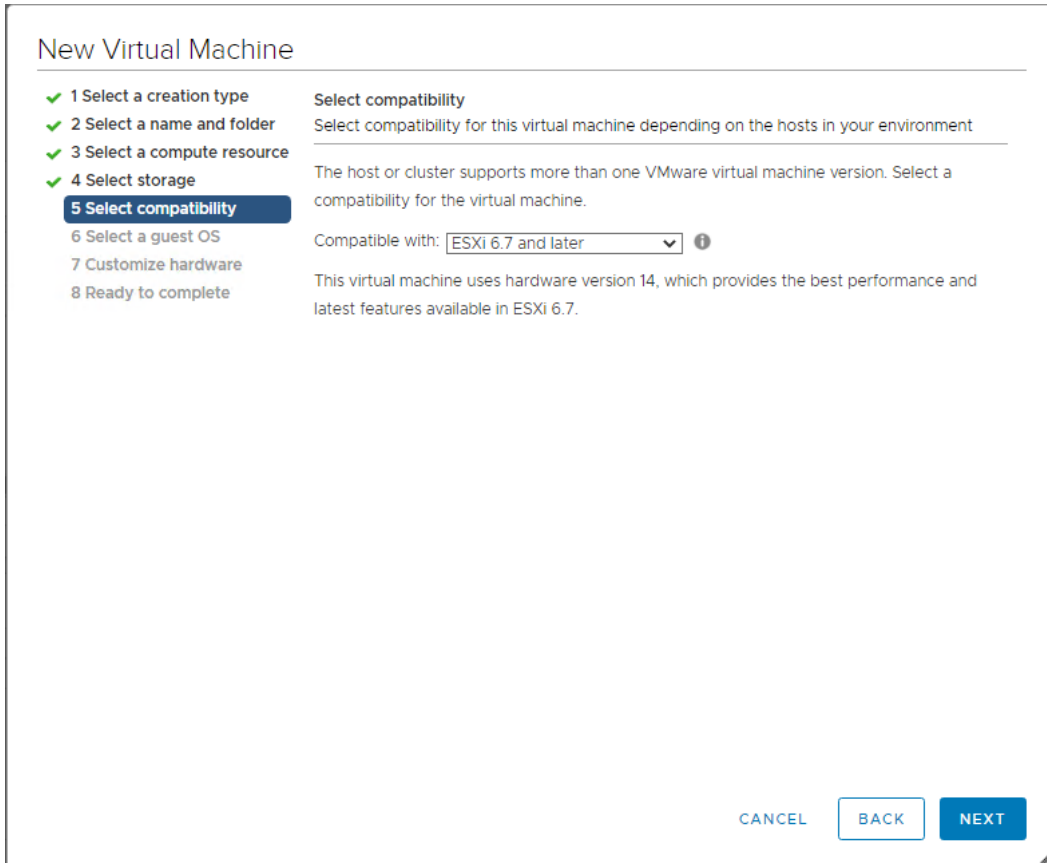


Figure 35.
Selecting virtual machine hardware Version 11 or later

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 36 and 37).

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.

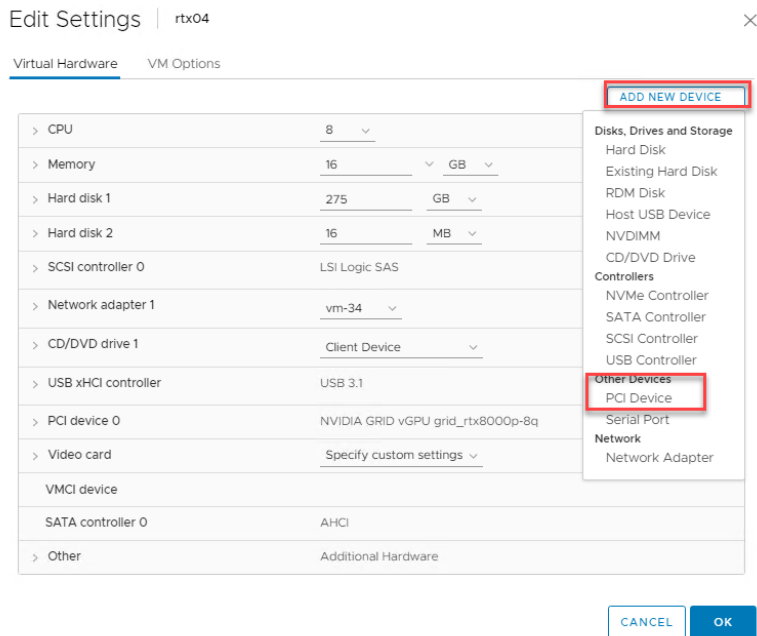


Figure 36. Adding a shared PCI device to the virtual machine to attach the GPU profile

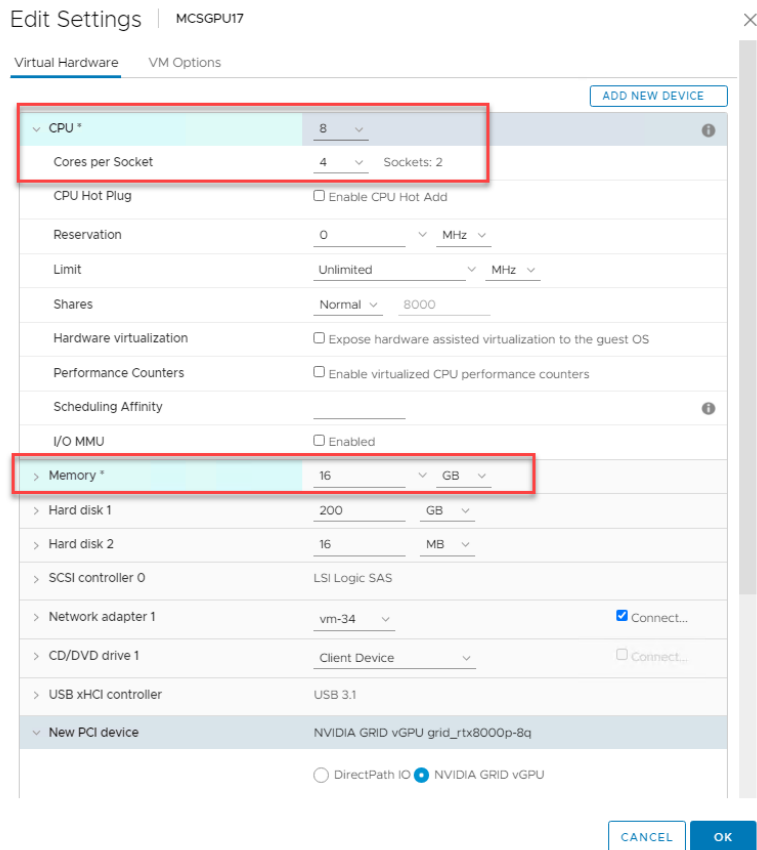


Figure 37. Attaching the GPU profile to a shared PCI device

- Go to the VM Options tab. From the drop-down menu under Advanced > Configuration Parameters, choose Edit Configuration. Add the name **pciPassthru.use64bitMMIO** and set the value to TRUE (Figure 38).

Configuration Parameters ×

pciBridge7.virtualDev	pcieRootPort
pciPassthru.64bitMMIOSizeGB	48
pciPassthru.use64bitMMIO	TRUE
sata0.pciSlotNumber	34
sched.cpu.latencySensitivity	normal
sched.mem.pin	TRUE
sched.swap.derivedName	/vmfs/volumes/5e3345

Figure 38.
Add pciPassthru.use64bitMMIO = 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 39):

```
#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 39.
Disable ECC

9. Install and configure Microsoft Windows on the virtual machine:

- For higher-end vGPU desktops such as those used on 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 vCPUs (4 per socket) and 16 GB of RAM; see Figure 40).

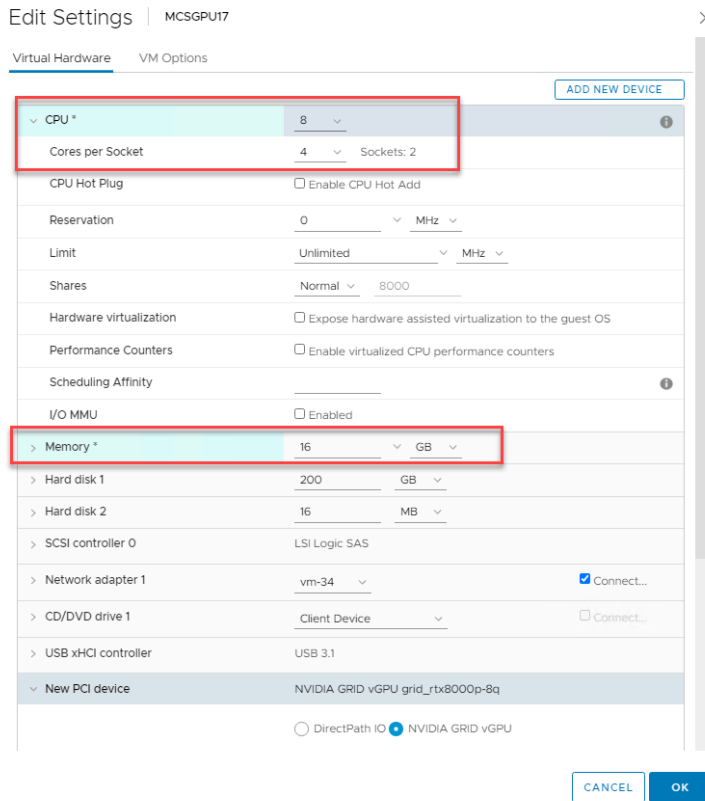


Figure 40.
Configure vCPUs and RAM

- Install VMware Tools.
- Join the virtual machine to the Microsoft Active Directory domain.
- Install or upgrade Citrix Virtual Desktop Agent (in this document, Version 1912 LTSR is used).

HDX 3D Pro mode is recommended for data center desktops with GPU hardware. This option is available with the desktop OS VDA to use GPU vendor drivers and APIs for hardware rendering and graphics acceleration.

There are different ways to configure a virtual desktop with HDX 3D Pro depending on the product version used. For VDA Versions 7.16 and later, the VDA will detect the presence of supported GPU drivers automatically at runtime and use the GPU for graphics rendering and acceleration if available.

Citrix optimizations can be enabled through Citrix policy by specifying “Optimize for 3D graphics workload.”

Note: There is no special installation mode for HDX 3D Pro.

Refer to the product documentation for additional information: See [GPU Acceleration for Windows Single-Session OS](#).

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 41).

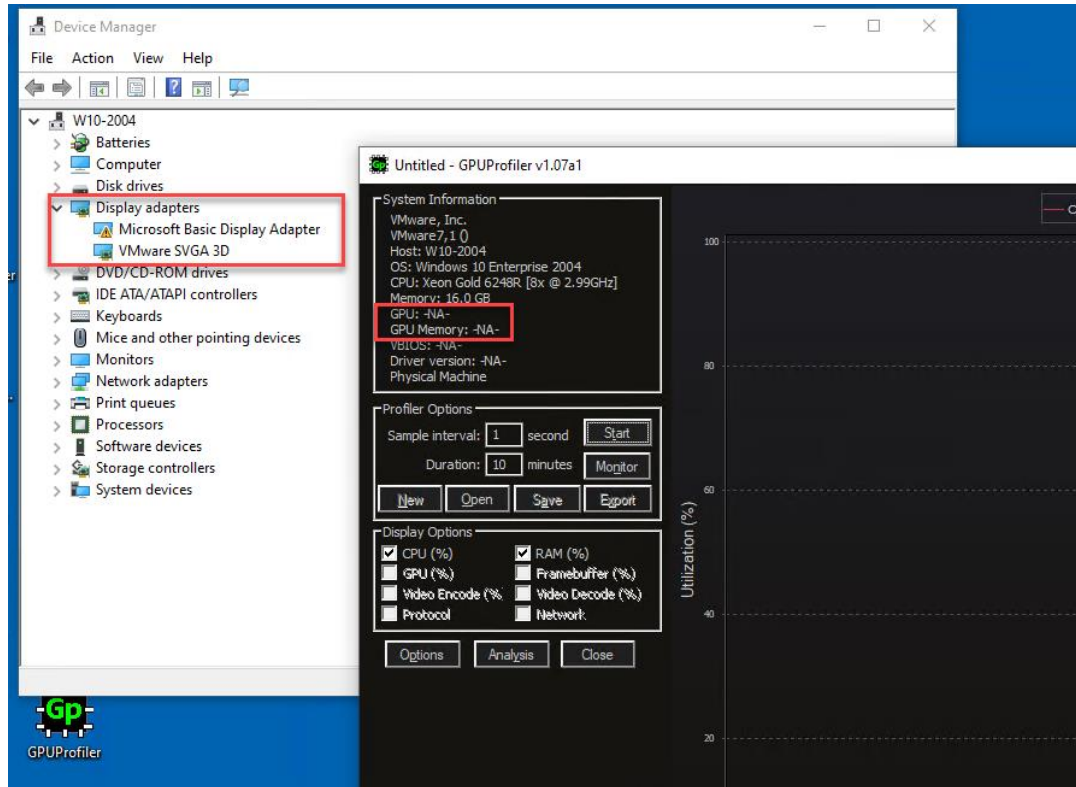


Figure 41.
Device Manager before the NVIDIA driver is installed

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 42).

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-vmware-vmware	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-Is-windows-2018.10.0.25098346 (1)	7/28/2020 12:59 PM
NVIDIA-Is-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 42.
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 43). After the installation has been completed successfully (Figure 44), 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.

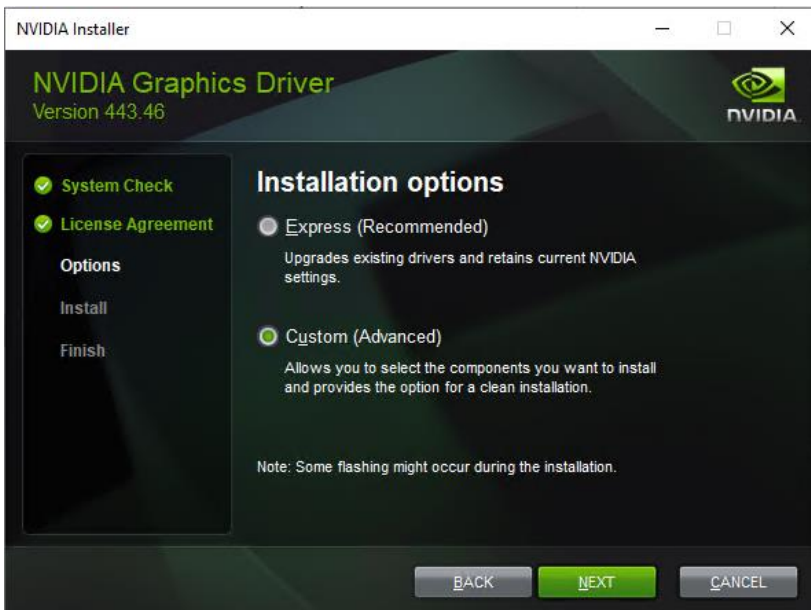


Figure 43.
Select the Express or Custom installation option

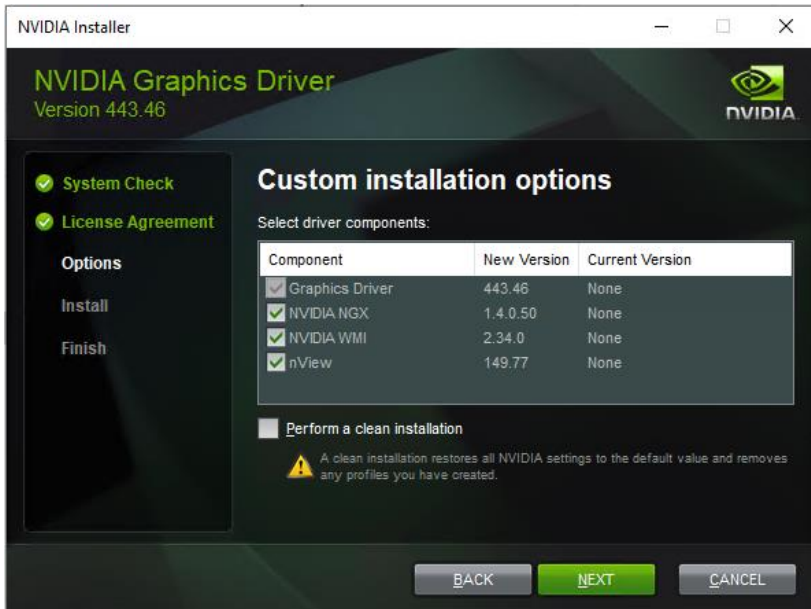


Figure 44.
Components to be installed during NVIDIA graphics driver installation process

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 45).

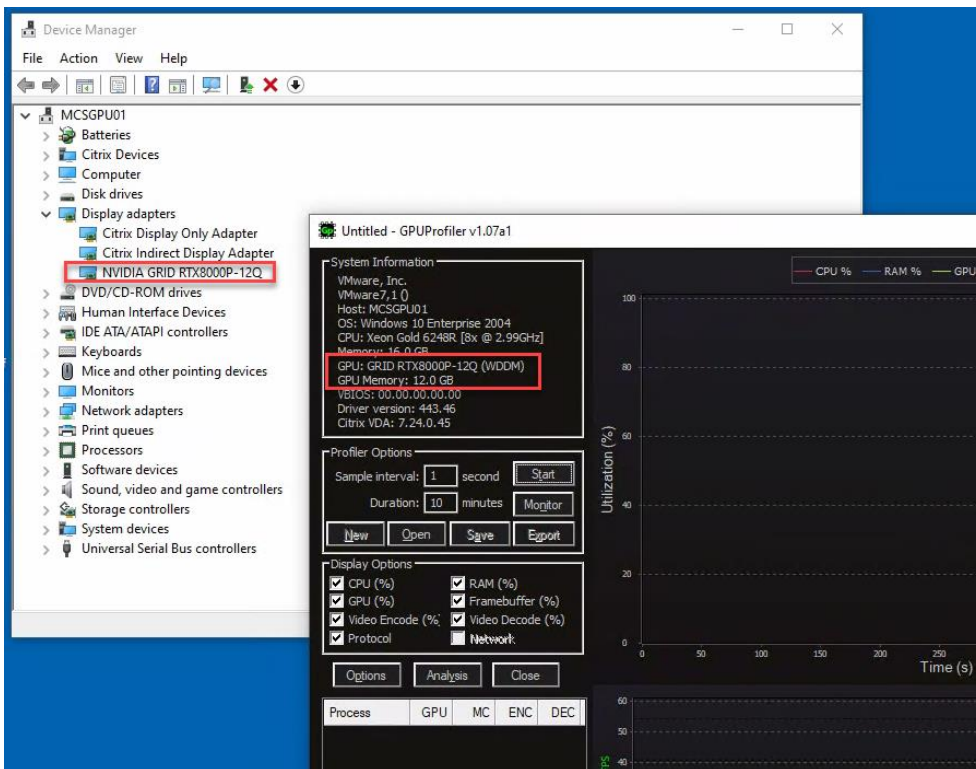
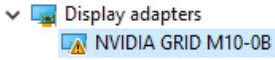


Figure 45.
Validating the driver installation

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



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).

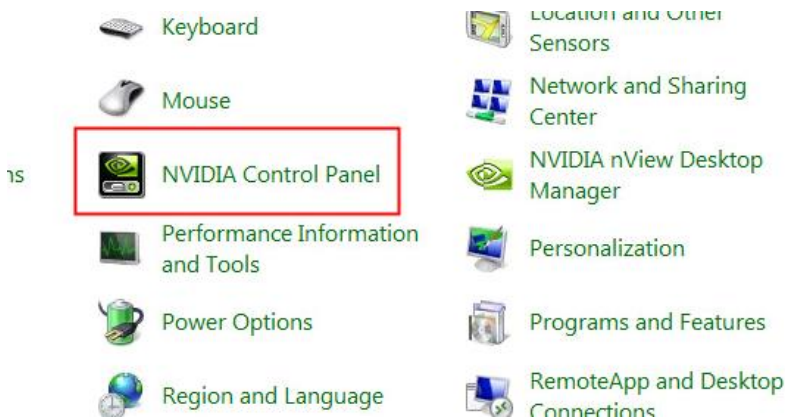


Figure 46.
Choosing NVIDIA Control Panel

2. Select Manage License from the left pane and enter your license server address and port (Figure 47).

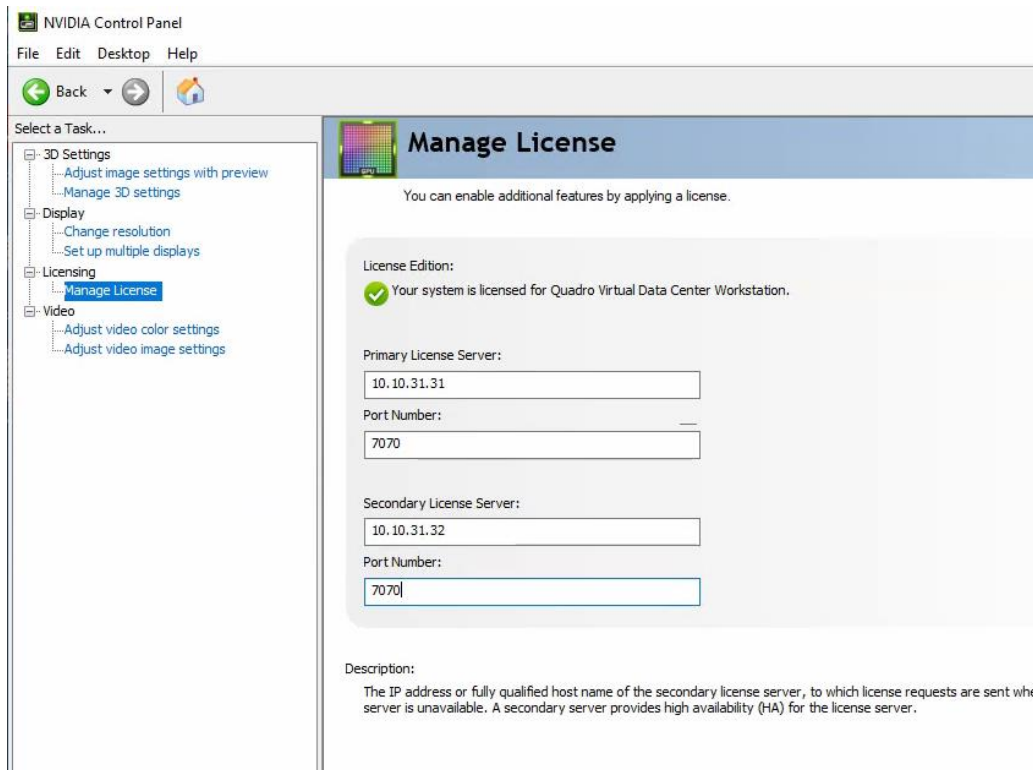


Figure 47.
Managing your license

3. Select Apply.

Verify vGPU deployment

After the desktops are provisioned, use the following steps to verify vGPU deployment in the Citrix XenDesktop 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 48).

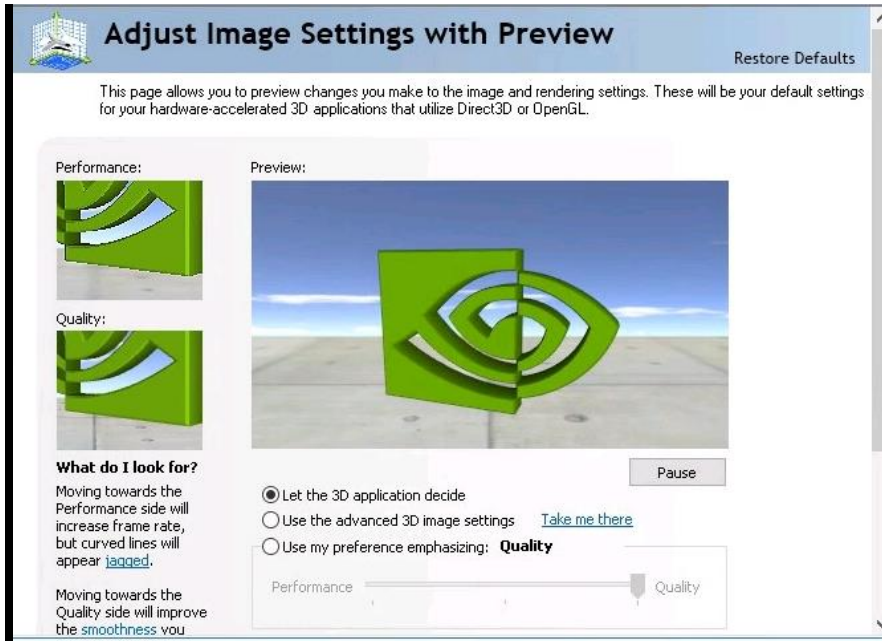


Figure 48.
NVIDIA Control Panel

Verify NVIDIA 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 49).

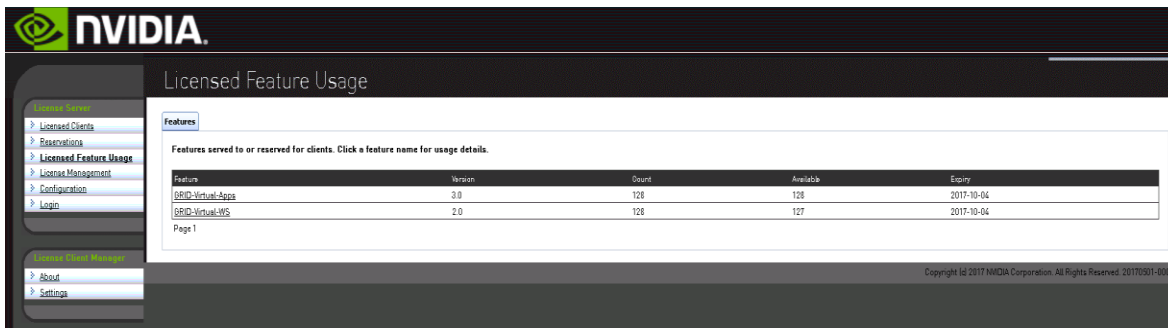


Figure 49.
NVIDIA License Server: Licensed Feature Usage

To view the details, select Licensed Clients in the left pane (Figure 50).

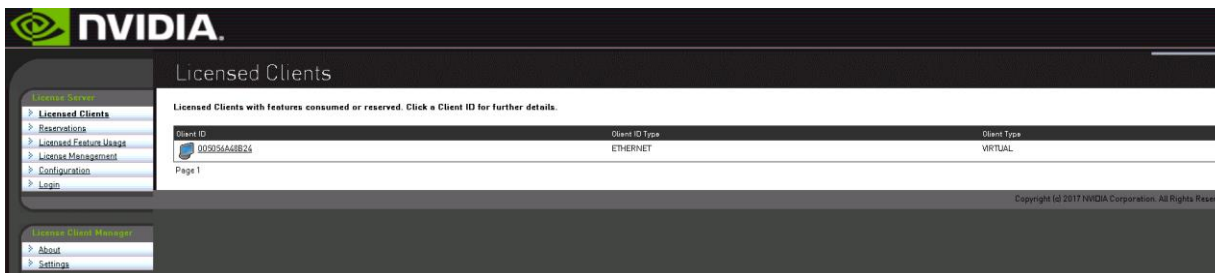


Figure 50.
NVIDIA License Server: Licensed Clients

Verify the NVIDIA configuration on the host

To obtain a hostwide overview of the NVIDIA GPUs, enter the `nvidia-smi` command without any arguments (Figure 51).

```
[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 51.

The `nvidia-smi` command output from the host with two NVIDIA RTX cards and 4 Microsoft Windows 10 desktops with RTX 8000-xQ 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 Citrix HDX Monitor

Use the Citrix HDX Monitor tool (which replaces the Health Check tool) to validate the operation and configuration of HDX visualization technology and to diagnose and troubleshoot HDX problems. To download the tool and learn more about it, go to <https://taas.citrix.com/hdx/download/>.

Optimize the Citrix HDX 3D Pro user experience

You can change various settings to optimize the HDX 3D Pro user experience.

Create a Citrix policy to optimize for 3D graphics workloads (Figure 52). Test other settings with your environment to see if they improve the end user experience.

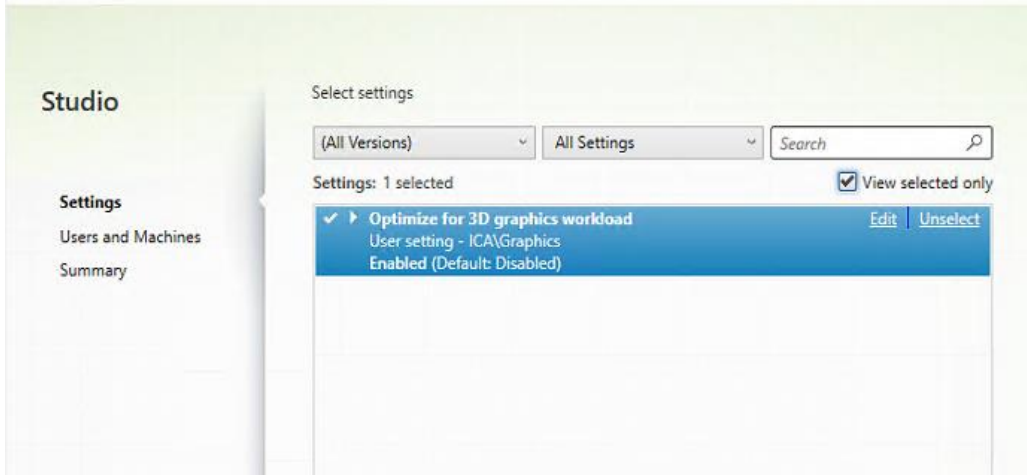


Figure 52.
Citrix HDX 3D Pro policy

To use HDX 3D Pro with multiple monitors, be sure that the host computer is configured with at least as many monitors as are attached to user devices. The monitors attached to the host computer can be either physical or virtual.

Do not attach a monitor (either physical or virtual) to a host computer while a user is connected to the virtual desktop or the application providing the graphical application. Doing so can cause instability for the duration of a user's session.

Let your users know that changes to the desktop resolution (by them or an application) are not supported while a graphical application session is running. After closing the application session, a user can change the resolution of the Desktop Viewer window in Citrix Receiver Desktop Viewer Preferences.

When multiple users share a connection with limited bandwidth (for example, at a branch office), Citrix recommends that you use the “Overall session bandwidth limit” policy setting to limit the bandwidth available to each user. This setting helps ensure that the available bandwidth does not fluctuate widely as users log on and off. Because HDX 3D Pro automatically adjusts to make use of all the available bandwidth, large variations in the available bandwidth over the course of user sessions can negatively affect performance.

For example, if 20 users share a 60-Mbps connection, the bandwidth available to each user can vary between 3 Mbps and 60 Mbps, depending on the number of concurrent users. To optimize the user experience in this scenario, determine the bandwidth required per user at peak periods and limit users to this amount at all times.

For users of a 3D mouse, Citrix recommends that you increase the priority of the generic USB redirection virtual channel to 0. For information about changing the virtual channel priority, see Citrix article CTX128190.

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 the desktops were provisioned, high-end graphics solutions were tested with SPECviewperf 13 Catia and SolidWorks workloads to determine the best practices for sizing high-end graphics workstations using RTX 6000 and 8000 cards.

In these tests, all virtual machines were configured identically with eight vCPUs (four per socket) and 16 GB of RAM.

Our recommendation is that the minimum vGPU profile size that should be used for the RTX cards is 8 GB of frame buffer, or the 8Q RTX profile (Table 6).

Table 6. Maximum recommended virtual machines per RTX card

Recommended Virtual Machines per RTX Card				
	8Q VMs	12Q VMs	24Q VMs	48Q VMs
RTX6000	3	2	1	0
RTX8000	6	4	2	1

Figures 53 through 63 show the results using the RTX 6000 and 8000 cards.

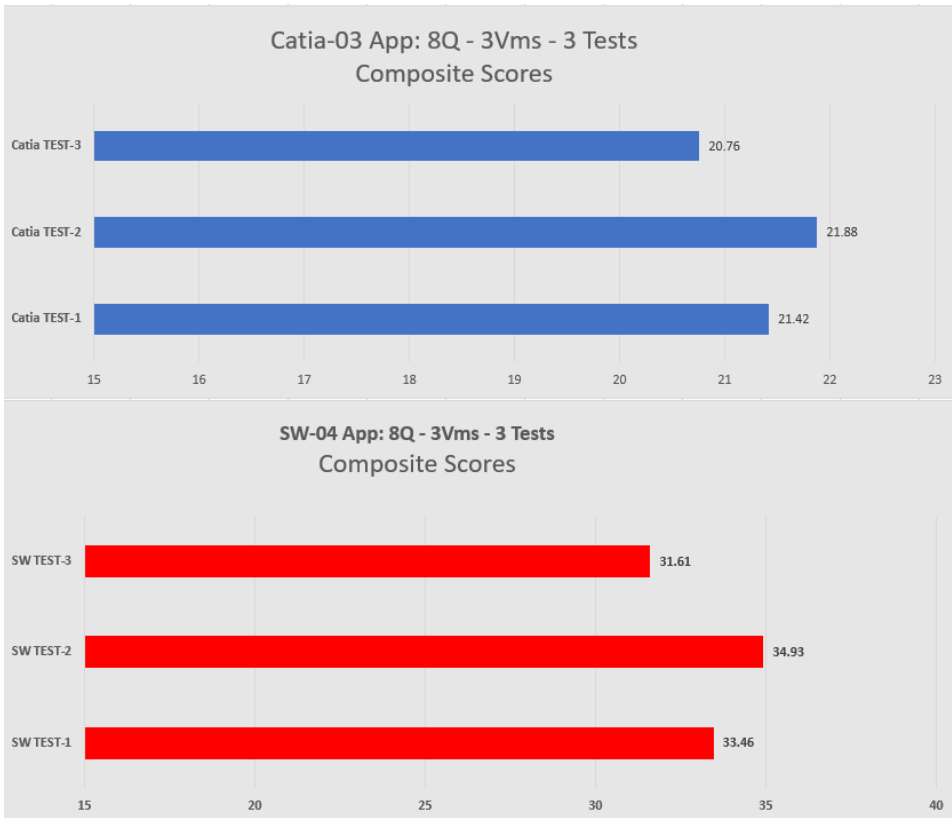


Figure 53.
Composite scores for RTX 6000 and 8000 cards with 8Q vDWS profile



Figure 54.
GPU Profiler: 8Q vDWS profile

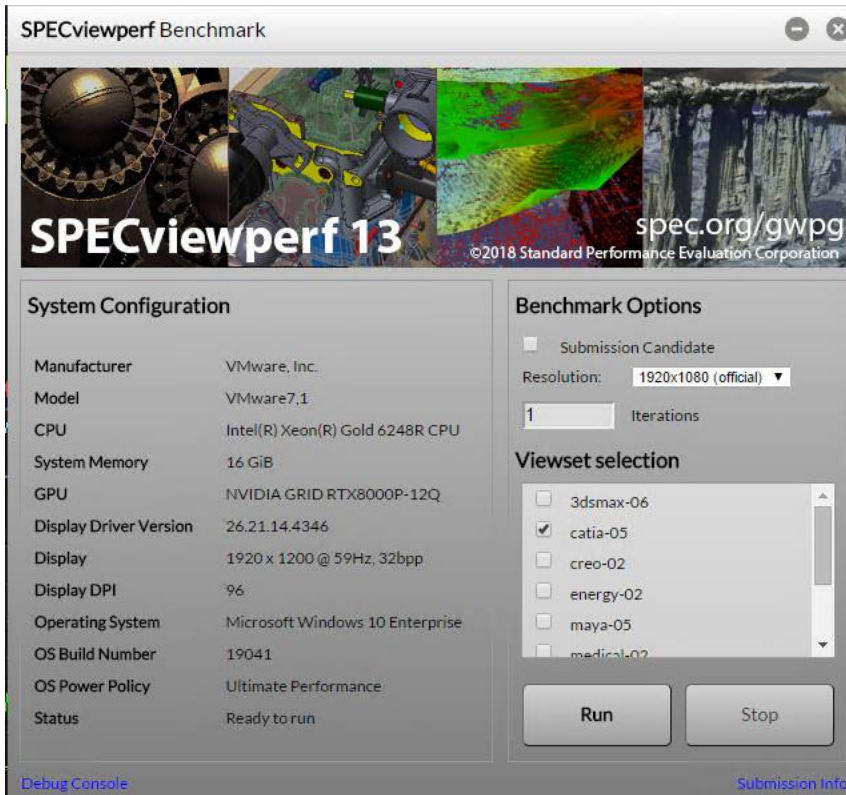


Figure 55.
RTX 6000 and 8000 cards with 12Q vDWS profile

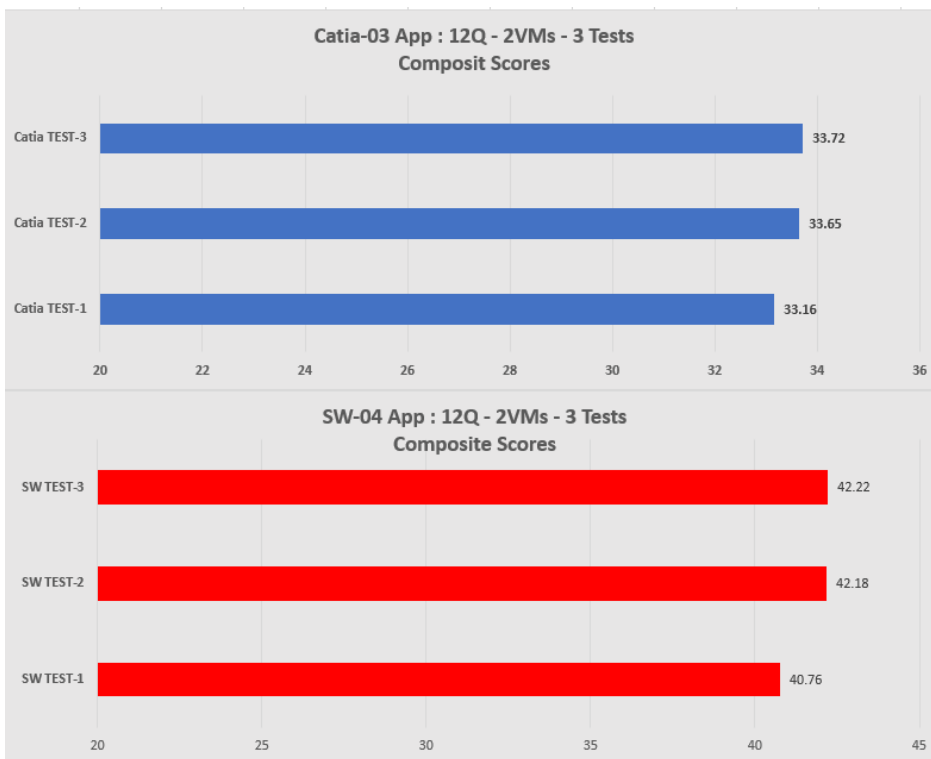


Figure 56.
Composite scores for RTX 6000 and 8000 cards with 12Q vDWS profile

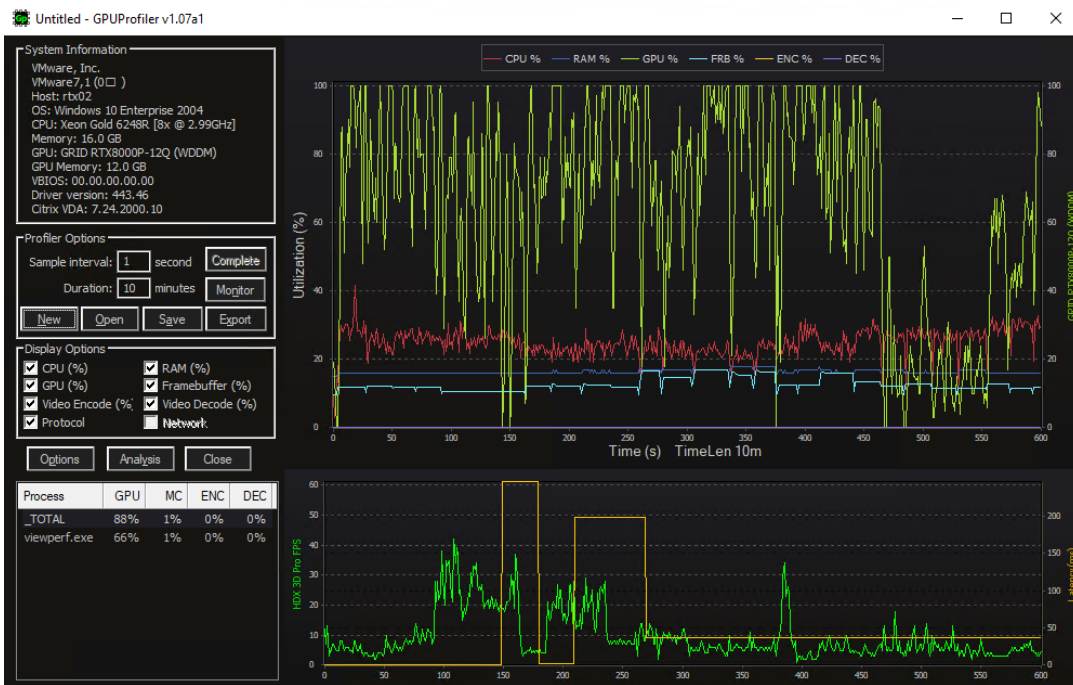


Figure 57.
GPU Profiler: 12Q vDWS profile

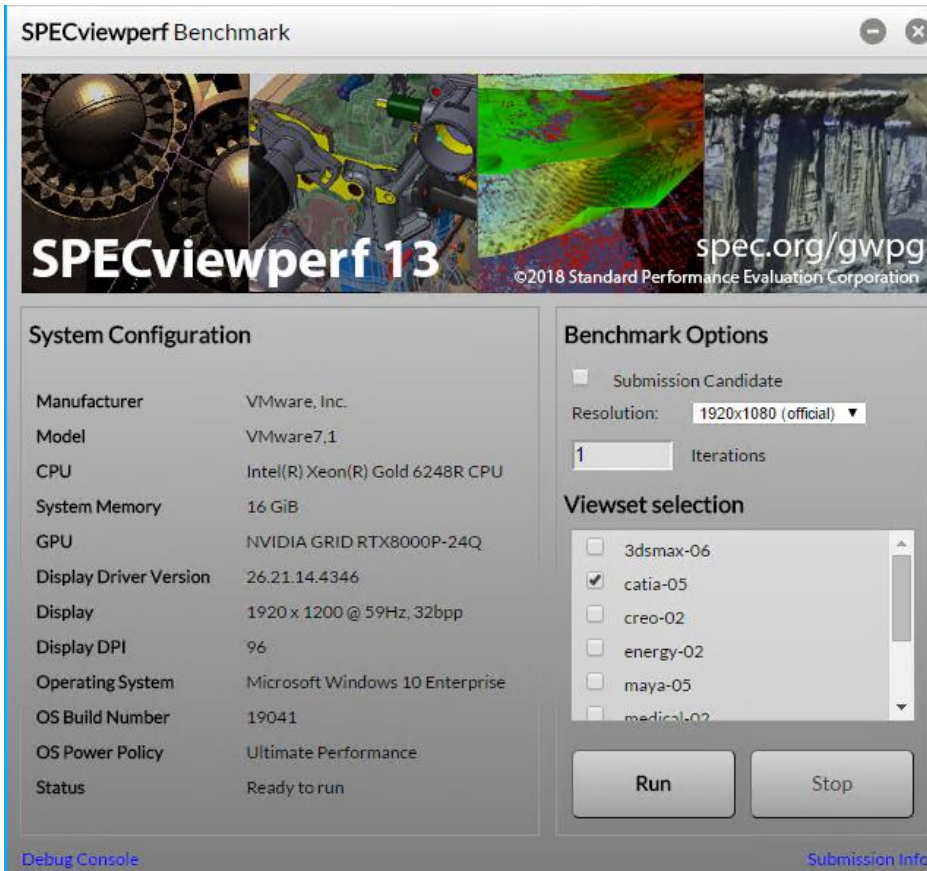


Figure 58.
RTX 6000 and 8000 cards with 24Q vDWS profile

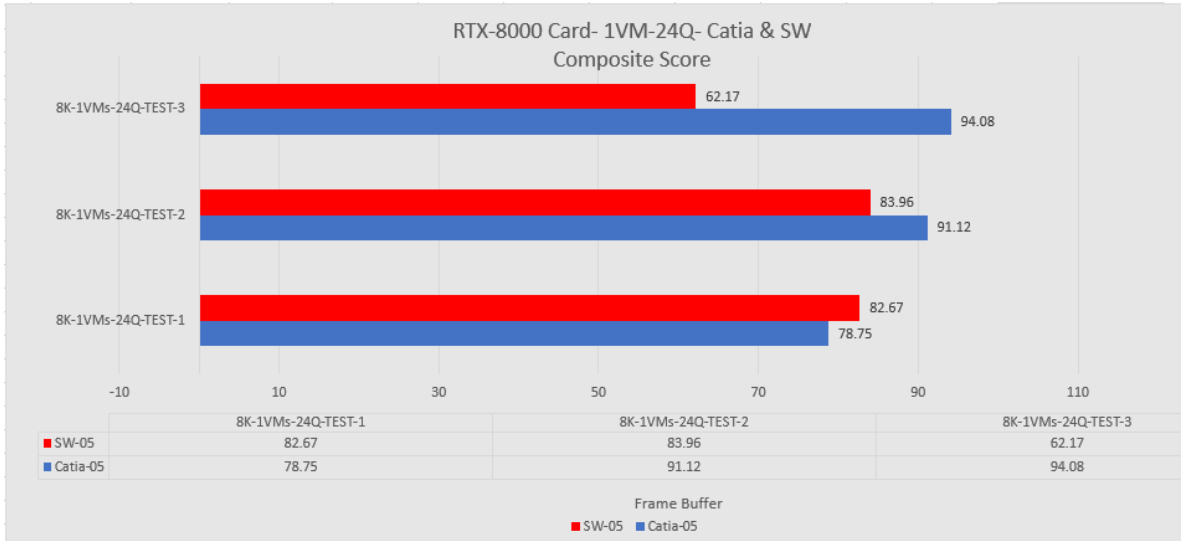


Figure 59.
Composite scores for RTX 6000 and 8000 cards with 24Q vDWS profile



Figure 60.
GPU Profiler: 24Q vDWS profile

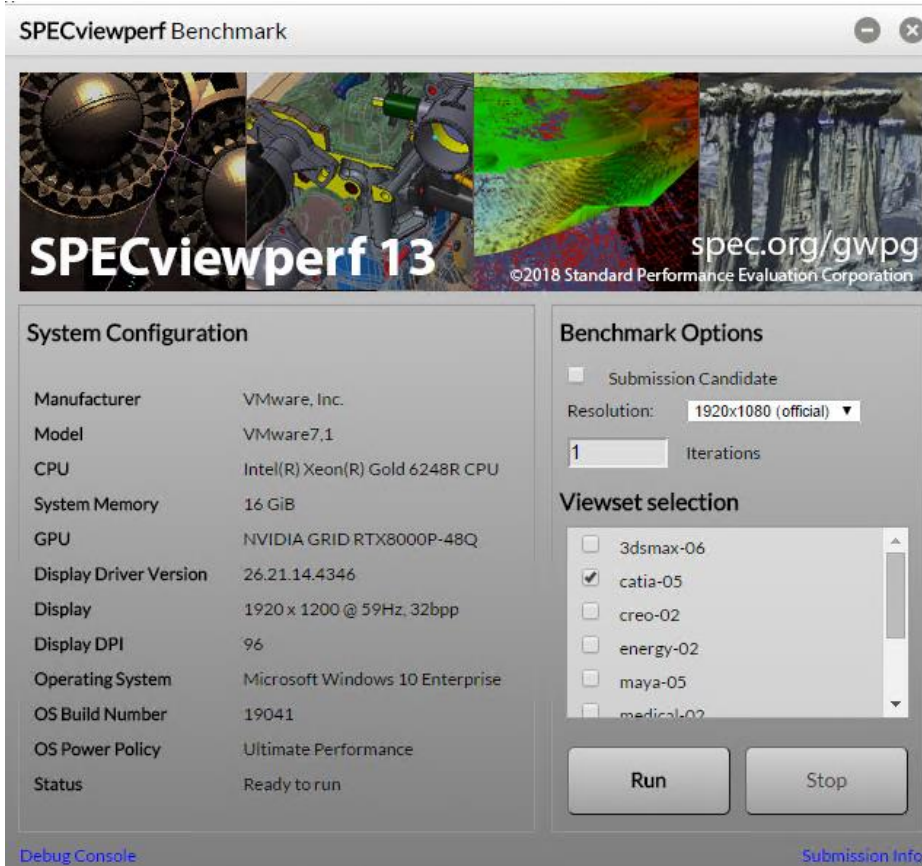


Figure 61.
RTX 8000 card with 48Q vDWS profile



Figure 62.
Composite scores for RTX 8000 card with 48Q vDWS profile



Figure 63.
GPU Profiler: 48Q vDWS 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 Citrix Virtual Apps and Desktops 1912 LTSR 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:
 - <http://www.cisco.com/en/US/products/ps10265/>
- NVIDIA:
 - <http://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>
 - GPUProfiler: <https://github.com/JeremyMain/GPUProfiler/releases/tag/v1.07a2>
- Citrix Virtual Apps and Desktops:

-
- <https://docs.citrix.com/en-us/citrix-virtual-apps-desktops/1912-ltsr.html>
 - <https://support.citrix.com/article/CTX131385>
 - <http://blogs.citrix.com/2014/08/13/citrix-hdx-the-big-list-of-graphical-benchmarks-tools-and-demos/>
 - **Microsoft Windows and Citrix optimization guides for virtual desktops:**
 - <http://support.citrix.com/article/CTX125874>
 - <https://support.citrix.com/article/CTX216252>
 - <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>.

Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: <https://www.cisco.com/go/trademarks>. Third-party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1110R)