

# Integrate Cisco UCS M5 Servers with NVIDIA GRID 5.0 on VMware vSphere 6.5 and Citrix XenDesktop 7.13



# Contents

## What you will learn

### NVIDIA GRID vGPU profiles

#### Cisco Unified Computing System

- Cisco UCS Manager
- Cisco UCS 6332 Fabric Interconnect
- Cisco UCS C-Series Rack Servers
- Cisco UCS C240 M5 Rack Server
- Cisco UCS B200 M5 Blade Server

#### NVIDIA Tesla cards

#### NVIDIA GRID

- NVIDIA GRID 5.0 GPU
- NVIDIA GRID 5.0 license requirements

#### VMware vSphere 6.5

#### Graphics acceleration in Citrix XenDesktop and XenApp

- GPU acceleration for Microsoft Windows desktops
- GPU acceleration for Microsoft Windows Server
- GPU sharing for Citrix XenApp RDS workloads

#### Citrix HDX 3D Pro requirements

#### Solution configuration

#### Configure Cisco UCS

- Create graphics card policy
- Install NVIDIA Tesla GPU card on Cisco UCS C240 M5
- Install NVIDIA Tesla GPU card on Cisco UCS B200 M5
- Configure the GPU card

#### Install the NVIDIA GRID license server

- Install the NVIDIA GRID 5.0 license server
- Configure the NVIDIA GRID 5.0 license server

#### Install NVIDIA GRID software on the VMware ESX host and Microsoft Windows virtual machine

- NVIDIA Tesla P6, P40, and M10 profile specifications
- Prepare a virtual machine for NVIDIA GRID vGPU support
- Install the NVIDIA GRID vGPU software driver
- Verify that applications are ready to support NVIDIA GRID vGPU
- Configure the virtual machine for an NVIDIA GRID vGPU license

#### Verify NVIDIA GRID vGPU deployment

- Verify that the NVIDIA driver is running on the desktop
- Verify NVIDIA license acquisition by desktops
- Verify the NVIDIA configuration on the host

## **Additional configurations**

- Install and upgrade NVIDIA drivers

- Use Citrix HDX Monitor

- Optimize the Citrix HDX 3D Pro user experience

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

- Use the OpenGL Software Accelerator

## **Conclusion**

## **For more information**

## What you will learn

Cisco recently introduced the fifth-generation blade and rack servers based on the Intel® Xeon® Scalable processor architecture. Nearly concurrently, NVIDIA launched new hardware and software designed specifically to use the new server architectures.

Using the increased processing power of the new Cisco UCS® B-Series Blade Servers and C-Series Rack 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 P6, P40, P100, and M10 cards in the Cisco Unified Computing System™ (Cisco UCS) portfolio of PCI Express (PCIe) and mezzanine form-factor cards for the C-Series Rack Servers and B-Series Blade Servers respectively.

With the addition of the new graphics processing capabilities, 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 new NVIDIA M10 high-density graphics card, deployable on Cisco UCS C240 M5 and C480 M5 Rack Servers.

The new graphics capabilities help enable organizations to centralize their graphics workloads and data in the data center. These capabilities greatly benefit 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 the C-Series Rack Servers offer these benefits:

- Support for two to four full-length, full-power NVIDIA GRID cards in 2-rack-unit (2RU) or 4RU form-factor servers
- Support on a single rack server for up to 48 users of high-performance graphics workstations that support graphics processing unit (GPU) processing
- 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 card in the B-Series Blade Servers offers these benefits:

- Support for two NVIDIA P6 mLOM cards per server
- Up to four times the density of GPU-supported user workstations on Cisco UCS B200 M5 Blade Server in a half-width blade than in the previous-generation B200 M4 with a single NVIDIA M6 card (32 versus 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 the design discussed in this document is VMware's support for the NVIDIA GRID virtual GPU (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.



vSphere 6.5 offers some additional new features. Now you can configure host graphics settings, customize vGPU graphics settings on a per virtual machine basis, and view GPU statistics on the Performance tab with the Advanced option.

The purpose of this document is to help our partners and customers integrate NVIDIA GRID 5.0 graphics processing cards, Cisco UCS B200 M5 Blade Servers, and Cisco UCS C240 M5 Rack Servers on VMware vSphere and Citrix XenDesktop 7.13 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 P6, P40, and M10 cards with VMware vSphere and Citrix products so that the servers, hypervisor, and virtual desktops are ready for installation of graphics applications.

## NVIDIA GRID vGPU profiles

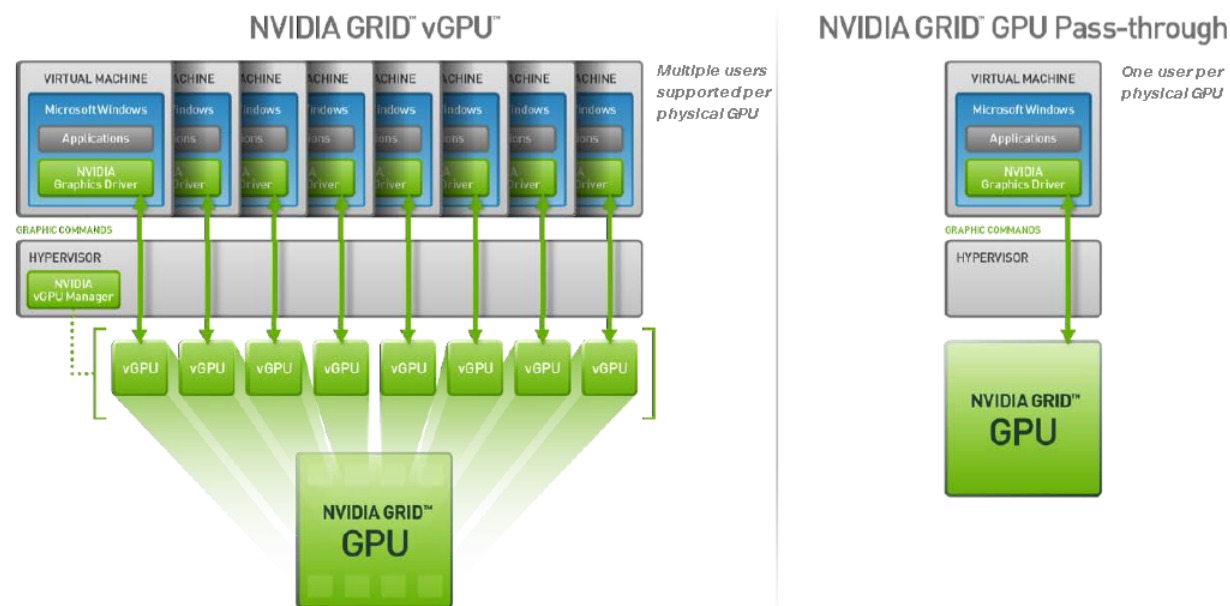
In any given enterprise, the needs of individual users vary widely. One of the main benefits of the GRID vGPU is the flexibility to use various vGPU profiles designed to serve the needs of different classes of end users.

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

- 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).

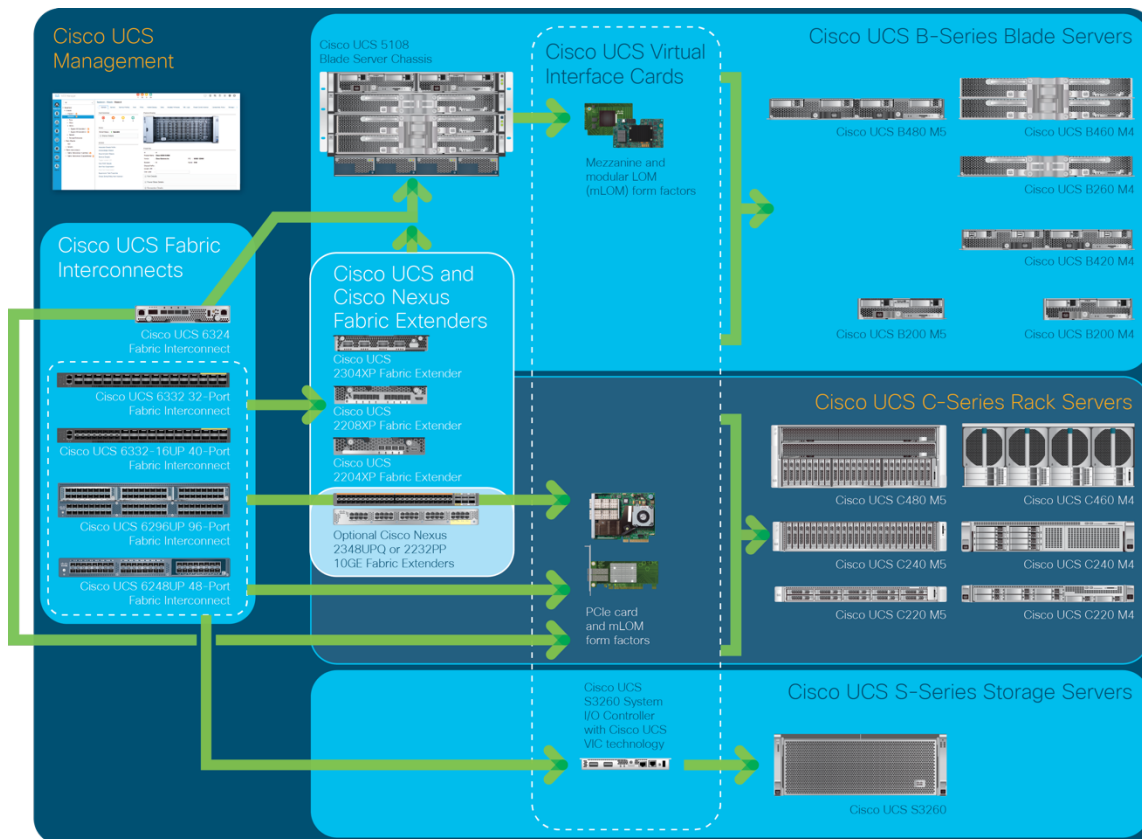
**Figure 1.** NVIDIA GRID vGPU GPU system architecture



## 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:

- **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:

- 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. Tight integration of Cisco UCS Manager and NVIDIA GPU cards enables better management of firmware and graphics card configuration.

### Cisco UCS 6332 Fabric Interconnect

The Cisco UCS 6332 Fabric Interconnect (Figure 3) is the management and communication backbone for Cisco UCS B-Series Blade Servers, C-Series Rack Servers, and 5100 Series Blade Server Chassis. All servers attached to these fabric interconnects become part of one highly available management domain.

Because they support unified fabric, Cisco UCS 6300 Series Fabric Interconnects provide both LAN and SAN connectivity for all servers within their domains. For more information see <https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-b-series-blade-servers/6332-specsheet.pdf>.

Features and capabilities of the fabric interconnects include:

- Bandwidth up to 2.56 Tbps with full-duplex throughput
- Thirty-two 40-Gbps Quad Enhanced Small Form-Factor Pluggable (QSFP+) ports in one 1 rack unit (1RU)
- Support for four 10-Gbps breakout cables
- Ports capable of line-rate, low-latency, lossless 40 Gigabit Ethernet and [FCoE](#)
- Centralized unified management with [Cisco UCS Manager](#)
- Efficient cooling and serviceability

**Figure 3.** Cisco UCS 6332 Fabric Interconnect**Front View****Rear View****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 an increase in processor frequency and improved security and availability features. With the increased performance provided by the [Intel Xeon Scalable processors](#), 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 a 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 (Figures 4 and 5 and Table 1) 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 Intel Xeon Scalable processors, 24 DIMM slots for 2666-MHz DDR4 DIMMs, and up to 128 GB of capacity, up to 6 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 UCS virtual interface card (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 delivers:

- 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
- 24 SFF front-facing SAS and SATA hard disk drives (HDDs) or SAS and SATA solid-state disks (SSDs)
- Optionally, up to two front-facing SFF Non-Volatile Memory Express (NVMe) PCIe SSDs (replacing the SAS and 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 and SATA HDDs and 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 for the following cards on the dedicated mLOM slot on the motherboard:
  - Cisco VICs
  - Four-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 GPUs, providing a graphics-rich 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 (micro-SD) card on PCIe riser 1 (options 1 and 1B)
  - The micro-SD card serves as a dedicated local resource for utilities such as Cisco 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 ministorage module connector on the motherboard supports either:
  - An SD card module with two SD card slots; SD cards with different capacities cannot be mixed
  - An M.2 module with two SATA M.2 SSD slots; M.2 modules with different capacities cannot be mixed

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

The C240 M5 also increases performance and customer choice for many types of storage-intensive applications, such as:

- 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 reduces 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 (front view)



**Figure 5.** Cisco UCS C240 M5 (rear view)



**Table 1.** Cisco UCS C240 M5 PCIe slots

PCIe slot	Length	Lane
1	Half	x8
2	Full	x16
3	half	x8
4	half	x8
5	Full	x16
6	Full	x8

### Cisco UCS VIC 1387

The Cisco UCS VIC 1387 (Figure 6) is a dual-port SFP+ 40-Gbps Ethernet and FCoE-capable PCIe mLOM adapter installed in the Cisco UCS C-Series Rack Servers. The mLOM slot can be 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 more than 256 PCIe standards-compliant interfaces to the host that can be dynamically configured as either NICs or host bus adapters (HBAs). The personality of the card is determined dynamically at boot time using the service profile associated with the server. The number, type (NIC or HBA), identity (MAC address and World Wide Name [WWN]), failover policy, bandwidth, and quality-of-service (QoS) policies of the PCIe interfaces are all determined using the service profile.

For more information about the VIC, see

<https://www.cisco.com/c/en/us/products/interfaces-modules/ucs-virtual-interface-card-1387/index.html>.

**Figure 6.** Cisco UCS VIC 1387



### Cisco UCS B200 M5 Blade Server

Delivering performance, versatility, and density without compromise, the Cisco UCS B200 M5 Blade Server (Figure 7) addresses a broad 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 B200 M5 harnesses the power of the latest Intel Xeon Scalable CPUs, with up to 3072 GB of RAM (using 128-GB DIMMs), two SSDs or HDDs, and up to 80-Gbps throughput connectivity.

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 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 allows you to enable an additional four ports of the VIC 1340, bringing the total capability of the VIC to a dual

native 40-Gbps interface or a dual 4 x 10-Gbps Ethernet port-channel interface. Alternatively, the same rear mezzanine adapter slot can be configured with an NVIDIA P6 GPU.

The B200 M5 has one front mezzanine slot. The 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/dam/en/us/products/collateral/servers-unified-computing/ucs-b-series-blade-servers/b200m5-specsheet.pdf>.

**Figure 7.** Cisco UCS B200 M5 Blade Server (front view)



#### Cisco UCS VIC 1340

The Cisco UCS VIC 1340 (Figure 8) is a 2-port 40-Gbps Ethernet or dual 4 x 10-Gbps Ethernet, 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 VIC 1340 is enabled for two ports of 40-Gbps Ethernet. The VIC 1340 enables a policy-based, stateless, agile server infrastructure that can present more than 256 PCIe standards-compliant interfaces to the host that can be dynamically configured as either NICs or HBAs. In addition, the VIC 1340 supports Cisco Virtual Machine Fabric Extender (VM-FEX) technology, which extends the Cisco UCS fabric interconnect ports to virtual machines, simplifying server virtualization deployment and management.

For more information, see <https://www.cisco.com/c/en/us/products/collateral/interfaces-modules/ucs-virtual-interface-card-1340/datasheet-c78-732517.html>.




**Figure 8.** Cisco UCS VIC 1340



## NVIDIA Tesla cards

For desktop virtualization applications, the NVIDIA Tesla P6, M10, and P40 cards are optimal choices for high-performance graphics (Table 2).

**Table 2.** Technical specifications for NVIDIA Tesla cards

	P6	M10	P40
<b>Number of GPUs</b>	Single midrange Pascal	Quad midlevel Maxwell	Single high-end Pascal
<b>NVIDIA Compute Unified Device Architecture (CUDA) cores</b>	2048	2560 (640 per GPU)	3840
<b>Memory size</b>	16-GB GDDR5	32-GB GDDR5 (8 GB per GPU)	24-GB GDDR5
<b>Maximum number of vGPU instances</b>	16	64	24
<b>Power</b>	75W	225W	250W
<b>Form factor</b>	MXM (blade servers), x16 lanes	PCIe 3.0 dual slot (rack servers), x16 lanes	PCIe 3.0 dual slot (rack servers), x16 lanes
<b>Cooling solution</b>	Bare board	Passive	Passive
<b>H.264 1080p30 streams</b>	24	28	24
<b>Maximum number of users per board</b>	16 (with 1-Gbps profile)	32 (with 1-Gbps profile)	24 (with 1-Gbps profile)
<b>Virtualization use case</b>	Blade optimized	User-density optimized	Performance optimized

## 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 offers the highest 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 5.0 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 5.0 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 information about GRID 5.0 license requirements, see <https://images.nvidia.com/content/grid/pdf/GRID-Licensing-Guide.pdf>

## VMware vSphere 6.5

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 vSphere Hypervisor, VMware virtual machines, vCenter Server, virtual storage, and virtual networking, further extending the core capabilities of the vSphere platform.

The vSphere 6.5 platform includes these features:

- Computing
  - **Increased scalability:** vSphere 6.5 supports larger maximum configuration sizes. Virtual machines support up to 128 virtual CPUs (vCPUs) and 6128 GB of virtual RAM (vRAM). Hosts support up to 576 CPUs and 12 TB of RAM, 1024 virtual machines per host, and 64 nodes per cluster.
  - **Expanded support:** Get expanded support for the latest x86 chip sets, devices, drivers, and guest operating systems. For a complete list of guest operating systems supported, see the VMware Compatibility Guide.
  - **Outstanding graphics:** The NVIDIA GRID vGPU delivers the full benefits of NVIDIA hardware-accelerated graphics to virtualized solutions.
  - **Instant cloning:** Technology built in to vSphere 6.0 lays the foundation for rapid cloning and deployment of virtual machines—up to 10 times faster than what is possible today.
- Storage
  - **Transformation of virtual machine storage:** vSphere Virtual Volumes enable your external storage arrays to become virtual machine aware. Storage policy-based management (SPBM) enables common management across storage tiers and dynamic storage class-of-service (CoS) automation. Together these features enable exact combinations of data services (such as clones and snapshots) to be instantiated more efficiently on a per-virtual machine basis.
- Network
  - **Network I/O control:** New support for per-virtual machine VMware vSphere Distributed Switch (VDS) bandwidth reservation helps ensure isolation and enforce limits on bandwidth.
  - **Multicast snooping:** Support for Internet Group Management Protocol (IGMP) snooping for IPv4 packets and Multicast Listener Discovery (MLD) snooping for IPv6 packets in VDS improves performance and scalability with multicast traffic.
  - **Multiple TCP/IP stacks for VMware vMotion:** Implement a dedicated networking stack for vMotion traffic, simplifying IP address management with a dedicated default gateway for vMotion traffic.



- Availability
  - **vMotion enhancements:** Perform nondisruptive live migration of workloads across virtual switches and vCenter Servers and over distances with a round-trip time (RTT) of up to 100 milliseconds (ms). This support for a dramatically longer RTT—a 10x increase in the supported time—for long-distance vMotion now enables data centers physically located in New York and London to migrate live workloads between one another.
  - **Replication-assisted vMotion:** Customers with active-active replication set up between two sites can perform more efficient vMotion migration, resulting in huge savings in time and resources, with up to 95 percent more efficient migration depending on the amount of data moved.
  - **Fault tolerance (up to four vCPUs):** Get expanded support for software-based fault tolerance for workloads with up to four vCPUs.
- Management
  - **Content library:** This centralized repository provides simple and effective management for content, including virtual machine templates, ISO images, and scripts. With vSphere Content Library, you can now store and manage content from a central location and share content through a publish-and-subscribe model.
  - **Cloning and migration across vCenter:** Copy and move virtual machines between hosts on different vCenter Servers in a single action.
  - **Enhanced user interface:** vSphere Web Client is more responsive, more intuitive, and simpler than ever before.

## Graphics acceleration in Citrix XenDesktop and XenApp

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 virtual delivery agent [VDA] supports GPU acceleration of DirectX only.)

Examples of 3D professional applications include:

- 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 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 the [Citrix vGPU 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.

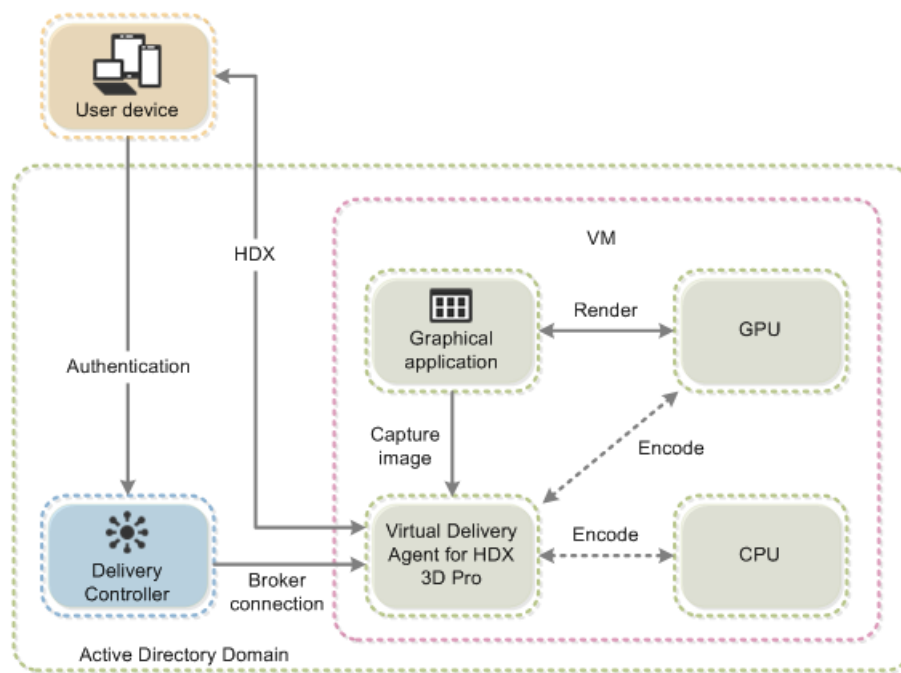
The lossless switch enables the user to change to Always Lossless mode at any time in the session. To select or deselect Always Lossless at any time in a session, right-click the icon or use the shortcut Alt+Shift+1.

- 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 services (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 9:

- 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 9.** 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 XenApp 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/>.

## 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 XenServer; VMware vSphere and ESX, where it is referred to as virtual direct graphics acceleration (vDGA); and Microsoft Hyper-V in Microsoft Windows Server 2016, where it is referred to as discrete device assignment (DDA).
- vGPU is available with Citrix XenServer and VMware vSphere; see <https://www.citrix.com/products/xenapp-xendesktop/hdx-3d-pro.html>.
- 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 the 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.

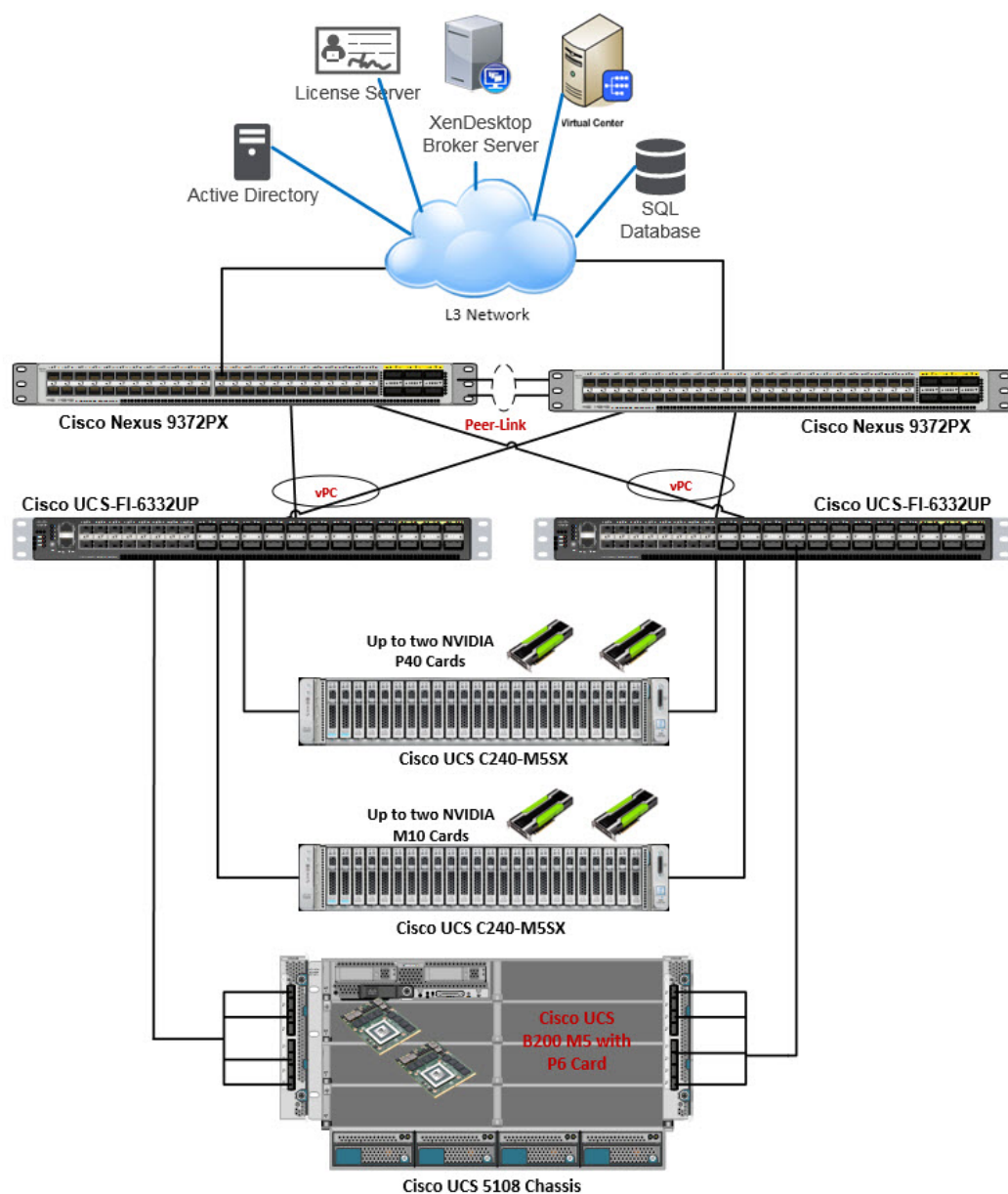
For more information, see the Citrix HDX 3D Pro articles at <http://docs.citrix.com/en-us/xenapp-and-xendesktop/7-12/hdx/hdx-3d-pro.html> and <http://www.citrix.com/xenapp/3>.



## Solution configuration

Figure 10 provides an overview of the solution configuration.

**Figure 10.** Reference architecture



The hardware components in the solution are:

- Cisco UCS C240 M5 Rack Server (two Intel Xeon Scalable processor Platinum 8176 CPUs at 2.10 GHz) with 768 GB of memory (32 GB x 24 DIMMs at 2666 MHz)
- Cisco UCS B200 M5 Blade Server (two Intel Xeon Scalable processor Platinum 8176 CPUs at 2.10 GHz) with 768 GB of memory (32 GB x 24 DIMMs at 2666 MHz)
- Cisco UCS VIC 1387 mLOM (Cisco UCS C240 M5)
- Cisco UCS VIC 1340 mLOM (Cisco UCS B200 M5)
- Two Cisco UCS 6332 third-generation fabric interconnects
- NVIDIA Tesla M10, P6, and P40 cards
- Two Cisco Nexus® 9372 Switches (optional access switches)

The software components of the solution are:

- Cisco UCS Firmware Release 3.2(1a)
- VMware ESXi 6.5 (5969303) for VDI hosts
- Citrix XenDesktop 7.13
- Microsoft Windows 10 64-bit
- Microsoft Server 2016
- NVIDIA GRID 5.0 software and licenses:
  - NVIDIA-VMware-384.73-1OEM.650.0.0.4598673.x86\_64.vib
  - 385.41\_grid\_win10\_server2016\_64bit\_international

## Configure Cisco UCS

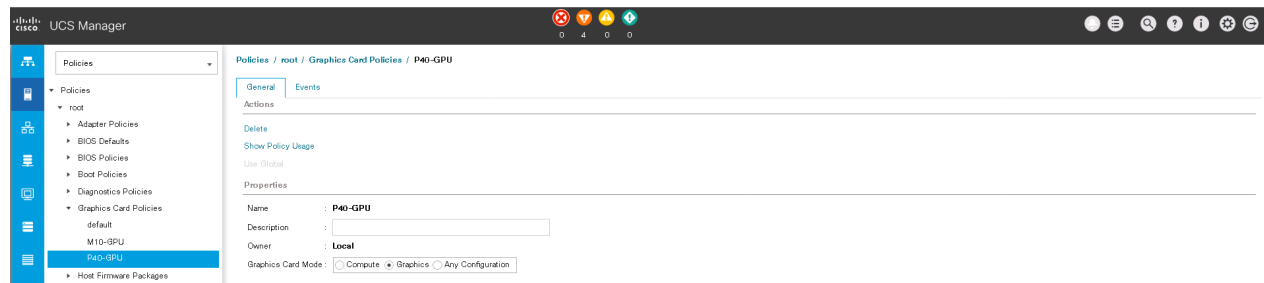
This section describes the Cisco UCS configuration.

### Create graphics card policy

Create a new graphics card policy with the desired mode for the graphics card.

For VDI deployment, select Graphics mode (Figure 11).

**Figure 11.** Graphics card policy



### Install NVIDIA Tesla GPU card on Cisco UCS C240 M5

Install the M10 or P40 GPU card on the Cisco UCS C240 M5 server.

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

**Table 3.** Minimum server firmware versions required for GPU cards

Cisco Integrated Management Controller (IMC)	BIOS minimum version
NVIDIA Tesla M10	Release 3.2(1d)
NVIDIA Tesla P40	Release 3.2(1d)

The rules for mixing NVIDIA GPU cards are as follows:

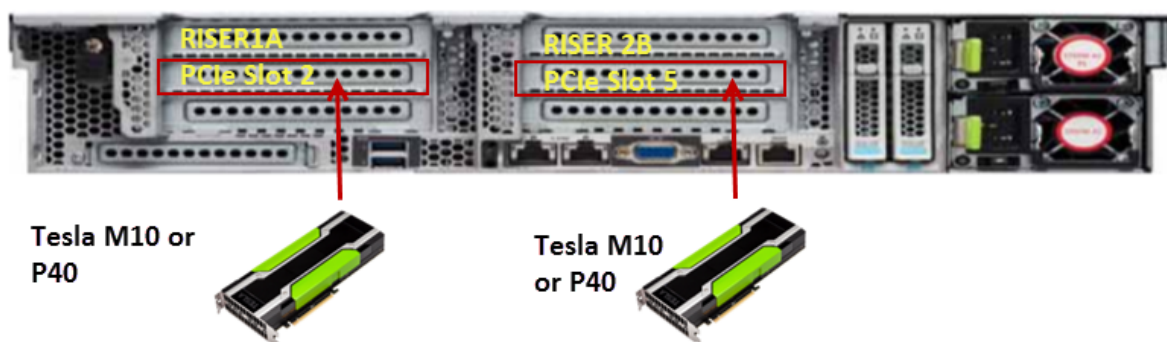
- Do not mix GRID GPU cards with Tesla GPU cards in the same server.
- Do not mix different models of Tesla GPU cards in the same server.

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/2B, slot 5	Riser 1A, slot 2, and riser 2A/2B, slot 5

**Figure 12.** One-GPU scenario**Figure 13.** Two-GPU scenario

For more information, refer to the C240 M5 server installation and configuration document:

[https://www.cisco.com/c/en/us/td/docs/unified\\_computing/ucs/c/hw/C240M5/install/C240M5.pdf](https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/c/hw/C240M5/install/C240M5.pdf).

## Install NVIDIA Tesla GPU card on Cisco UCS B200 M5

Install the P6 GPU card on the Cisco UCS B200 M5 server.

Table 5 lists the minimum firmware version required for the GPU card. Figure 16 shows the card in the server.

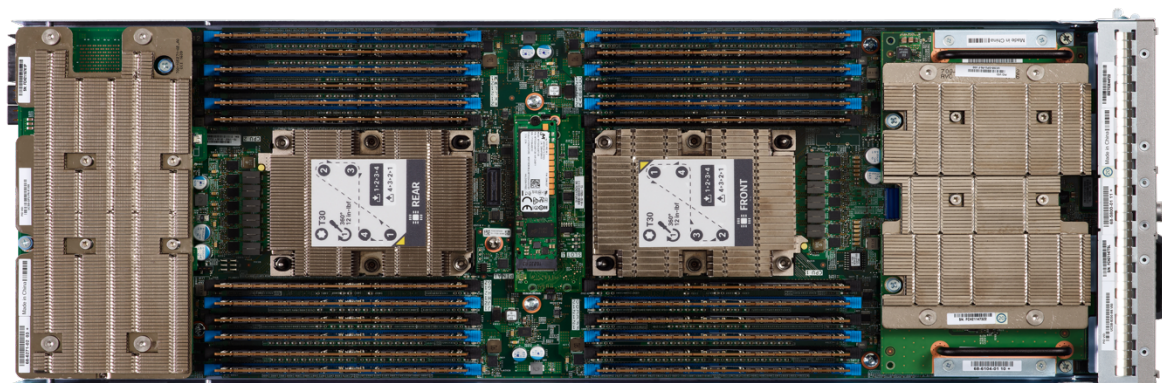
**Table 5.** Minimum server firmware versions required for GPU card

Cisco IMC	BIOS minimum version
NVIDIA Tesla P6	Release 3.2(1d)

Before installing the NVIDIA P6 GPU, do the following:

- Remove any adapter card, such as a Cisco UCS VIC 1380 or 1280 or port extender card from mLOM slot 2. You cannot use any other card in slot 2 when the NVIDIA P6 GPU is installed.
- Upgrade your Cisco UCS instance to a version of Cisco UCS Manager that supports this card. Refer to the latest version of the release notes for Cisco UCS software at the following URL for information about supported hardware:  
<http://www.cisco.com/c/en/us/support/servers-unified-computing/ucs-manager/products-release-notes-list.html>

**Figure 14.** Cisco UCS B200 M5 Blade Server with two P6 GPU cards



## Configure the GPU card

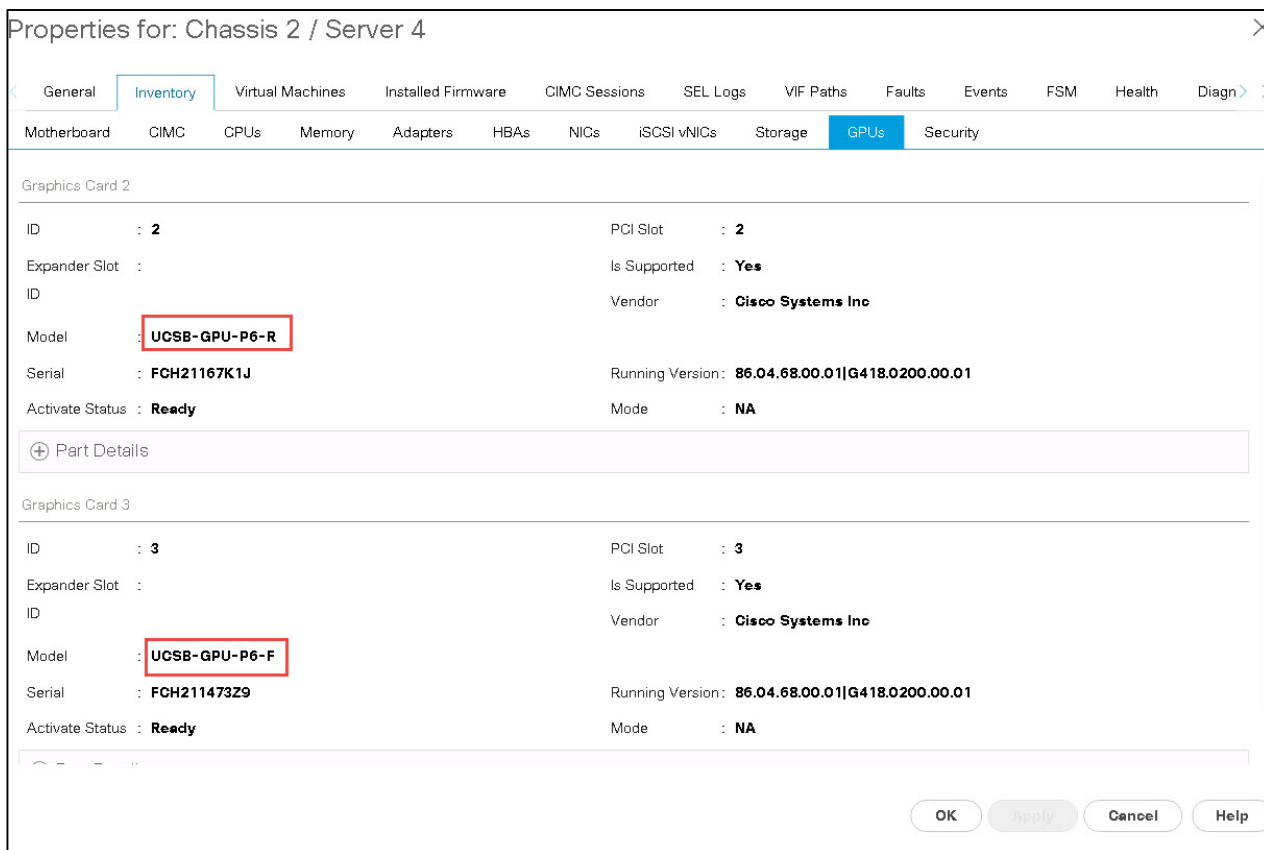
Follow the steps in this section to configure the GPU card.

### NVIDIA P6 installation

After the NVIDIA P6 GPU card is physically installed and the Cisco UCS B200 M5 Blade Server is discovered in Cisco UCS Manager, select the server and choose Inventory > GPUs. As shown in Figure 15, PCIe slots 2 and 3 are used with two GRID P6 cards.

**Note:** Notice, as highlighted in the red boxes in Figure 17, that the NVIDIA P6 GPU card has different part numbers for the front and rear GPU cards.

**Figure 15.** NVIDIA GRID P6 card inventory displayed in Cisco UCS Manager



Properties for: Chassis 2 / Server 4

General Inventory Virtual Machines Installed Firmware CIMC Sessions SEL Logs VIF Paths Faults Events FSM Health Diagn >

Motherboard CIMC CPUs Memory Adapters HBAs NICs iSCSI vNICs Storage **GPUs** Security

Graphics Card 2

ID	: 2	PCI Slot	: 2
Expander Slot	:	Is Supported	: Yes
ID	:	Vendor	: Cisco Systems Inc
Model	: <b>UCSB-GPU-P6-R</b>		
Serial	: FCH21167K1J	Running Version	: 86.04.68.00.01 G418.0200.00.01
Activate Status	: Ready	Mode	: NA

+ Part Details

Graphics Card 3

ID	: 3	PCI Slot	: 3
Expander Slot	:	Is Supported	: Yes
ID	:	Vendor	: Cisco Systems Inc
Model	: <b>UCSB-GPU-P6-F</b>		
Serial	: FCH21147329	Running Version	: 86.04.68.00.01 G418.0200.00.01
Activate Status	: Ready	Mode	: NA

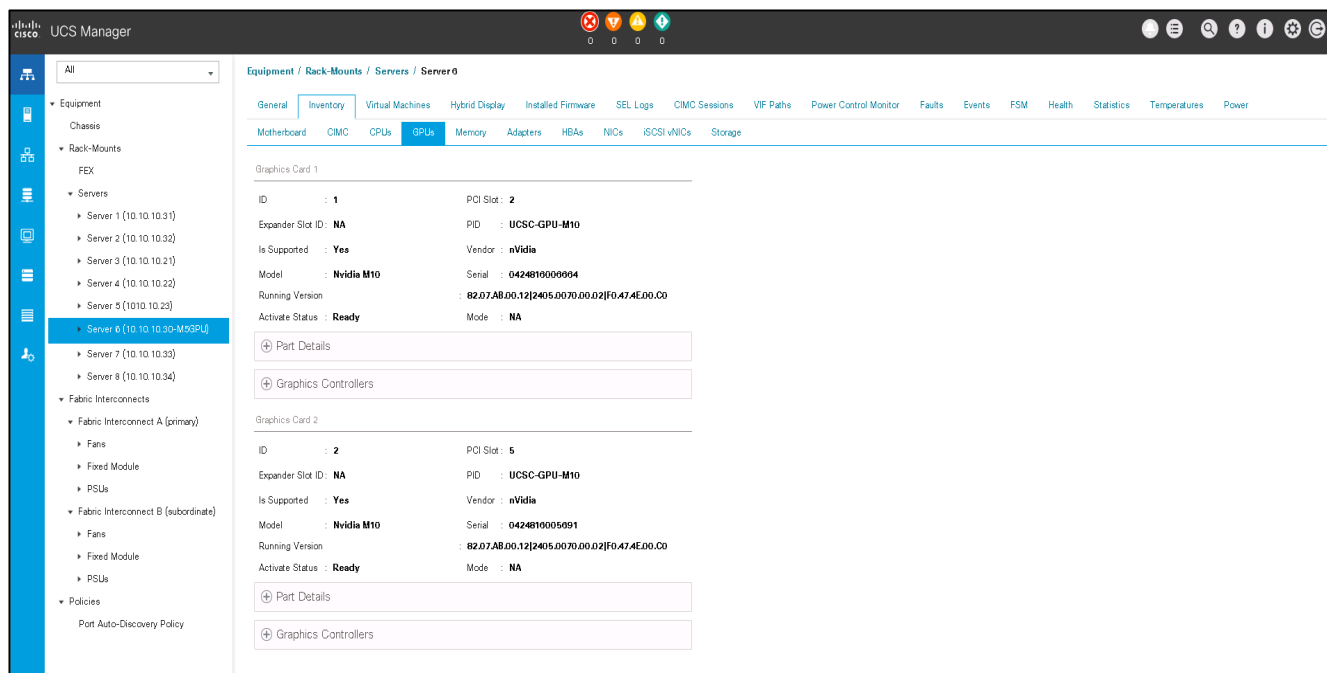
OK Apply Cancel Help



## NVIDIA M10 installation

After the NVIDIA M10 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 16, PCIe slots 2 and 5 are used with two GRID M10 cards.

**Figure 16.** NVIDIA GRID M10 card inventory displayed in Cisco UCS Manager



The screenshot displays the Cisco UCS Manager interface. On the left, a navigation pane shows the hierarchy: Equipment > Servers > Server 6 (10.10.10.30-M50GPU). The main content area is titled 'Equipment / Rack-Mounts / Servers / Server 6' and shows the 'Inventory' tab for 'GPUs'. It lists two graphics cards:

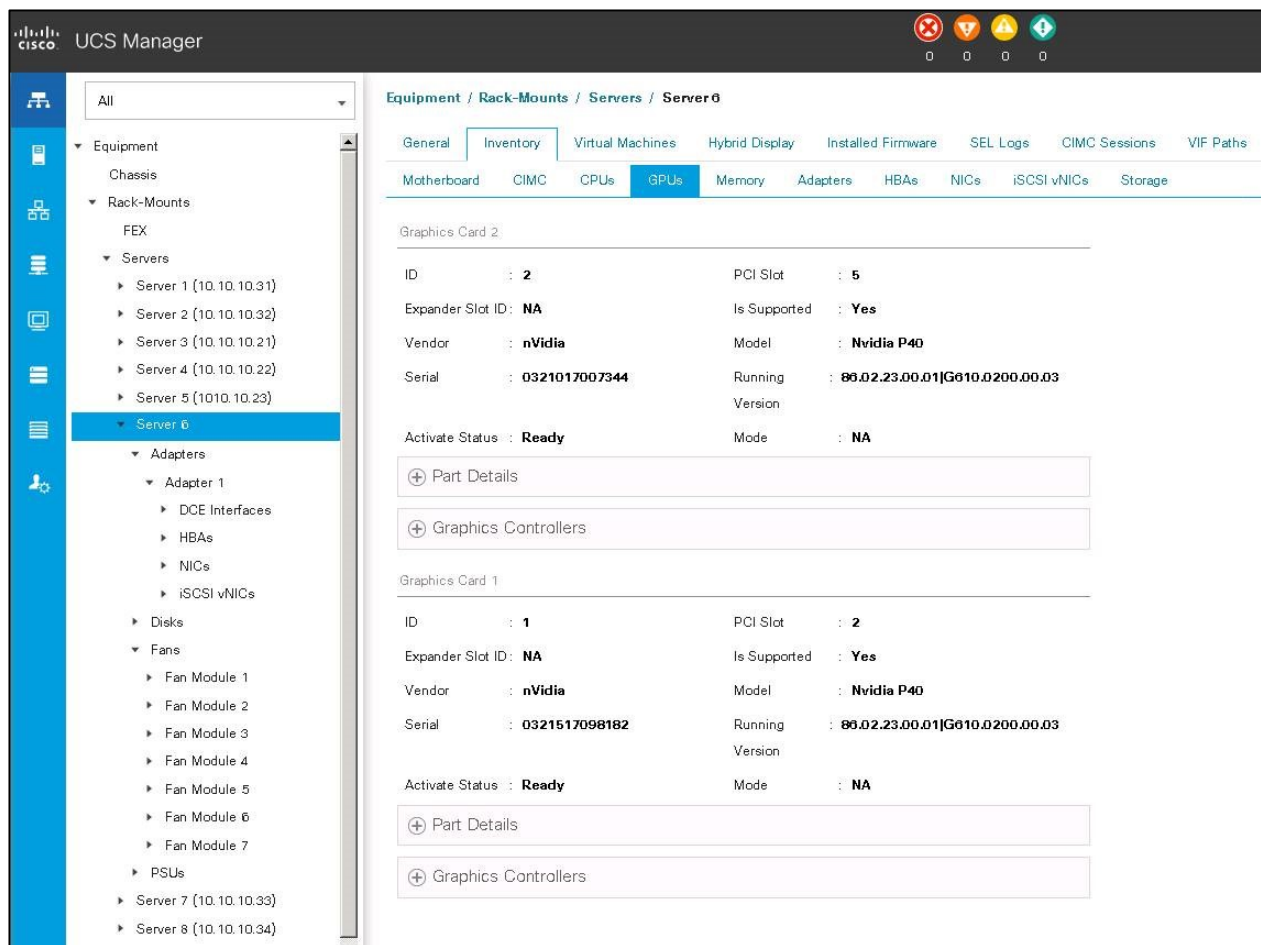
Graphics Card 1	
ID	1
PCI Slot	2
Expander Slot ID	NA
PID	UCSC-GPU-M10
Is Supported	Yes
Vendor	nVidia
Model	Nvidia M10
Serial	0424819009994
Running Version	82.07.AB.00.12[2405.00.70.00.02]F0.47.4E.00.C0
Activate Status	Ready
Mode	NA

Below the first card, there are buttons for 'Part Details' and 'Graphics Controllers'. The second card, 'Graphics Card 2', has identical information, with PCI Slot 5 and Serial 0424819005091.

## NVIDIA M40 installation

After the NVIDIA P40 GPU card is 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 17, PCIe slots 2 and 5 are used with the two GRID P40 cards.

**Figure 17.** NVIDIA GRID P40 card inventory displayed in Cisco UCS Manager



The screenshot shows the Cisco UCS Manager interface. On the left, a navigation pane lists the hierarchy: Equipment > Rack-Mounts > Servers > Server 6. The main content area is titled 'Equipment / Rack-Mounts / Servers / Server 6' and has tabs for General, Inventory, Virtual Machines, Hybrid Display, Installed Firmware, SEL Logs, CIMC Sessions, and VIF Paths. Under the 'Inventory' tab, there are sub-tabs for Motherboard, CIMC, CPUs, GPUs, Memory, Adapters, HBAs, NICs, iSCSI vNICs, and Storage. The 'GPUs' sub-tab is selected, showing details for two graphics cards.

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
+ Part Details			
+ Graphics Controllers			

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
+ Part Details			
+ Graphics Controllers			

Firmware management for NVidia GPU cards is no different than firmware management for Cisco UCS server components. You can use the same Cisco UCS Manager policy-based configuration to manage the firmware versions of NVIDIA cards as shown in the support matrix for Cisco UCS Manager.

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 5.0 license server.

The NVIDIA GRID vGPU is a licensed feature on Tesla P6, P40, and M10. A software license is required to use the full vGPU feature set 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 - utm_source=shorturl&utm_medium=referrer&utm_campaign=grideval).

Three packages are required for VMware ESXi host setup, as shown in Figure 18:

- 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

**Figure 18.** Software required for NVIDIA GRID 5.0 setup on the VMware ESXi host

Name	Date modified	Type	Size
 NVIDIA-v5GA-vSphere-6.0-Host_Driver-384.73	9/1/2017 8:13 AM	Compressed (zippe...	
 NVIDIA-vGPU-VMware_ESXi_6.0_Host_Driver_384.73-1OEM.600.0.0.2494585-offline_bundle	8/29/2017 1:08 AM	Compressed (zippe...	
 NVIDIA-vGPU-VMware_ESXi_6.0_Host_Driver_384.73-1OEM.600.0.0.2494585.vib	8/29/2017 1:08 AM	VIB File	
 NVIDIA-Is-windows-2017.08-0001	9/6/2017 8:34 AM	Compressed (zippe...	
 NVIDIA-Linux-x86_64-384.73-grid.run	8/21/2017 1:59 PM	RUN File	
 385.41_grid_win10_server2016_64bit_international	8/31/2017 1:09 AM	Application	
 385.41_grid_win10_32bit_international	8/31/2017 1:07 AM	Application	
 385.41_grid_win8_win7_server2012R2_server2008R2_64bit_international	8/31/2017 1:07 AM	Application	
 385.41_grid_win8_win7_32bit_international	8/31/2017 1:06 AM	Application	
 384.73-385.41-grid-vgpu-user-guide	8/30/2017 1:59 PM	Firefox HTML Docu...	
 384.73-385.41-grid-vgpu-release-notes-vmware-vsphere	8/30/2017 1:58 PM	Firefox HTML Docu...	
 384.73-385.41-grid-software-quick-start-guide	8/30/2017 1:55 PM	Firefox HTML Docu...	
 384.73-385.41-grid-licensing-user-guide	8/30/2017 2:02 PM	Firefox HTML Docu...	
 384.73-385.41-grid-license-server-user-guide	8/30/2017 2:01 PM	Firefox HTML Docu...	
 384.73-385.41-grid-license-server-release-notes	8/30/2017 2:00 PM	Firefox HTML Docu...	
 384.73-385.41-grid-gpumodeswitch-user-guide	8/30/2017 2:00 PM	Firefox HTML Docu...	





## Install the NVIDIA GRID 5.0 license server

The steps shown here use the Microsoft Windows version of the license server installed on Windows Server 2012 R2. A Linux version of the license server is also available.

The GRID 5.0 license server requires Java Version 7 or later. Go to [Java.com](http://Java.com) and install the latest version.

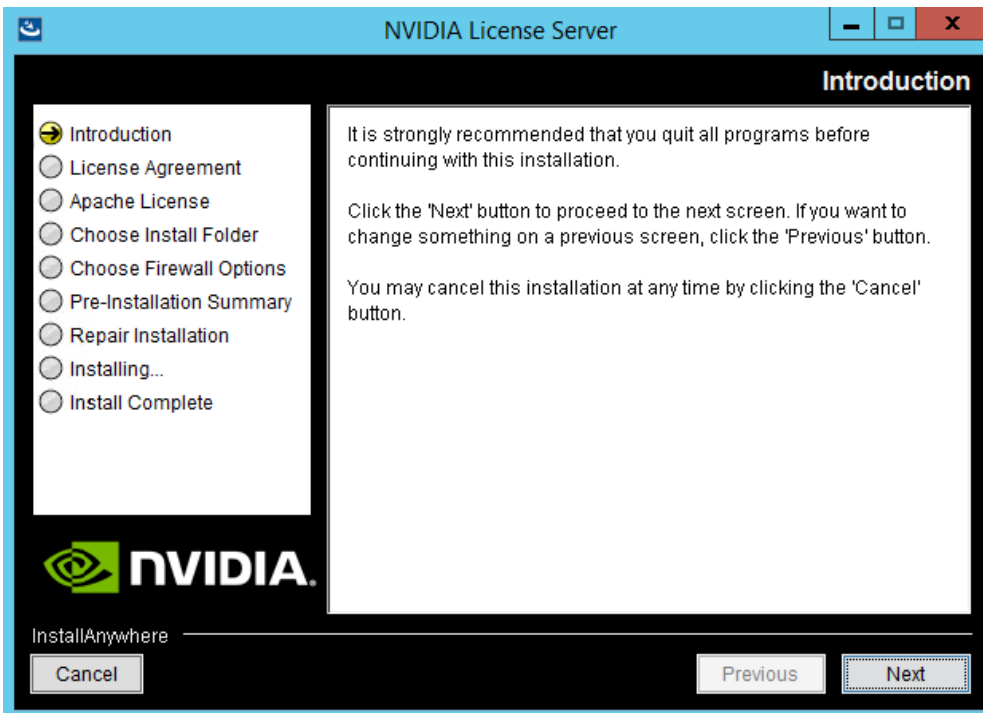
1. Extract and open the NVIDIA-Is-windows-2017.06-0001 folder. Run setup.exe (Figure 19).

**Figure 19.** Run setup.exe

Name ^	Type	Compressed size	Password p...	Size
 384.73-385.4-grid-license-server-release-notes	Firefox HTML Document	1,492 KB	No	
 384.73-385.4-grid-license-server-user-guide	Firefox HTML Document	2,984 KB	No	
 384.73-385.4-grid-licensing-user-guide	Firefox HTML Document	1,950 KB	No	
 setup	Application	237,356 KB	No	

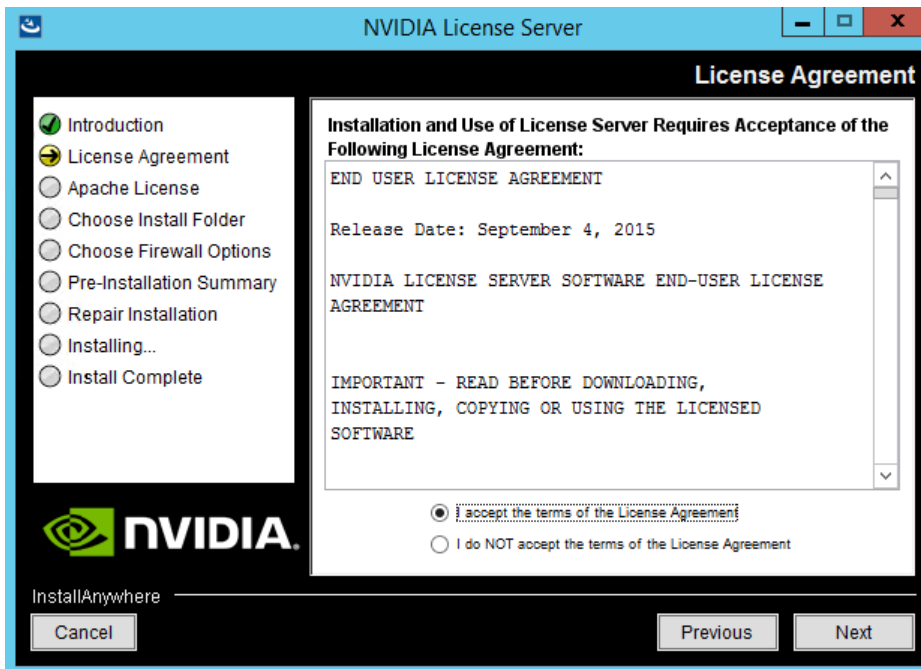
2. Click Next (Figure 20).

**Figure 20.** NVIDIA License Server



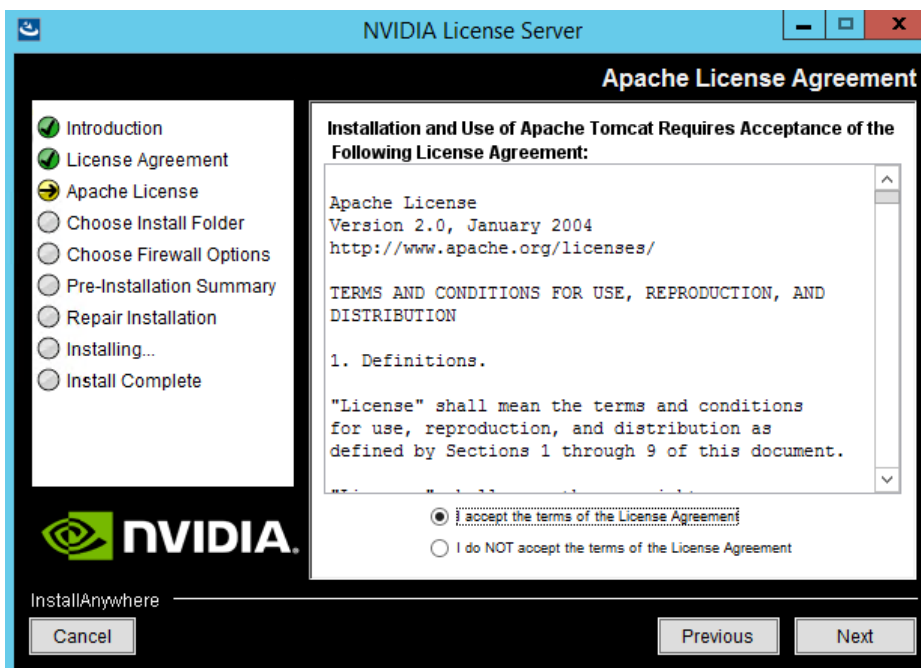
3. Accept the license agreement and click Next (Figure 21).

**Figure 21.** Accepting the NVIDIA license agreement



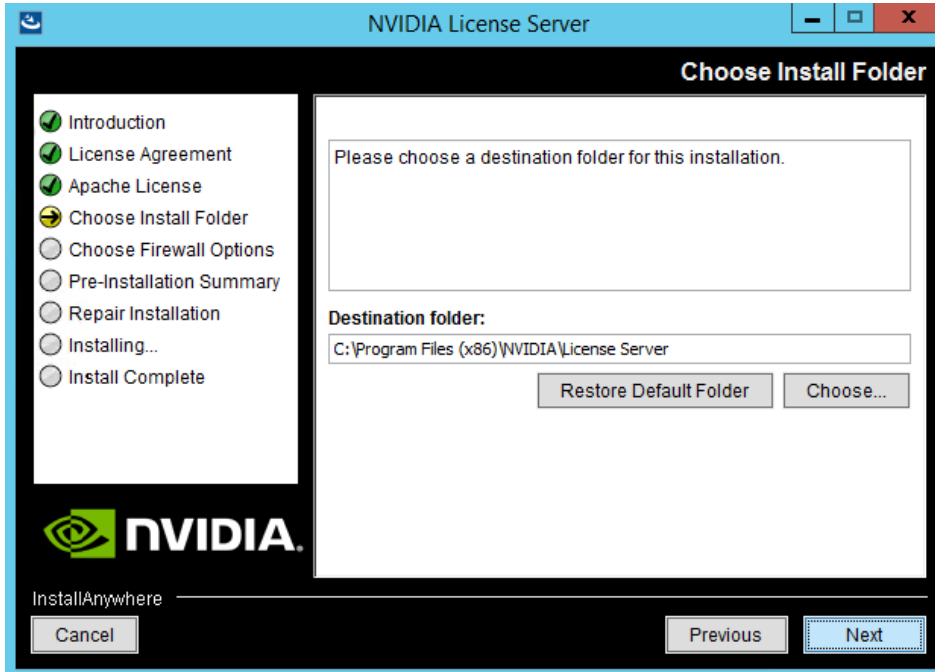
4. Accept the Apache license agreement and click Next (Figure 22).

**Figure 22.** Accepting the Apache license agreement

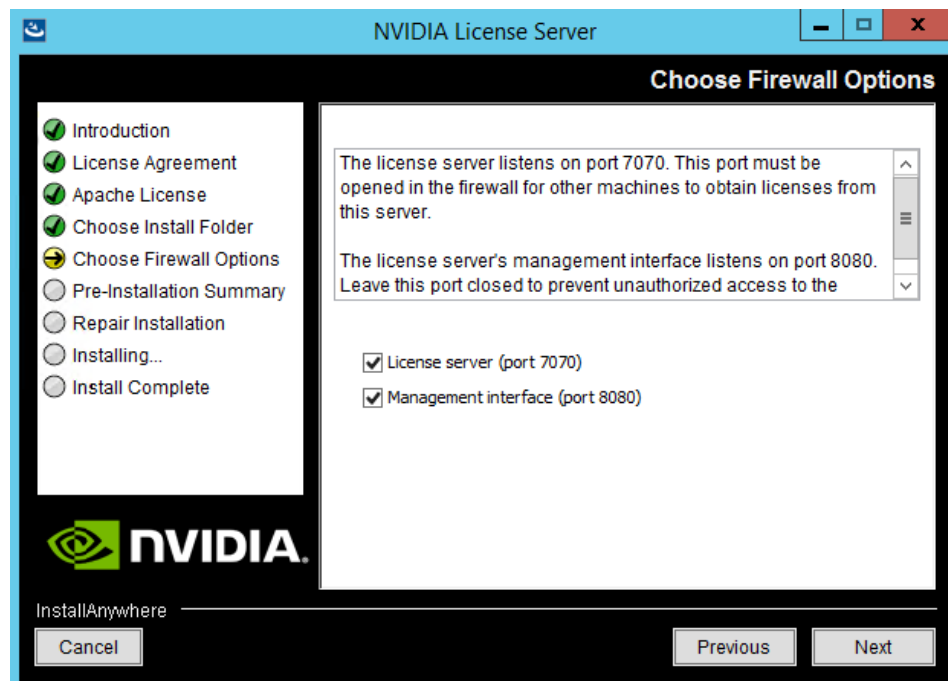


5. Choose the desired installation folder and click Next (Figure 23).

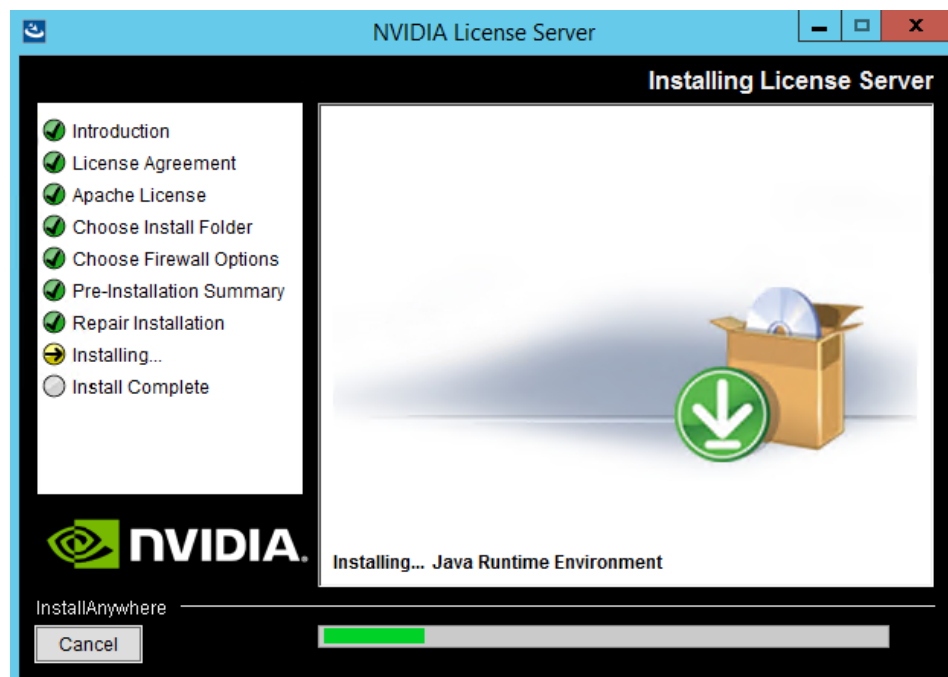
**Figure 23.** Choosing a destination folder



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 up port 8080. Select the "Management interface (port 8080)" option.
8. Click Next (Figure 24).

**Figure 24.** Setting firewall options

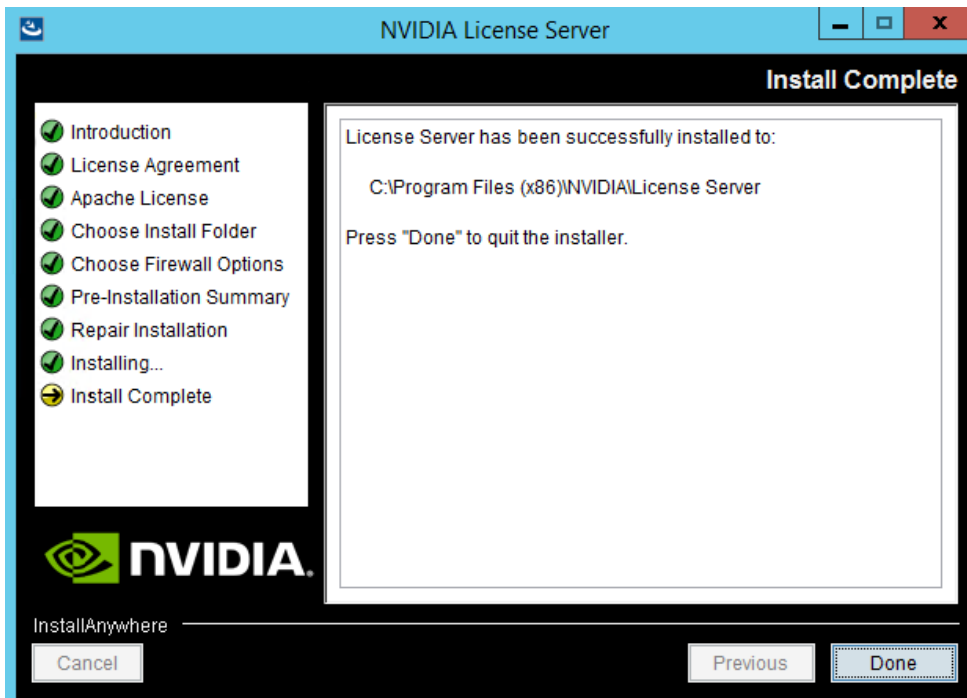
9. The Pre-installation Summary and Repair Installation options automatically progress without user input (Figure 25).

**Figure 25.** Installing the license server



10. When the installation process is complete, click Done (Figure 26).

**Figure 26.** Installation complete

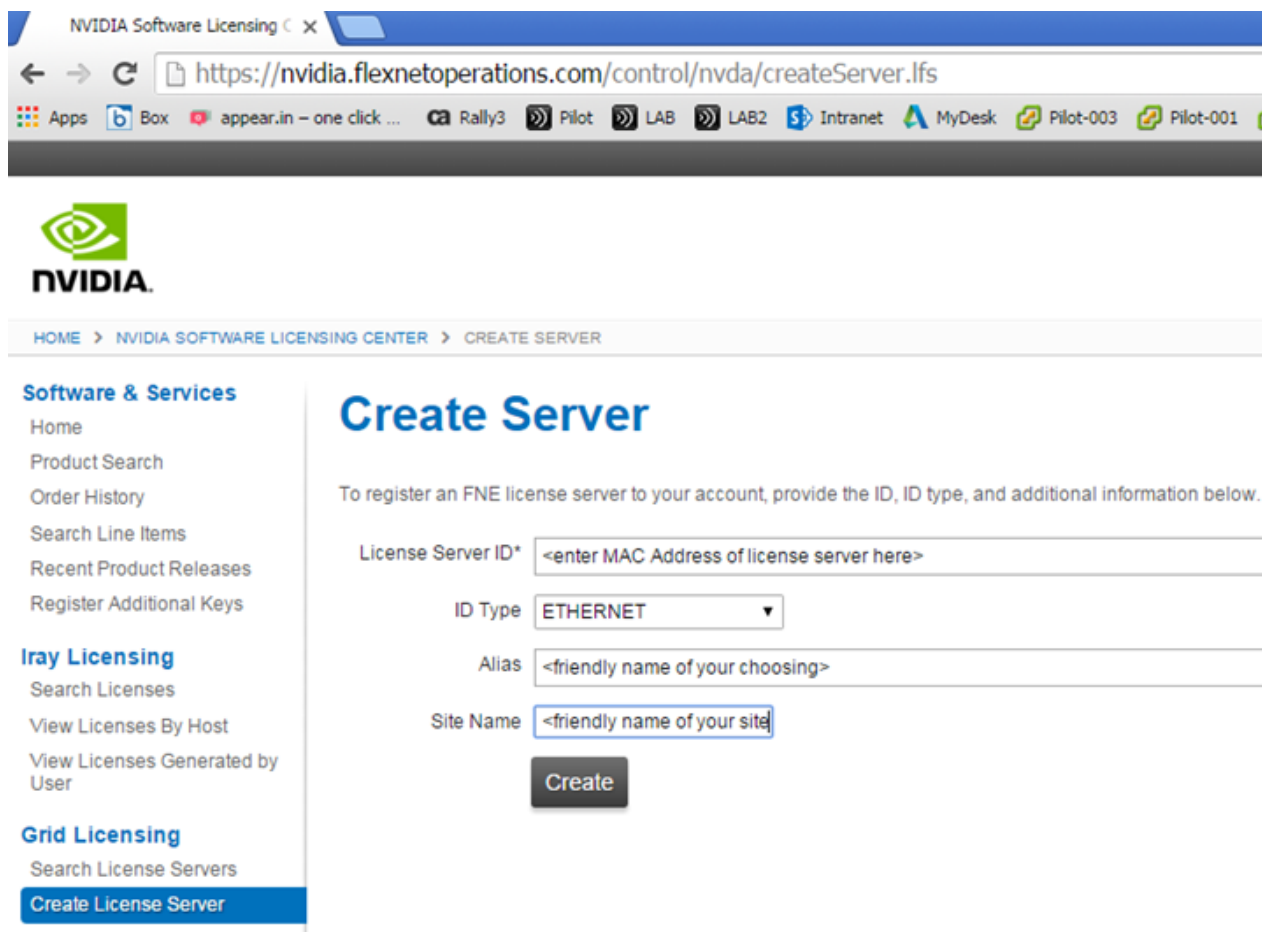


## Configure the NVIDIA GRID 5.0 license server

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](https://nvidia.com/grideval). A license file is generated from <https://nvidia.flexnetoperations.com/>.
2. After you are logged in, click Create License Server.
3. Specify the fields as shown in Figure 27. In the License Server ID field, enter the MAC address of your local license server's NIC. Leave ID Type set to Ethernet. For Alias and Site Name, choose user-friendly names. Then click Create.


**Figure 27.** Creating the license server



NVIDIA Software Licensing C X

← → ↻ <https://nvidia.flexnetoperations.com/control/nvda/createServer.lfs>

Apps Box appear.in – one click ... Rally3 Pilot LAB LAB2 Intranet MyDesk Pilot-003 Pilot-001

 NVIDIA

HOME > NVIDIA SOFTWARE LICENSING CENTER > CREATE SERVER

**Software & Services**

- Home
- Product Search
- Order History
- Search Line Items
- Recent Product Releases
- Register Additional Keys

**Iray Licensing**

- Search Licenses
- View Licenses By Host
- View Licenses Generated by User

**Grid Licensing**

- Search License Servers
- Create License Server**

## Create Server

To register an FNE license server to your account, provide the ID, ID type, and additional information below.

License Server ID\*


ID Type

Alias

Site Name

4. Click the Search License Servers node.
5. Click your license server ID (Figure 28).

**Figure 28.** Selecting the license server ID



HOME > NVIDIA SOFTWARE LICENSING CENTER > SEARCH SERVERS

## Search Servers

License Server ID  Alias

ID Type  Site Name

Activation Code

**Filter**

1 to 1 of 1 Entries per page: 25

License Server ID	ID Type
	ETHERNET

**Software & Services**

- Home
- Product Search
- Order History
- Search Line Items
- Recent Product Releases
- Register Additional Keys

**Iray Licensing**

- Search Licenses
- View Licenses By Host
- View Licenses Generated by User

**Grid Licensing**

- Search License Servers**
- Create License Server

6. Click Map Add-Ons and choose the number of license units out of your total pool to allocate to this license server (Figure 29).

**Figure 29.** Choosing the number of license units from the pool

## View Server

License Server ID

ID Type ETHERNET

Alias  **Update Alias**

Site Name

**Map Add-Ons** | [Remove Add-Ons](#) | [View History](#) | [View Served Clients](#) | [Download License File](#)

### Add-Ons

Add-On Name	Status
GRID Evaluation Edition	License generated

After the add-ons are mapped, the interface will look like Figure 30, showing 128 units mapped, for example.

**[[I DON'T SEE THIS IN THE FOLLOWING FIGURE. PLS VERIFY. THIS FIGURE LOOKS IDENTICAL TO FIG 29.]]**

**Figure 30.** The interface after the add-ons are mapped: here showing 128 units mapped

## View Server

License Server ID

ID Type ETHERNET

Alias

Site Name

[Map Add-Ons](#) | 
 [Remove Add-Ons](#) | 
 [View History](#) | 
 [View Served Clients](#) | 
 [Download License File](#)

### Add-Ons

Add-On Name	Status
GRID Evaluation Edition	License generated

- Click Download License File and save the .bin file to your license server (Figure 31).

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

**Figure 31.** Saving the .bin File

HOME > NVIDIA SOFTWARE LICENSING CENTER > VIEW SERVER

**Software & Services**

- Home
- Product Search
- Order History
- Search Line Items
- Recent Product Releases
- Register Additional Keys

**Iray Licensing**

- Search Licenses
- View Licenses By Host
- View Licenses Generated by User

**Grid Licensing**

## View Server

License Server ID

ID Type ETHERNET

Alias

Site Name

[Map Add-Ons](#) | 
 [Remove Add-Ons](#) | 
 [View History](#) | 
 [View Served Clients](#) | 
 [Download License File](#)

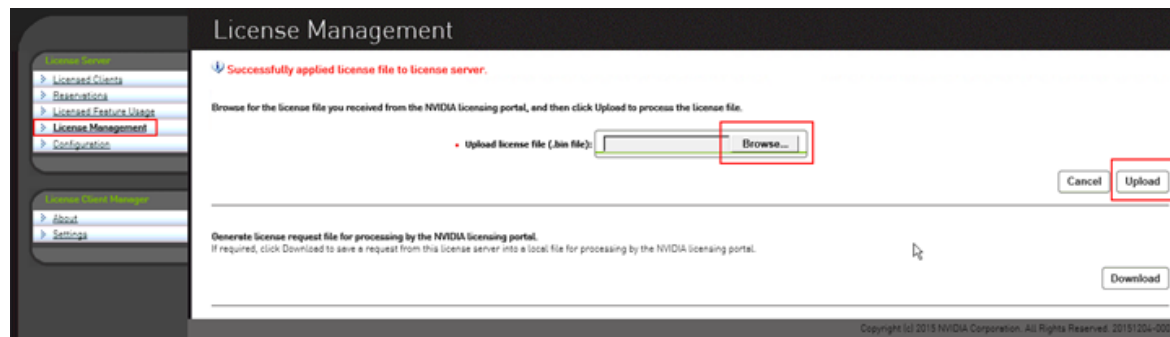
### Add-Ons

Add-On Name	Status
No add-ons are currently mapped.	

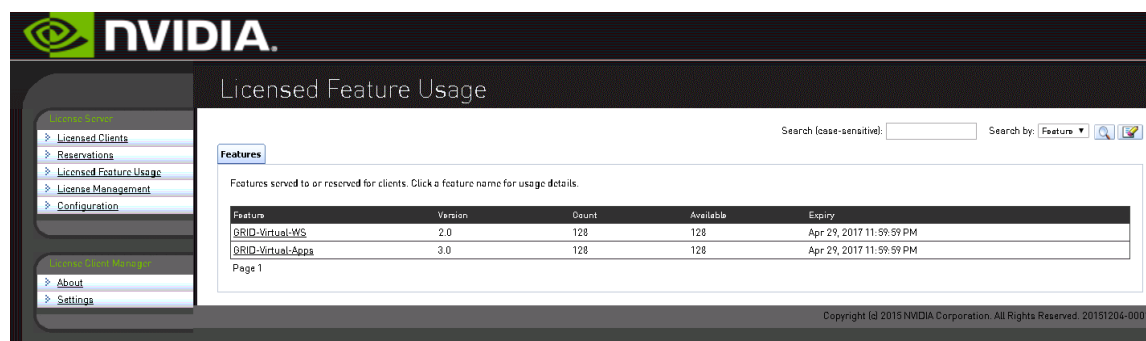
- On the local license server, browse to `http://<FQDN>:8080/licserver` to display the License Server Configuration page.
- Click License Management in the left pane.
- Click Browse to locate your recently download .bin license file. Select the .bin file and click OK.

11. Click Upload. The message “Successfully applied license file to license server” should appear on the screen (Figure 32). The features are available (Figure 33).

**Figure 32.** License file successfully applied



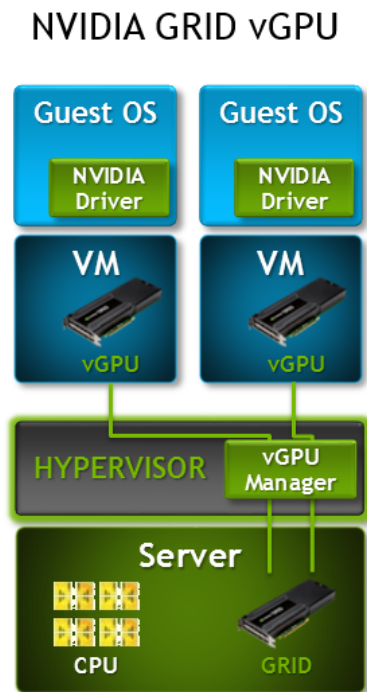
**Figure 33.** NVIDIA license server



## Install NVIDIA GRID software on the VMware ESX host and Microsoft Windows virtual machine

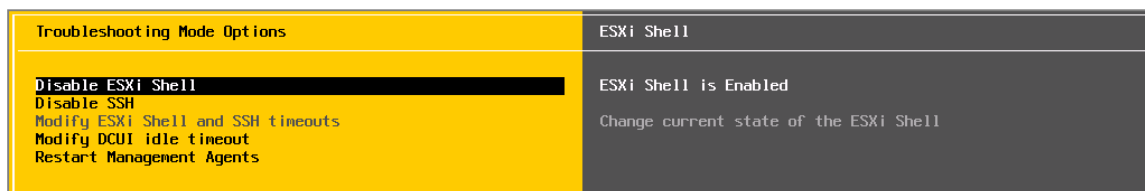
This section summarizes the installation process for configuring an ESXi host and virtual machine for vGPU support. Figure 34 shows the components used for vGPU support.

**Figure 34.** NVIDIA GRID vGPU cComponents



1. Download the NVIDIA GRID GPU driver pack for VMware vSphere ESXi 6.5.
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 35).

**Figure 35.** VMware ESXi configuration console



3. Upload the NVIDIA driver (vSphere Installation Bundle [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.

- The ESXi host must be in maintenance mode for you to install the VIB module. To place the host in maintenance mode, use the following command:

```
esxcli system maintenanceMode set -enable true
```

- 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-vGPU-
VMware_ESXi_6.5_Host_Driver_384.73-1OEM.650.0.0.4598673.x86_64.vib --no-sig-check
Installation Result
    Message: Operation finished successfully.
    Reboot Required: false
    VIBs Installed: NVIDIA_bootbank_NVIDIA-VMware_ESXi_6.5_Host_Driver_384.73-
1OEM.650.0.0.4598673
    VIBs Removed:
    VIBs Skipped:
```

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

- 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
```

- 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_6.5_Host_Driver  384.73-1OEM.650.0.0.4598673      NVIDIA
VMwareAccepted      2017-07-31
```

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](http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2033434).



- 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**

The command should return output similar to that shown in Figures 36, 37, or 38, depending on the card used in your environment.

**Figure 36.** VMware ESX SSH console report for GPU P40 card detection on Cisco UCS C240 M5 Rack Server

```

-sh: nvidia-smi: not found
[root@M5:~] nvidia-smi
Wed Sep  6 00:43:04 2017

+-----+
| NVIDIA-SMI 384.73                 Driver Version: 384.73 |
+-----+

+-----+
| GPU   Name           Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf    Pwr:Usage/Cap|      Memory-Usage | GPU-Util  Compute M. |
+-----+-----+
|  0   Tesla P40      On          | 0000:5E:00.0  Off  |    0%      0         |
| N/A   28C    P8      19W / 250W | 45MiB / 23039MiB |           Default     |
+-----+-----+
|  1   Tesla P40      On          | 0000:AF:00.0  Off  |    0%      0         |
| N/A   22C    P8      18W / 250W | 45MiB / 23039MiB |           Default     |
+-----+-----+

+-----+
| Processes:                                GPU Memory |
|  GPU       PID    Type    Process name      Usage    |
+-----+-----+
| No running processes found |
+-----+

[root@C240-M5:~]

```

**Figure 37.** VMware ESX SSH console report for GPU M10 card detection on Cisco UCS C240 M5 Rack Server

```

-sh: nvidia-smi: not found
[root@M5:~] nvidia-smi
Wed Sep  6 00:43:04 2017
+-----+
| NVIDIA-SMI 384.73                Driver Version: 384.73          |
+-----+
+-----+
| GPU  Name          Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf    Pwr:Usage/Cap|      Memory-Usage | GPU-Util  Compute M. |
+-----+-----+
|  0   Tesla M10      On          | 0000:60:00.0   Off  |           N/A       |
| N/A   29C    P8      10W /  53W |  18MiB / 8191MiB |      0%      Default |
+-----+-----+
|  1   Tesla M10      On          | 0000:61:00.0   Off  |           N/A       |
| N/A   30C    P8      10W /  53W |  18MiB / 8191MiB |      0%      Default |
+-----+-----+
|  2   Tesla M10      On          | 0000:62:00.0   Off  |           N/A       |
| N/A   26C    P8      10W /  53W |  18MiB / 8191MiB |      0%      Default |
+-----+-----+
|  3   Tesla M10      On          | 0000:63:00.0   Off  |           N/A       |
| N/A   26C    P8      10W /  53W |  18MiB / 8191MiB |      0%      Default |
+-----+-----+
|  4   Tesla M10      On          | 0000:88:00.0   Off  |           N/A       |
| N/A   27C    P8      10W /  53W |  18MiB / 8191MiB |      0%      Default |
+-----+-----+
|  5   Tesla M10      On          | 0000:89:00.0   Off  |           N/A       |
| N/A   28C    P8      10W /  53W |  18MiB / 8191MiB |      0%      Default |
+-----+-----+
|  6   Tesla M10      On          | 0000:8A:00.0   Off  |           N/A       |
| N/A   25C    P8      10W /  53W |  18MiB / 8191MiB |      0%      Default |
+-----+-----+
|  7   Tesla M10      On          | 0000:8B:00.0   Off  |           N/A       |
| N/A   24C    P8      10W /  53W |  18MiB / 8191MiB |      0%      Default |
+-----+-----+
+-----+
| Processes:                                     GPU Memory |
|  GPU       PID    Type    Process name      Usage    |
+-----+-----+
| No running processes found                     |
+-----+

```

**Figure 38.** VMware ESX SSH console report for GPU P6 card detection on Cisco UCS B200 M5 Blade Server

```
[root@M5:~] nvidia-smi
Wed Sep  6 00:43:04 2017

+-----+
| NVIDIA-SMI 384.73                  Driver Version: 384.73          |
+-----+-----+
| GPU  Name            Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf    Pwr:Usage/Cap|      Memory-Usage | GPU-Util  Compute M. |
+-----+-----+
|  0   Tesla P6             On      | 00000000:18:00.0 Off  |          Off          |
| N/A   21C    P8             9W /  90W |  41MiB / 16383MiB |      0%      Default  |
+-----+-----+
|  1   Tesla P6             On      | 00000000:D8:00.0 Off  |          Off          |
| N/A   35C    P8            10W /  90W |  41MiB / 16383MiB |      0%      Default  |
+-----+-----+

+-----+
| Processes:                                     GPU Memory |
|  GPU       PID    Type    Process name                        Usage      |
+-----+-----+
| No running processes found                     |
+-----+

[root@M5:~] █
```

**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 Tesla P6, P40, and M10 profile specifications

The Tesla P6 and P40 card has a single physical GPU, and the Tesla M10 card has multiple physical GPUs. Each physical GPU can support several different types of virtual GPU. 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 6 lists the vGPU types supported by GRID GPUs.

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

**Table 6.** End-user profile specifications for NVIDIA Tesla cards

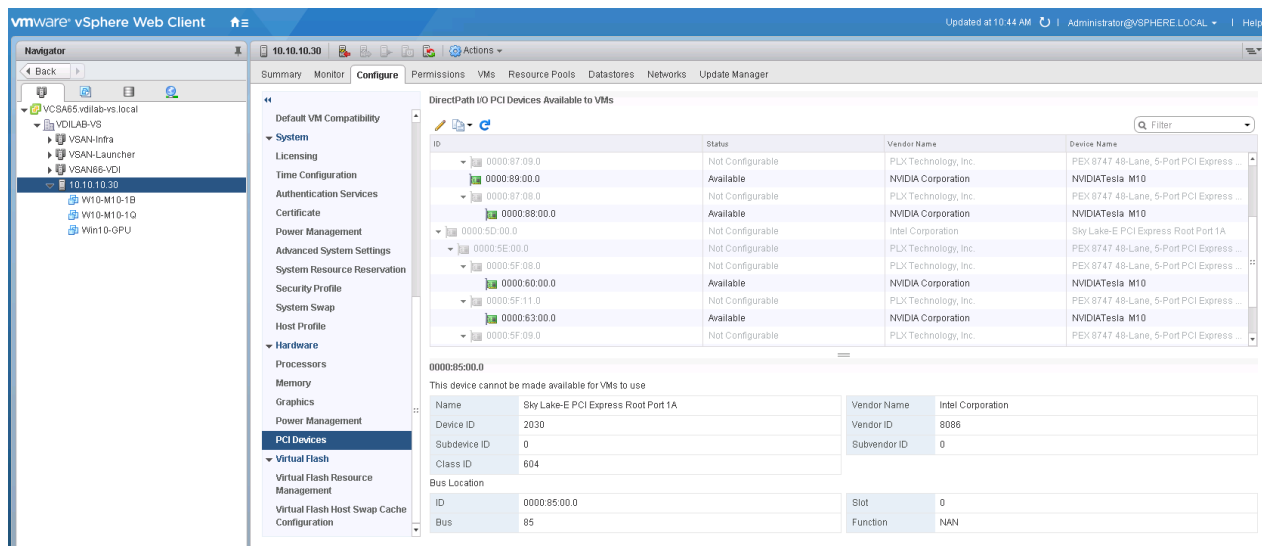
End-user and NVIDIA GRID options			
End-user profile	GRID Virtual Application profiles	GRID Virtual PC profiles	GRID Virtual Workstation profiles
<b>1 GB</b>	P6: 1A	P6: 1B	P6: 1Q
	M10: 1A	M10: 1B	M10: 1Q
	P40: 1A	P40: 1B	P40: 1Q
<b>2 GB</b>	P6: 2A	—	P6: 2Q
	M10: 2A		M10: 2Q
	P40: 2A		P40: 2Q
<b>3 GB</b>	P40: 3A	—	P40: 3Q
<b>4 GB</b>	P6: 4A	—	P6: 4Q
	M10: 4A		M10: 4Q
	P40: 4A		P40: 4Q
<b>6 GB</b>	P40: 6A	—	P40: 6Q
<b>8 GB</b>	P6: 8A	—	P6: 8Q
	M10: 8A		M10: 8Q
	P40: 8A		P40: 8Q
<b>12 GB</b>	P40: 12A	—	P40: 12Q
<b>16 GB</b>	P6: 16A	—	P6: 16Q
<b>24 GB</b>	P40: 24A	—	P40: 24Q
<b>Pass-through</b>	—	—	—

## Prepare a virtual machine for NVIDIA GRID vGPU support

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

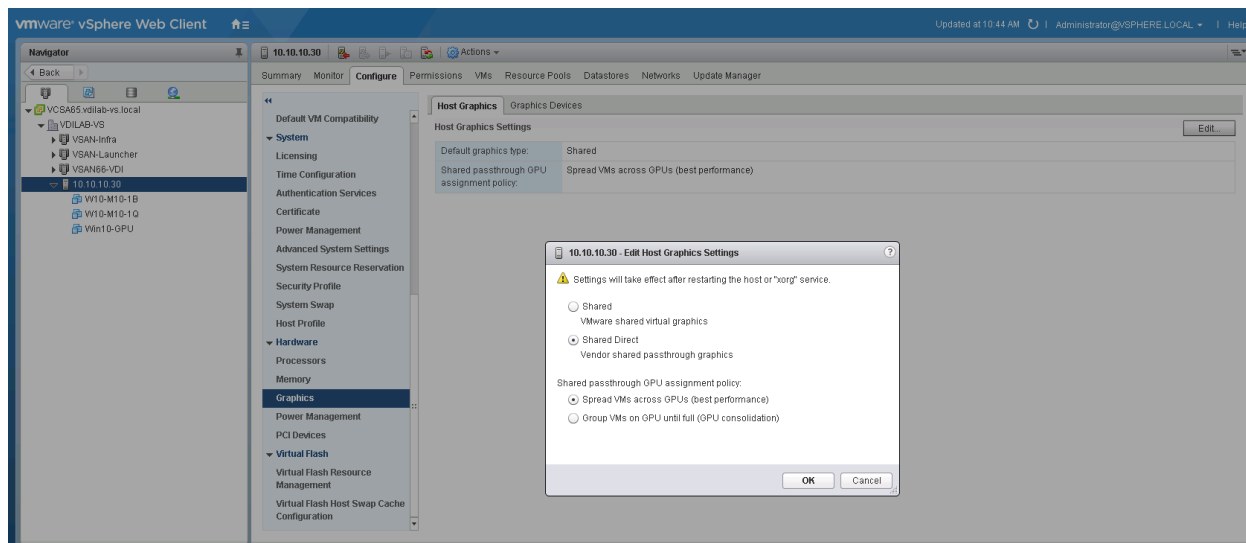
1. In vSphere Web Client, go to the host settings and select PCI Devices. Make GPUs available for pass-through (Figure 39).

**Figure 39.** Make GPUs available for pass-through



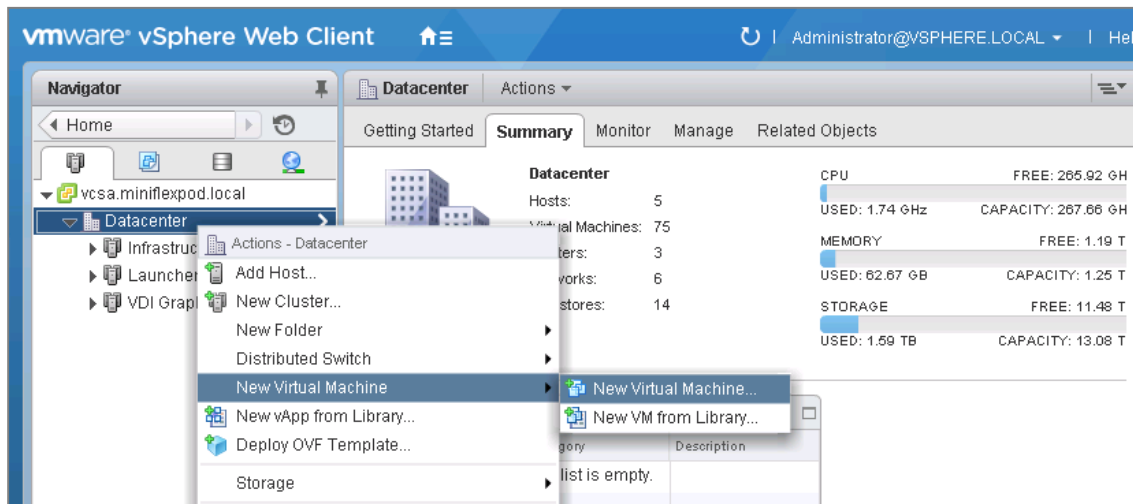
2. Select the ESXi host and click the Configure tab.
3. From the list of options at the left, select Graphics. Click Edit Host Graphics Settings. **[[WHERE DO WE DO THIS?]]** Then select “Shared Direct: Vendor shared passthrough graphics” (Figure 40).

**Figure 40.** Editing the host graphics settings



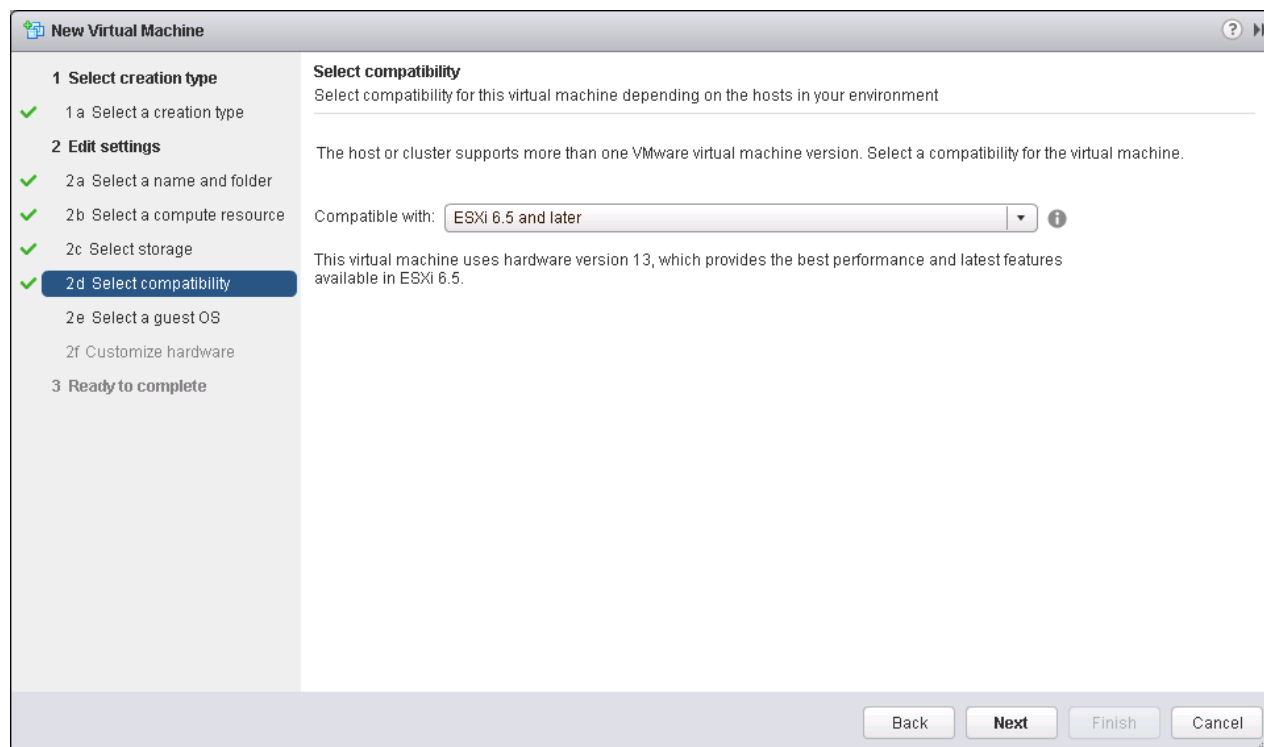
4. Reboot to make the changes take effect.
5. Using the vSphere Web Client, create a new virtual machine. To do this, right-click a host or cluster and choose New Virtual Machine. Work through the New Virtual Machine wizard. Unless another configuration is specified, select the configuration settings appropriate for your environment (Figure 41).

**Figure 41.** Creating a new virtual machine in VMware vSphere Web Client



6. From the “Compatible with” drop-down menu, choose “ESXi 6.0 and later” to use the latest features, including the mapping of shared PCI devices, which is required for the vGPU feature (Figure 42). The study in this document uses “ESXi 6.5 and later,” which provides the latest features available in ESXi 6.5, and virtual machine hardware Version 13.

**Figure 42.** Selecting Virtual Machine Hardware Version 11 or higher



**New Virtual Machine**

**1 Select creation type**

- ✓ 1 a Select a creation type

**2 Edit settings**

- ✓ 2 a Select a name and folder
- ✓ 2 b Select a compute resource
- ✓ 2 c Select storage
- ✓ 2 d Select compatibility**
- 2 e Select a guest OS
- 2 f Customize hardware

**3 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 6.5 and later**

This virtual machine uses hardware version 13, which provides the best performance and latest features available in ESXi 6.5.

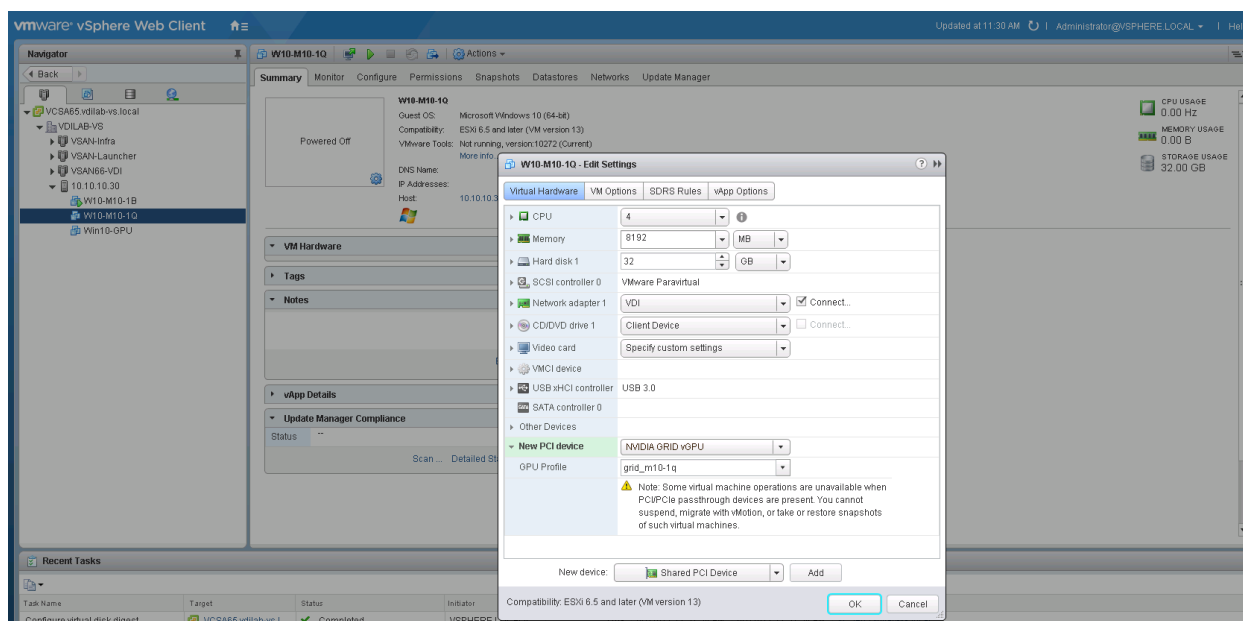
Back Next Finish Cancel



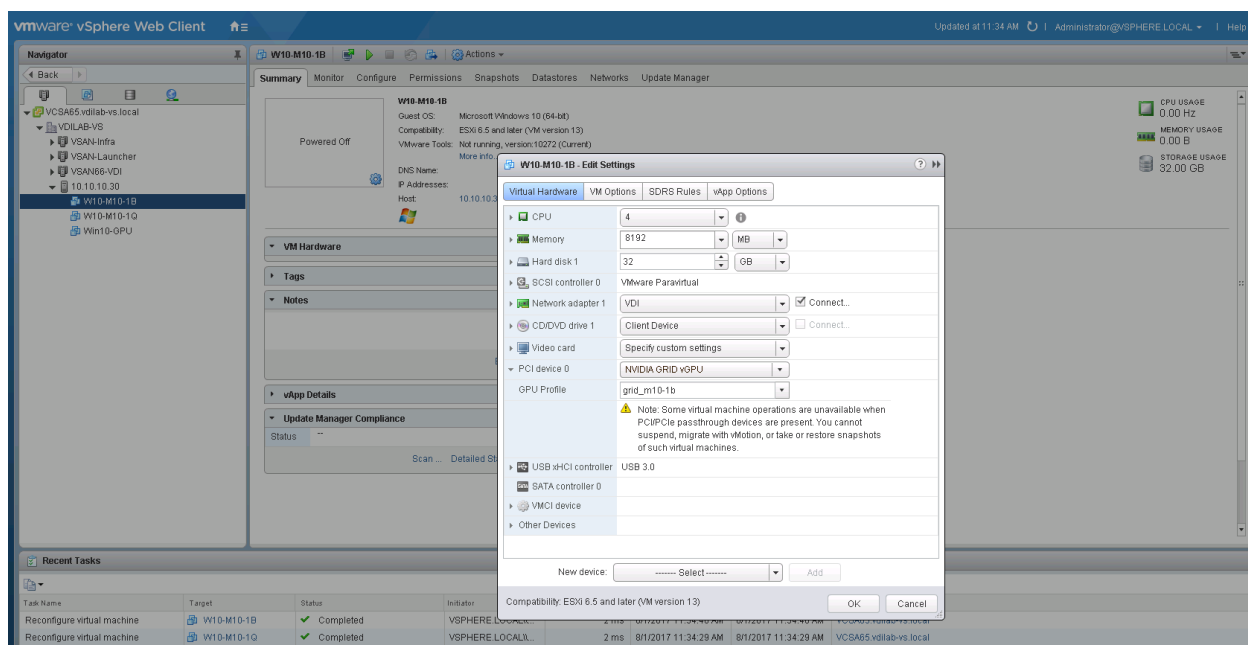
- In customizing the hardware of the new virtual machine, add a new shared PCI device, select the appropriate GPU profile, and reserve all virtual machine memory (Figures 43 and 44).

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 43.** Adding a shared PCI device to the virtual machine to attach the GPU profile

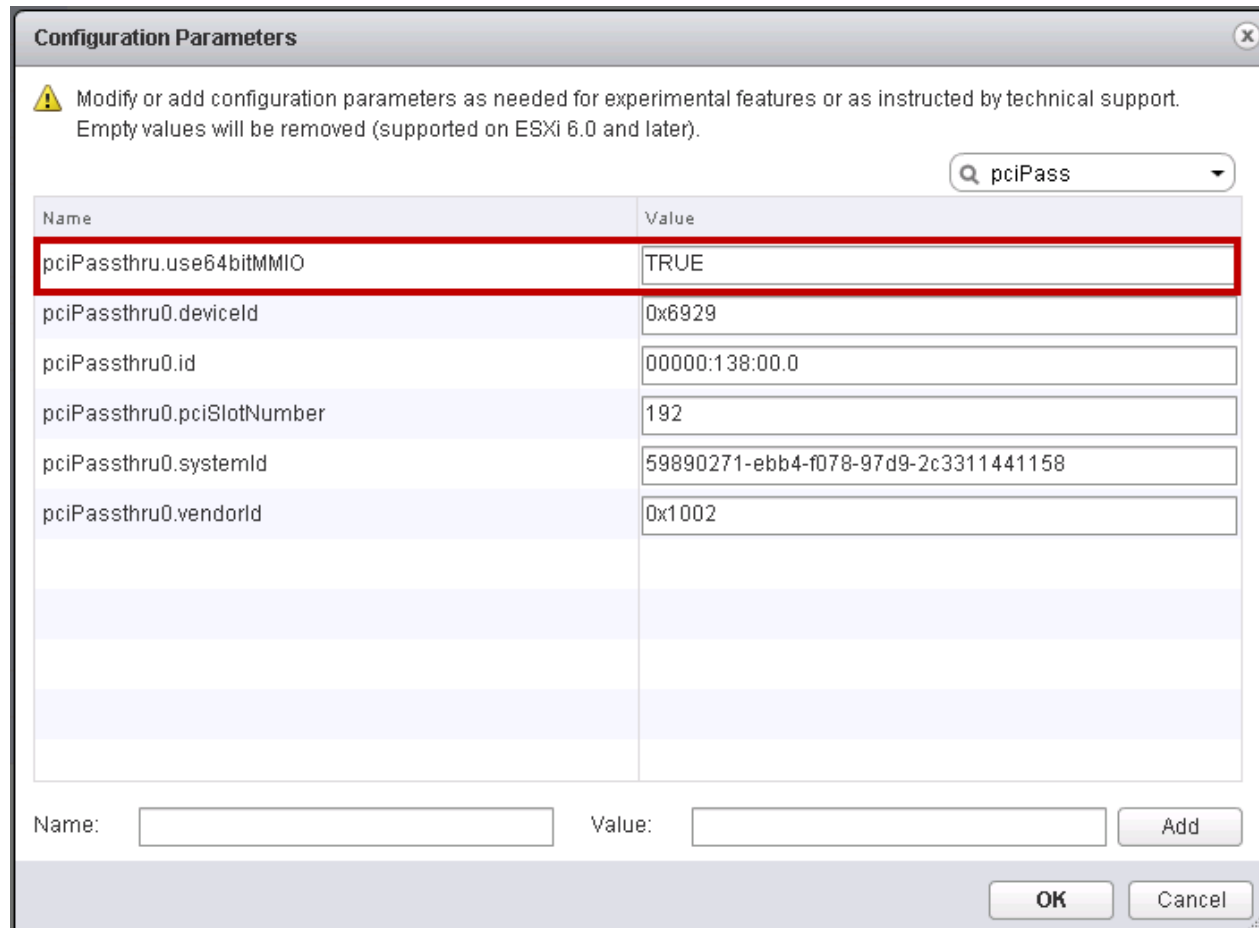


**Figure 44.** Attaching the GPU profile to a shared PCI device



8. Select the VM Options tab. From the drop-down menu under Advanced, choose Configuration Parameters. Click Edit Configuration. Specify a name such as `pciPassthru.use64bitMMIO` and set the value to `TRUE` (Figure 45).

**Figure 45.** Adding `pciPassthru.use64bitMMIO = TRUE`



**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).

pciPass

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

Name:  Value:  Add

OK Cancel

9. Note that a virtual machine with vGPU assigned will fail to start if ECC is enabled. As a workaround, disable ECC by entering the following commands (Figure 46):

```
#nvidia-smi -i 0 -e 0
#nvidia-smi -i 1 -e 0
```

Use -i to target a specific GPU. If two cards are installed in a server, run the command twice as shown here, where 0 and 1 identify the two GPU cards.

**Figure 46.** Disabling ECC

```
-sh: nvidia-smi: not found
[root@M5:~] nvidia-smi
Wed Sep  6 00:43:04 2017

+-----+
| NVIDIA-SMI 384.73                 Driver Version: 384.73 |
+-----+

+-----+
| GPU  Name          Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf    Pwr:Usage/Cap|     Memory-Usage | GPU-Util  Compute M. |
+-----+-----+
|  0   Tesla P6             On     | 0000:18:00.0  Off  |           0         |
| N/A   22C    P8      9W /  90W |  39MiB / 15359MiB|      0%    Default   |
+-----+-----+
|  1   Tesla P6             On     | 0000:D8:00.0  Off  |           0         |
| N/A   37C    P8     10W /  90W |  39MiB / 15359MiB|      0%    Default   |
+-----+-----+

+-----+
| Processes:                      GPU Memory |
| GPU       PID    Type    Process name                        Usage |
+-----+-----+
| No running processes found |
+-----+

[root@M5:~] esxtop -a -b -d 10 -n 600 > /vmfs/volumes/594d8376-1531284a-003b-0025b5000a2f/215U-003.csv
[root@M5:~] nvidia-smi -i 0 -e 0
-sh: nvidia-smi: not found
[root@M5:~] nvidia-smi -i 0 -e 0
Disabled ECC support for GPU 0000:18:00.0.
All done.
Reboot required.
[root@M5:~] nvidia-smi -i 1 -e 0
Disabled ECC support for GPU 0000:D8:00.0.
All done.
Reboot required.
[root@M5:~]
```

10. Install and configure Microsoft Windows on the virtual machine:
- Configure the virtual machine with the appropriate amount of vCPU and RAM according to the GPU profile selected.
  - Install VMware Tools.
  - Join the virtual machine to the Microsoft Active Directory domain.
  - Install or upgrade Citrix HDX 3D Pro Virtual Desktop Agent using the CLI.
    - When you use the installer's GUI to install a VDA for a Windows desktop, simply select Yes on the HDX 3D Pro page. When you use the CLI, include the /enable\_hdx\_3d\_pro option with the XenDesktop VdaSetup.exe command.
    - To upgrade HDX 3D Pro, uninstall both the separate HDX 3D for Professional Graphics component and the VDA before installing the VDA for HDX 3D Pro. Similarly, to switch from the standard VDA for a Windows desktop to the HDX 3D Pro VDA, uninstall the standard VDA and then install the VDA for HDX 3D Pro.

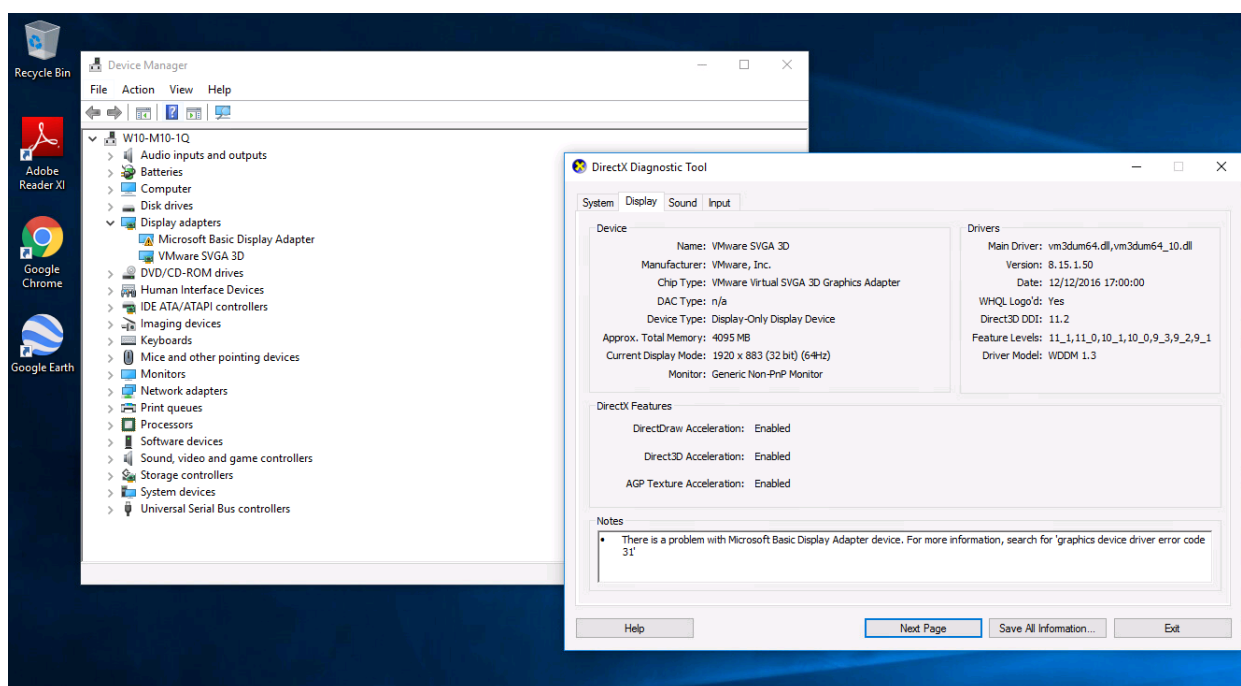
- e. Optimize the Windows OS. [VMware OSOT](#), the optimization tool, includes customizable templates to enable or disable Windows system services and features using LoginVSI recommendations and best practices across multiple systems. Because most Windows system services are enabled by default, the optimization tool can be used to easily disable unnecessary services and features to improve performance.

### Install the NVIDIA GRID 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.

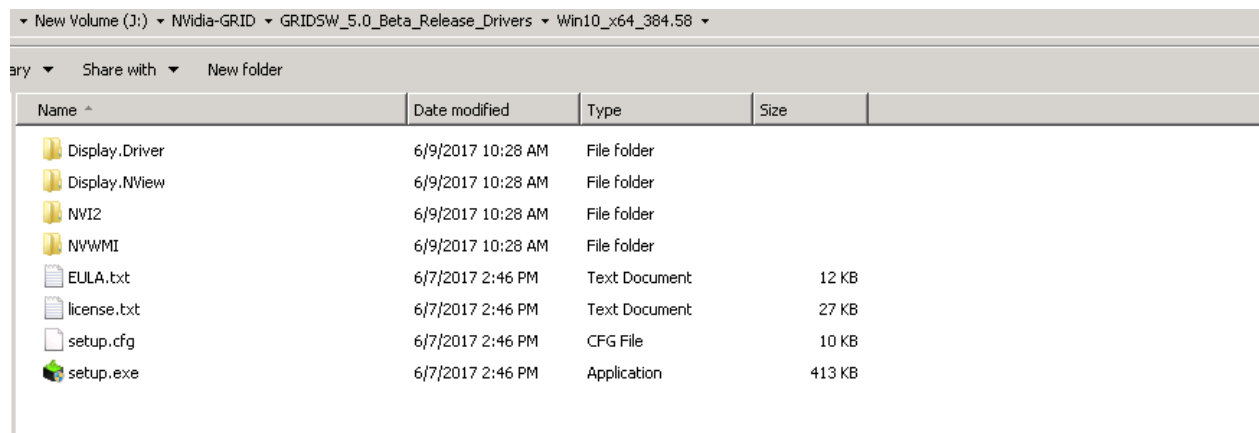
Before the NVIDIA driver is installed on the guest virtual machine, the Device Manager shows the standard VGA graphics adapter (Figure 47).

**Figure 47.** 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 48).

**Figure 48.** NVIDIA driver pack



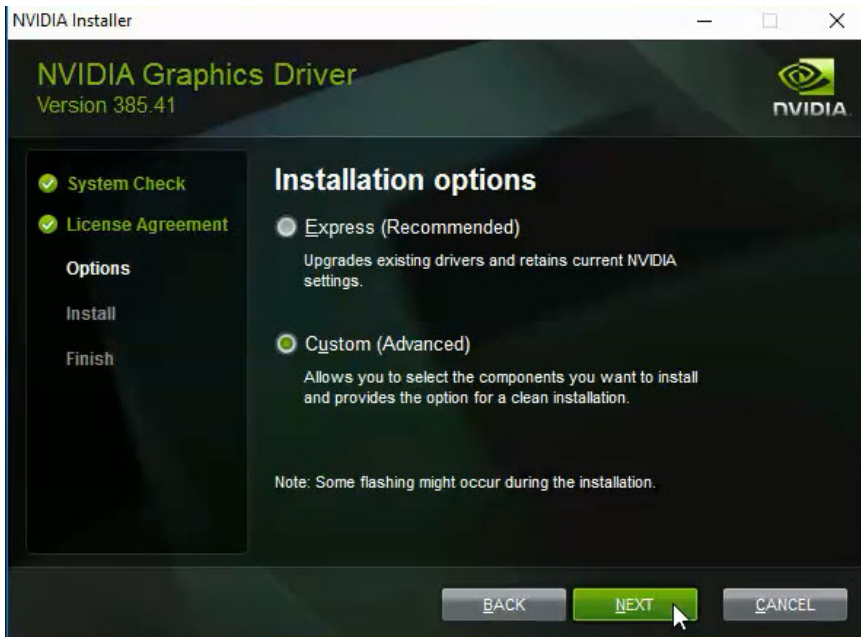
Name ^	Date modified	Type	Size
Display.Driver	6/9/2017 10:28 AM	File folder	
Display.NView	6/9/2017 10:28 AM	File folder	
NVI2	6/9/2017 10:28 AM	File folder	
NVWMI	6/9/2017 10:28 AM	File folder	
EULA.txt	6/7/2017 2:46 PM	Text Document	12 KB
license.txt	6/7/2017 2:46 PM	Text Document	27 KB
setup.cfg	6/7/2017 2:46 PM	CFG File	10 KB
setup.exe	6/7/2017 2:46 PM	Application	413 KB

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 pass-through driver.

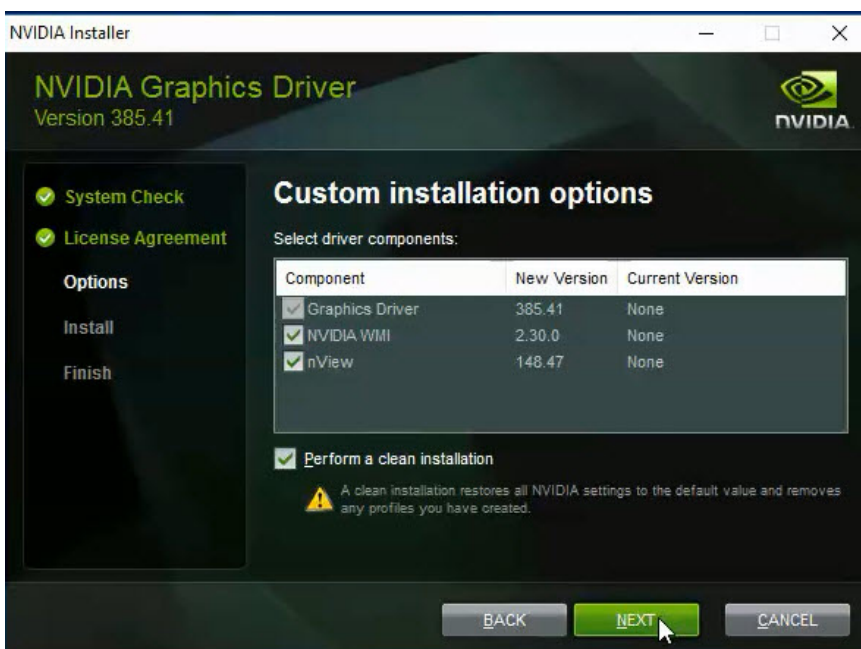
3. Install the graphics drivers using the Express/Custom option (Figure 49). After the installation has been completed successfully (Figure 50), restart the virtual machine.

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 49.** Select Express/Custom installation option



**Figure 50.** Components installed during NVidia Graphics Driver installation process.

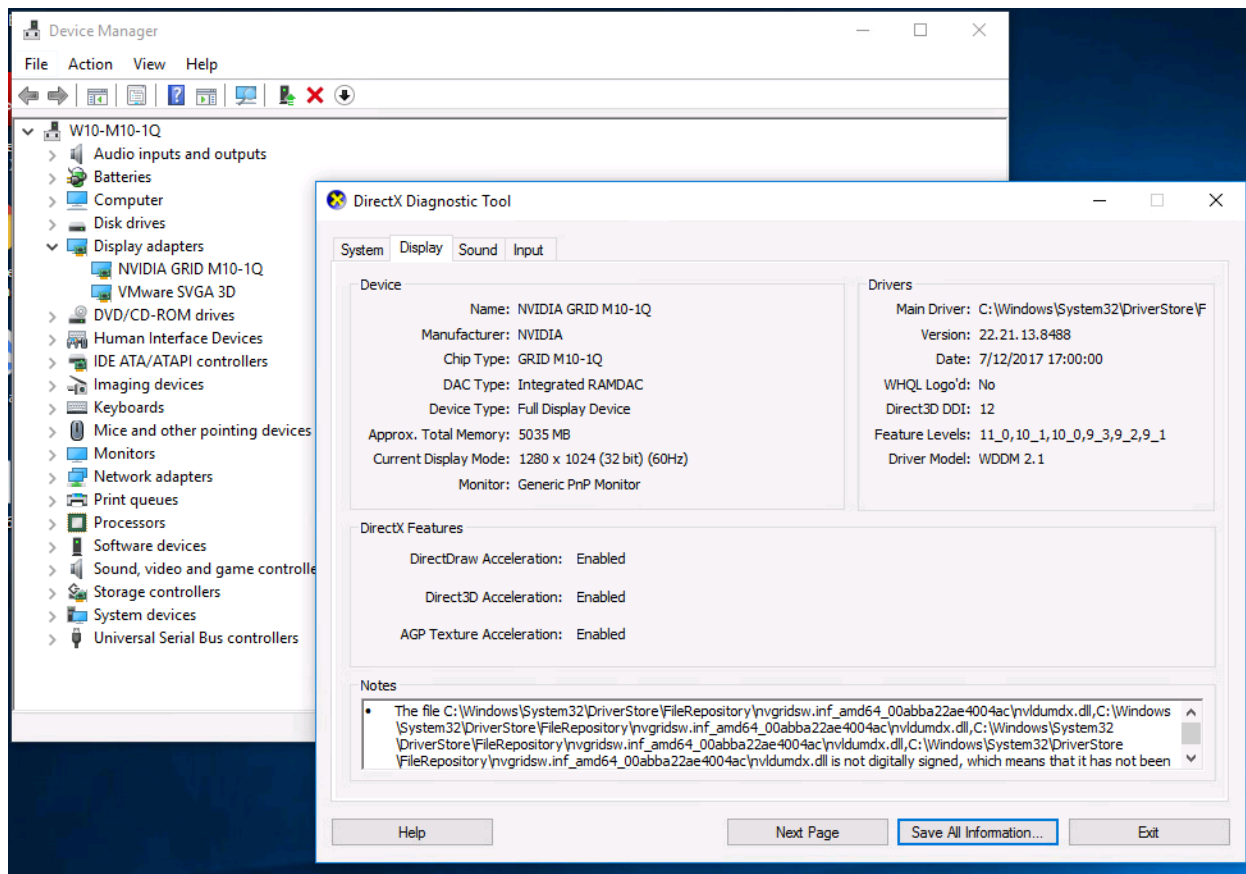


## Verify that applications are ready to support NVIDIA GRID vGPU

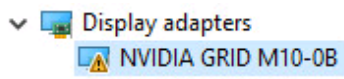
Validate the successful installation of the graphics drivers and the vGPU device.

Open Windows Device Manager and expand the Display Adapter section. The device will reflect the chosen profile (Figure 51).

**Figure 51.** Validating the driver installation



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



The following are the most likely the reasons:

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

## Configure the virtual machine for an NVIDIA GRID vGPU license

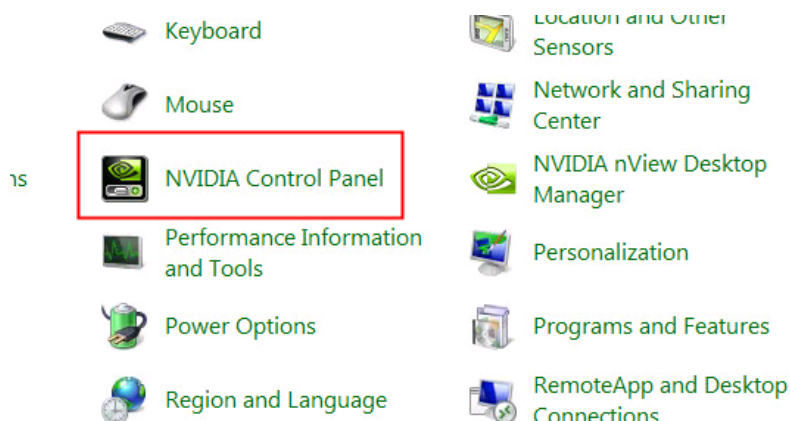
You need to point the master image to the license server so the virtual machines with vGPUs can obtain a license.

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



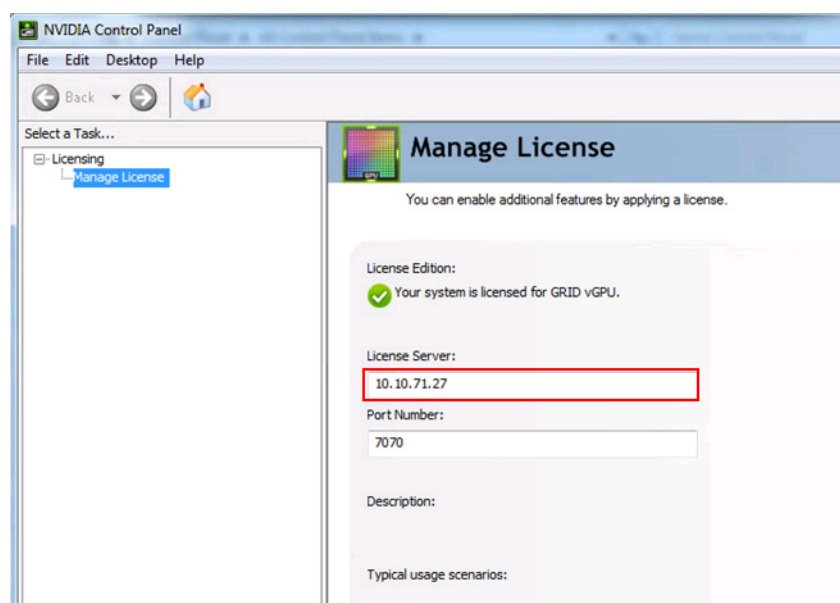
1. In the Microsoft Windows Control Panel, double-click NVIDIA Control Panel (Figure 52).

**Figure 52.** Choosing NVIDIA Control Panel



2. Select Manage License from the left pane and enter your license server address and port (Figure 53).

**Figure 53.** Managing your license



3. Click Apply.

## Verify NVIDIA GRID vGPU deployment

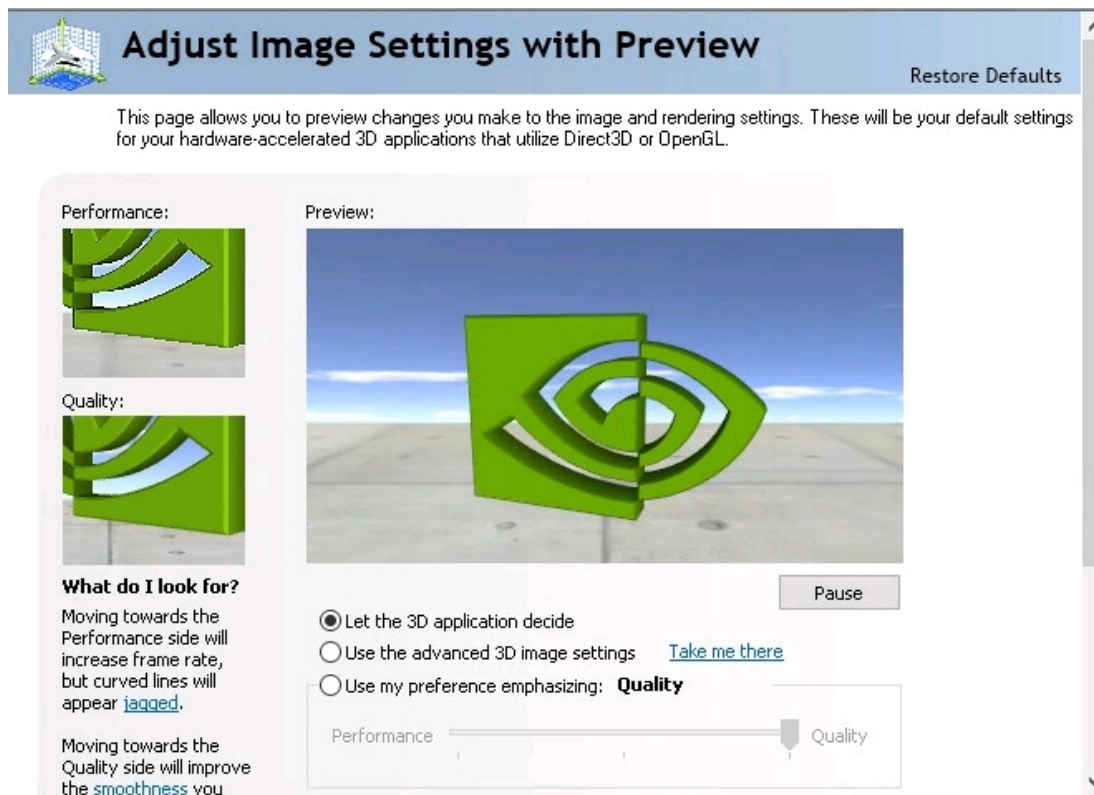
After the desktops are provisioned, use the following steps to verify vGPU deployment in the Citrix XenDesktop environment.

### 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 54).


**Figure 54.** 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 55).

**Figure 55.** NVIDIA license server: Licensed Feature Usage



The screenshot shows the NVIDIA License Server web interface. The left sidebar contains navigation links: License Server, Licensed Clients, Reservations, Licensed Feature Usage (selected), License Management, Configuration, and Login. Below these are links for License Client Manager, About, and Settings. The main content area is titled "Licensed Feature Usage" and contains a table with the following data:

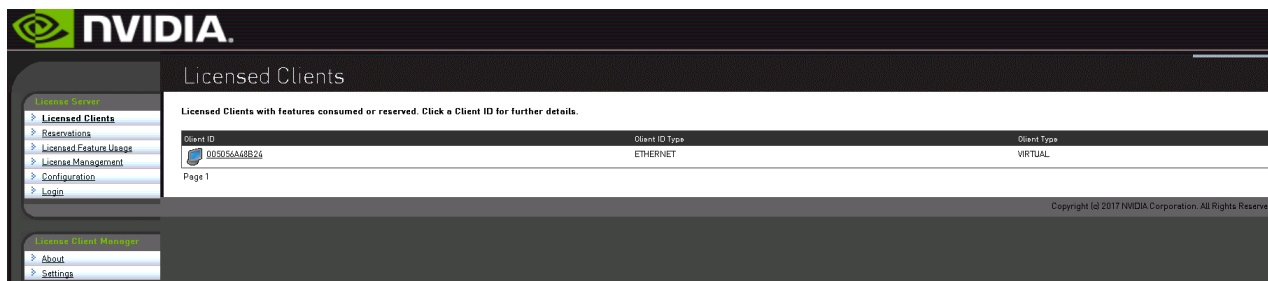
Feature	Version	Count	Available	Expiry
GRID-Virtual-Apps	3.0	128	128	2017-10-04
GRID-Virtual-WS	2.0	128	127	2017-10-04

Page 1

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

To view the details, select Licensed Clients in the left pane (Figure 56).

**Figure 56.** NVIDIA license server: Licensed Clients



The screenshot shows the NVIDIA License Server web interface with the "Licensed Clients" section selected in the left sidebar. The main content area is titled "Licensed Clients" and contains a table with the following data:

Client ID	Client ID Type	Client Type
00000000000000000000000000000000	ETHERNET	VIRTUAL

Page 1

Copyright (c) 2017 NVIDIA Corporation. All Rights Reserved

## Verify the NVIDIA configuration on the host

To obtain a hostwide overview of the NVIDIA GPUs, enter the `nvidia-smi` command without any arguments (Figures 57 and 58).

**Figure 57.** The `nvidia-smi` command output from the host with two NVIDIA P40 cards and 48 Microsoft Windows 10 desktops with P40-1B vGPU profile

```
-sh: nvidia-smi: not found
[root@M5:~] nvidia-smi
Wed Sep  6 00:43:04 2017
```

-----+-----					
NVIDIA-SMI		384.73			
		Driver Version: 384.73			
-----+-----					
GPU	Name	Bus-Id			GPU-Util
vGPU ID	Name	VM ID	VM Name		vGPU-Util
=====+=====					
0	Tesla P40	0000:5E:00.0			1%
38050	GRID P40-1B	38054	P40a-001		0%
38053	GRID P40-1B	38066	P40a-004		0%
46441	GRID P40-1B	46443	P40a-036		0%
46468	GRID P40-1B	46476	P40a-035		0%
46471	GRID P40-1B	46485	P40a-010		0%
46474	GRID P40-1B	46495	P40a-048		0%
46475	GRID P40-1B	46496	P40a-009		0%
46473	GRID P40-1B	46502	P40a-024		0%
46687	GRID P40-1B	46704	P40a-028		0%
46690	GRID P40-1B	46713	P40a-026		0%
46691	GRID P40-1B	46714	P40a-037		0%
46692	GRID P40-1B	46724	P40a-043		0%
46694	GRID P40-1B	46722	P40a-038		0%
46958	GRID P40-1B	46971	P40a-016		0%
46955	GRID P40-1B	46975	P40a-005		0%
46960	GRID P40-1B	46993	P40a-031		0%
46959	GRID P40-1B	46995	P40a-011		0%
47198	GRID P40-1B	47207	P40a-042		0%
47199	GRID P40-1B	47209	P40a-045		0%
47201	GRID P40-1B	47229	P40a-046		0%
47202	GRID P40-1B	47232	P40a-047		0%
47203	GRID P40-1B	47246	P40a-007		0%
47436	GRID P40-1B	47440	P40a-027		0%
47438	GRID P40-1B	47449	P40a-018		0%
-----+-----					
1	Tesla P40	0000:AF:00.0			1%
38051	GRID P40-1B	38059	P40a-002		0%
38052	GRID P40-1B	38065	P40a-003		0%
46442	GRID P40-1B	46450	P40a-014		0%
46470	GRID P40-1B	46480	P40a-022		0%
46472	GRID P40-1B	46487	P40a-008		0%
46469	GRID P40-1B	46481	P40a-006		0%
46686	GRID P40-1B	46696	P40a-044		0%
46688	GRID P40-1B	46699	P40a-041		0%
46689	GRID P40-1B	46708	P40a-013		0%
46693	GRID P40-1B	46716	P40a-033		0%
46695	GRID P40-1B	46719	P40a-034		0%
46953	GRID P40-1B	46963	P40a-032		0%
46954	GRID P40-1B	46967	P40a-021		0%
46956	GRID P40-1B	46973	P40a-020		0%
46957	GRID P40-1B	46970	P40a-039		0%
46962	GRID P40-1B	46999	P40a-015		0%
46961	GRID P40-1B	47020	P40a-017		0%
47196	GRID P40-1B	47206	P40a-019		0%
47197	GRID P40-1B	47208	P40a-030		0%
47200	GRID P40-1B	47226	P40a-029		0%
47205	GRID P40-1B	47247	P40a-040		0%
47204	GRID P40-1B	47250	P40a-012		0%
47437	GRID P40-1B	47442	P40a-023		0%
47439	GRID P40-1B	47450	P40a-025		0%
-----+-----					

**Figure 58.** The `nvidia-smi` command output from the host with two NVIDIA P6 cards and 32 Microsoft Windows 10 desktops with P6-1B vGPU profile

```

-sh: nvidia-smi: not found
[root@M5:~] nvidia-smi
Wed Sep  6 00:43:04 2017

```

NVIDIA-SMI		Driver Version: 384.73	
GPU	Name	Bus-Id	GPU-Util
vGPU ID	Name	VM ID	vGPU-Util
0	Tesla P6	0000:18:00.0	3%
39511	GRID P6-1B	39521	0%
39509	GRID P6-1B	39526	0%
39516	GRID P6-1B	39539	0%
39515	GRID P6-1B	39547	0%
39514	GRID P6-1B	39545	0%
39791	GRID P6-1B	39800	0%
39792	GRID P6-1B	39801	0%
39793	GRID P6-1B	39813	0%
39796	GRID P6-1B	39812	0%
39797	GRID P6-1B	39828	0%
40178	GRID P6-1B	40188	0%
40180	GRID P6-1B	40193	0%
40184	GRID P6-1B	40207	0%
40182	GRID P6-1B	40212	0%
40187	GRID P6-1B	40214	0%
40411	GRID P6-1B	40412	0%
1	Tesla P6	0000:D8:00.0	3%
38583	GRID P6-1B	38602	0%
39508	GRID P6-1B	39518	0%
39510	GRID P6-1B	39528	0%
39512	GRID P6-1B	39538	0%
39513	GRID P6-1B	39544	0%
39517	GRID P6-1B	39546	0%
39794	GRID P6-1B	39814	0%
39798	GRID P6-1B	39827	0%
39795	GRID P6-1B	39826	0%
39799	GRID P6-1B	39838	0%
40181	GRID P6-1B	40195	0%
40186	GRID P6-1B	40215	0%
40185	GRID P6-1B	40213	0%
40433	GRID P6-1B	40434	0%
40556	GRID P6-1B	40558	0%
40557	GRID P6-1B	40559	0%



**Figure 59.** The `nvidia-smi` command output from the host with two NVIDIA M10 cards and 32 Microsoft Windows 10 desktops with M10-1B vGPU profile

NVIDIA-SMI 384.37		Driver Version: 384.37			
GPU	Name	Bus-Id	GPU-Util		
vGPU ID	Name	VM ID	VM Name	vGPU-Util	
=====+=====+=====+=====+=====+=====					
0	Tesla M10	0000:60:00.0	0%		
1881359	GRID M10-1B	1881367	M10-VM2	0%	
1881363	GRID M10-1B	1881402	M10-VM7	0%	
1883906	GRID M10-1B	1883957	M10-VM21	0%	
1884913	GRID M10-1B	1884926	M10-VM28	0%	
1885756	GRID M10-1B	1885789	M10-VM34	0%	
1886654	GRID M10-1B	1886696	M10-VM46	0%	
1887660	GRID M10-1B	1887701	M10-VM53	0%	
1888536	GRID M10-1B	1888563	M10-VM57	0%	
-----+-----+-----+-----+-----+-----					
1	Tesla M10	0000:61:00.0	0%		
1881362	GRID M10-1B	1881383	M10-VM3	0%	
1882496	GRID M10-1B	1882531	M10-VM11	0%	
1883907	GRID M10-1B	1883965	M10-VM23	0%	
1884925	GRID M10-1B	1884971	M10-VM27	0%	
1885752	GRID M10-1B	1885773	M10-VM35	0%	
1885751	GRID M10-1B	1885781	M10-VM40	0%	
1887659	GRID M10-1B	1887703	M10-VM56	0%	
1888535	GRID M10-1B	1888537	M10-VM63	0%	
-----+-----+-----+-----+-----+-----					
2	Tesla M10	0000:62:00.0	0%		
1881360	GRID M10-1B	1881378	M10-VM6	0%	
1883908	GRID M10-1B	1883973	M10-VM22	0%	
1884911	GRID M10-1B	1884934	M10-VM31	0%	
1884910	GRID M10-1B	1884915	M10-VM15	0%	
1886650	GRID M10-1B	1886699	M10-VM44	0%	
1886655	GRID M10-1B	1886729	M10-VM45	0%	
1887658	GRID M10-1B	1887702	M10-VM52	0%	
1889738	GRID M10-1B	1889755	M10-VM64	0%	
-----+-----+-----+-----+-----+-----					
3	Tesla M10	0000:63:00.0	0%		
1881364	GRID M10-1B	1881399	M10-VM8	0%	
1882498	GRID M10-1B	1882539	M10-VM13	0%	
1883900	GRID M10-1B	1883910	M10-VM17	0%	
1885749	GRID M10-1B	1885757	M10-VM33	0%	
1886649	GRID M10-1B	1886691	M10-VM48	0%	
1887654	GRID M10-1B	1887669	M10-VM55	0%	
1888208	GRID M10-1B	1888210	M10-VM60	0%	
1889739	GRID M10-1B	1889765	M10-VM67	0%	
-----+-----+-----+-----+-----+-----					

## 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

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 graphics 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.



## Conclusion

The combination of Cisco UCS Manager, Cisco UCS C240 M5 Rack Servers and B200 M5 Blade Servers, and NVIDIA Tesla cards running on VMware vSphere 6.5 and Citrix XenDesktop 7.13 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

- 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>
  - [http://us.download.nvidia.com/Windows/Quadro\\_Certified/GRID/367.106/ESXi-6.5/367.106-370.12-grid-vgpu-release-notes-vmware-vsphere.pdf](http://us.download.nvidia.com/Windows/Quadro_Certified/GRID/367.106/ESXi-6.5/367.106-370.12-grid-vgpu-release-notes-vmware-vsphere.pdf)
- Citrix XenApp and XenDesktop 7.13:
  - <http://docs.citrix.com/en-us/xenapp-and-xendesktop/7-13.html>
  - <https://www.citrix.com/products/xenapp-xendesktop/hdx/hdx-3d-pro.html>
  - <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 6.5:
  - [http://kb.vmware.com/selfservice/microsites/search.do?language=en\\_US&cmd=displayKC&externalId=2033434http://pubs.vmware.com/vsphere-6-5/index.jsp](http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2033434http://pubs.vmware.com/vsphere-6-5/index.jsp)
  - <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 [www.cisco.com/go/offices](http://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: [www.cisco.com/go/trademarks](http://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)