# Integrate Cisco HyperFlex Systems and Cisco UCS M5 Servers with NVIDIA GRID 5.0 on VMware vSphere 6.5 and Citrix XenDesktop 7.15

February 2018

# Contents

## What you will learn

Cisco recently introduced the fifth-generation Cisco UCS® B-Series Blade Servers and C-Series Rack Servers and the Cisco HyperFlex™ hyperconverged 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 today's Cisco UCS B-Series Blade Servers and C-Series Rack Servers and Cisco HyperFlex hyperconverged servers, applications with demanding graphics requirements are now 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 B-Series and C-Series servers.

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. 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 Rack Servers and Cisco HyperFlex hyperconverged servers.

This new graphics capability helps enable organizations to centralize their graphics workloads and data in the data center. This capability greatly benefits organizations that need to be able to shift work 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 Cisco UCS servers offer these benefits:

- Support for full-length, full-power NVIDIA GRID cards in a 2-rack-unit (2RU) or 4RU form factor
- Support for a mezzanine form-factor adapter graphics processing unit (GPU) card in half-width and full-width blade servers
- 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 blade and rack servers with two NVIDIA GRID cards than with the 2-slot, 2.5-inch equivalent rack unit: the HP ProLiant WS460c Gen9 Graphics Server Blade with the GRID card in a second slot

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

- 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

An important element of this document's design is VMware's support for the NVIDIA GRID virtual graphics processing unit (vGPU) feature in VMware vSphere 6.5. Prior to Release 6.0 vSphere supported only virtual direct graphics acceleration (vDGA) and virtual shared graphics acceleration (vSGA), so support for vGPU in vSphere 6.0 and later releases greatly expands the range of deployment scenarios using the most versatile and efficient configuration of the GRID cards.

The purpose of this document is to help our partners and customers integrate NVIDIA GRID 5.0 graphics processing cards, Cisco HyperFlex systems, Cisco UCS B200 M5 Blade Servers, Cisco UCS C240 M5 Rack Servers, VMware vSphere, and Citrix XenDesktop 7.15 in vGPU mode.

Please contact our partners NVIDIA, Citrix, 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.

## Why use NVIDIA GRID vGPU for graphic deployments on Citrix XenDesktop

The NVIDIA GRID vGPU allows multiple virtual desktops to share a single physical GPU, and it allows multiple GPUs to reside on a single physical PCI card. All provide the 100 percent application compatibility of vDGA pass-through graphics, but with lower cost because multiple desktops share a single graphics card. With Citrix XenDesktop, you can centralize, pool, and more easily manage traditionally complex and expensive distributed workstations and desktops. Now all your user groups can take advantage of the benefits of virtualization.

The GRID vGPU capability brings the full benefits of NVIDIA hardware-accelerated graphics to virtualized solutions. This technology provides exceptional graphics performance for virtual desktops equivalent to PCs with an onboard graphics processor.

The GRID vGPU uses the industry's most advanced technology for sharing true GPU hardware acceleration among multiple virtual desktops—without compromising the graphics experience. Application features and compatibility are exactly the same as they would be at the user's desk.

With GRID vGPU technology, the graphics commands of each virtual machine are passed directly to the GPU, without translation by the hypervisor. By allowing multiple virtual machines to access the power of a single GPU in the virtualization server, enterprises can increase the number of users with access to true GPU-based graphics acceleration on virtual machines.

The physical GPU in the server can be configured with a specific vGPU profile. Organizations have a great deal of flexibility in how best to configure their servers to meet the needs of various types of end users.

vGPU support allows businesses to use the power of the NVIDIA GRID technology to create a whole new class of virtual machines designed to provide end users with a rich, interactive graphics experience.

### 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.

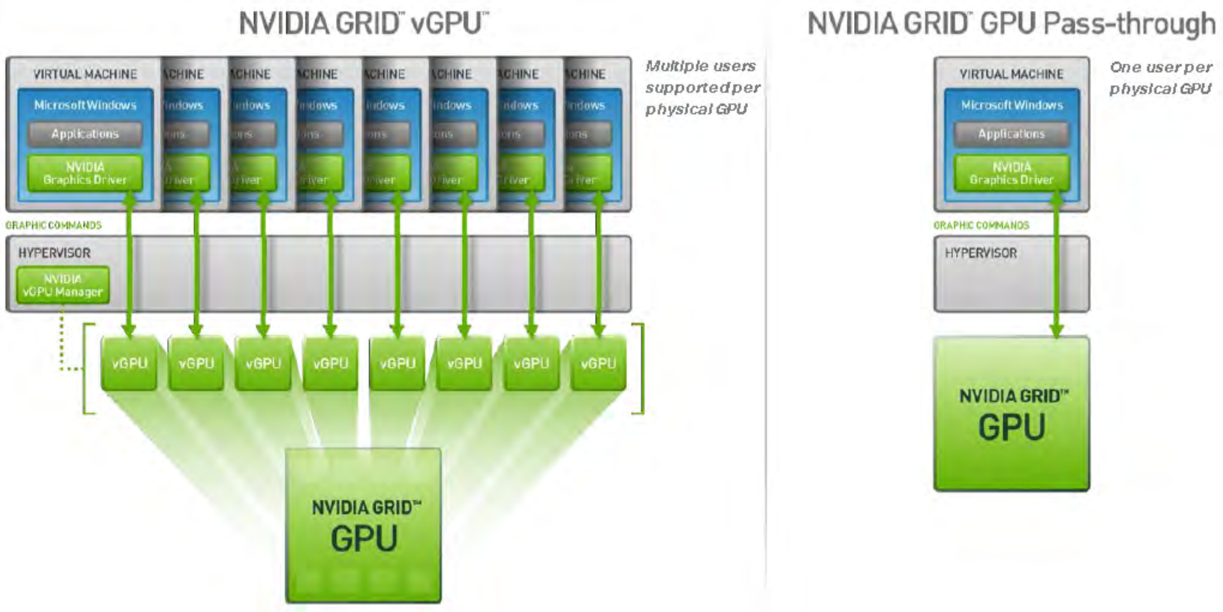
Although the needs of end users can be diverse, for simplicity users can be grouped into the following categories: knowledge workers, designers, and power users.

- 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).

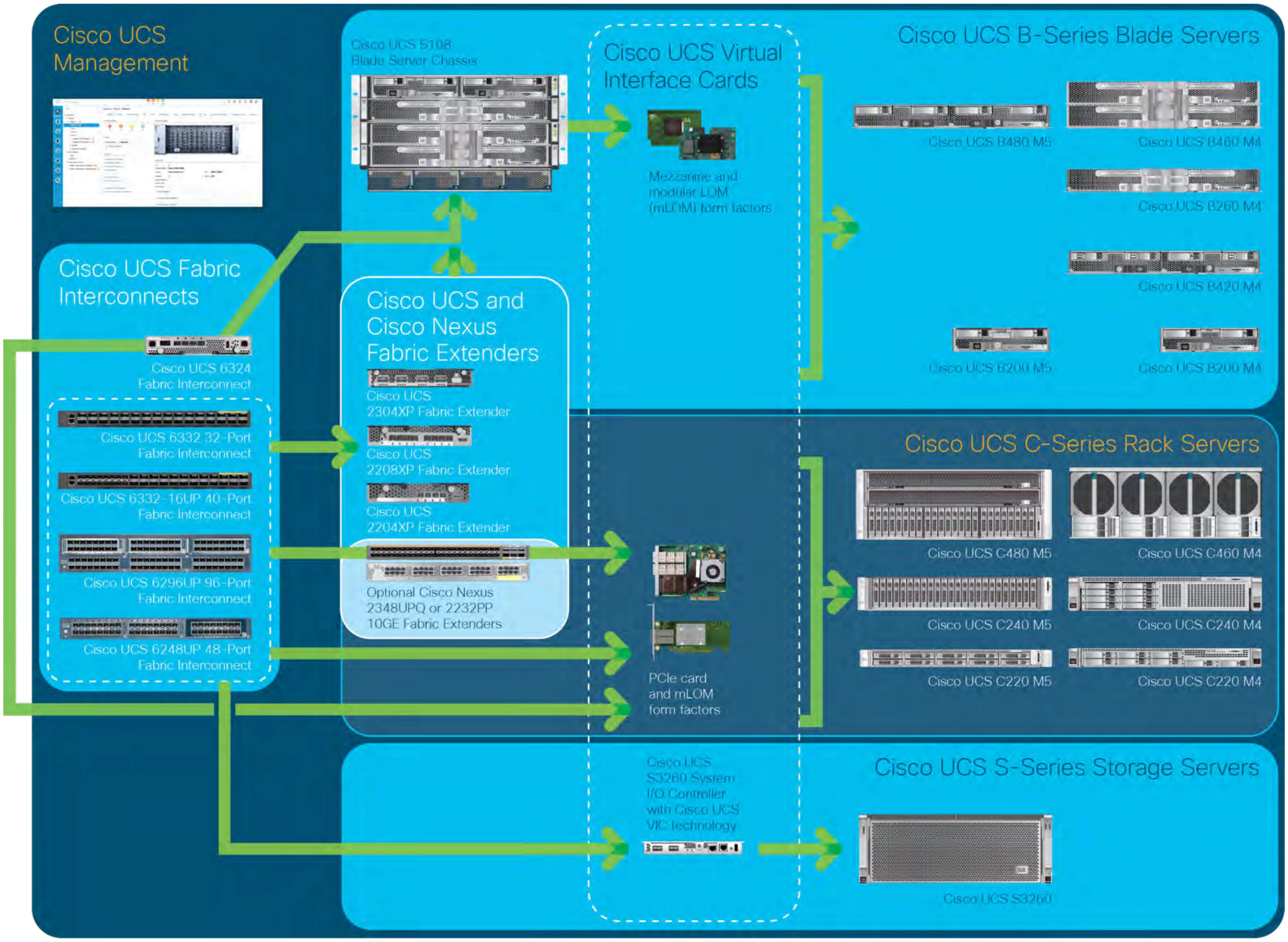
**Figure 1.** NVIDIA GRID vGPU system architecture



## 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


Cisco UCS Management
Cisco UCS Fabric Interconnects
Cisco UCS 6324 Fabric Interconnect
Cisco UCS 6332 32-Port Fabric Interconnect
Cisco UCS 6332-16UP 40-Port Fabric Interconnect
Cisco UCS 6296UP 96-Port Fabric Interconnect
Cisco UCS 6248UP 48-Port Fabric Interconnect
Cisco UCS 5108 Blade Server Chassis
Cisco UCS and Cisco Nexus Fabric Extenders
Cisco UCS 2304XP Fabric Extender
Cisco UCS 2208XP Fabric Extender
Cisco UCS 2204XP Fabric Extender
Optional Cisco Nexus 2348UPQ or 2232PP 10GE Fabric Extenders
Cisco UCS Virtual Interface Cards
Mezzanine and modular LOM (mLOM) form factors
PCIe card and mLOM form factors
Cisco UCS S3260 System I/O Controller with Cisco UCS VIC technology
Cisco UCS B-Series Blade Servers
Cisco UCS B480 M5
Cisco UCS B460 M4
Cisco UCS B260 M4
Cisco UCS B420 M4
Cisco UCS B200 M5
Cisco UCS B200 M4
Cisco UCS C-Series Rack Servers
Cisco UCS C480 M5
Cisco UCS C460 M4
Cisco UCS C240 M5
Cisco UCS C240 M4
Cisco UCS C220 M5
Cisco UCS C220 M4
Cisco UCS S-Series Storage Servers
Cisco UCS S3260

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 a 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. Tightly integrated Cisco UCS manager and NVidia GPU cards provide better management of firmware and graphics card configuration.

### Cisco UCS 6332 Fabric Interconnect

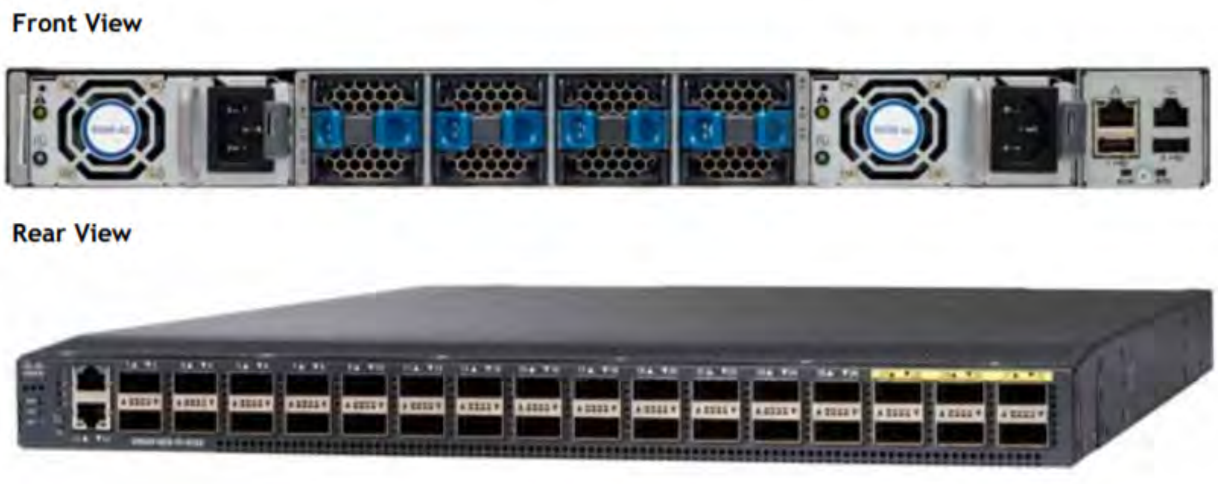
The Cisco UCS 6332 Fabric Interconnect (Figure 3) is the management and communication backbone for Cisco UCS B-Series Blade Servers, C-Series Rack Servers, and 5100 Series Blade Server Chassis. All servers attached to 6332 Fabric Interconnects become part of one highly available management domain.

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 details see, https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-b-series-blade-servers/6332-specsheet.pdf.

The 6332 Fabric Interconnect provides these features and capabilities:

- Bandwidth of up to 2.56-Tbps full-duplex throughput
- Thirty-two 40-Gbps Enhanced Quad Small Form-Factor Pluggable (QSFP+) ports in a 1RU form factor
- 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, Cisco UCS 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. Cisco UCS 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. Cisco UCS 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 the Intel Xeon Scalable processor family, 24 DIMM slots for 2666-MHz DDR4 DIMMs, and up to 128-GB capacity points, 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-GB SAS storage controller card. The C240 M5 server includes a dedicated internal mLOM slot for installation of a Cisco 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 2 Intel Xeon Scalable CPUs
- Up to 6 PCIe 3.0 slots (4 full-height, full-length for GPU)
- Six hot-swappable fans for front-to-rear cooling
- Twenty-four SFF front-facing SAS/SATA hard-disk drives (HDDs) or SAS/SATA solid-state drives (SSDs)
- Optionally, up to two front-facing SFF NVMe PCIe SSDs (replacing SAS/SATA drives). These drives must be placed in front drive bays 1 and 2 only and are controlled from Riser 2 option C.
- Optionally, up to two SFF, rear-facing SAS/SATA HDDs/SSDs or up to two rear-facing SFF NVMe PCIe SSDs
    - Rear facing SFF NVMe drives connected from Riser 2, Option B or C
    - Support for 12-Gbps SAS drives
- Dedicated mLOM slot on the motherboard, which can flexibly accommodate the following cards:
    - Cisco VICs
    - Quad-port Intel i350 1 Gigabit Ethernet RJ-45 mLOM NIC
- Two 1 Gigabit Ethernet embedded LOM ports
- Support for up to two double-wide NVIDIA 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 MicroSD card on PCIe Riser 1 (Option 1 and 1B)
    - The MicroSD card serves as a dedicated local resource for utilities such as the Cisco Host Upgrade Utility (HUU).
    - Images can be pulled from a file share (Network File System [NFS] or Common Internet File System [CIFS]) and uploaded to the cards for future use.
- A mini-storage module connector on the motherboard supports either:
    - SD card module with two Secure Digital (SD) card slots; mixing different capacity SD cards is not supported
    - M.2 module with two SATA M.2 SSD slots; mixing different capacity M.2 modules is not supported

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

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

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

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

**Figure 4.** Cisco UCS C240 M5 Rack Server

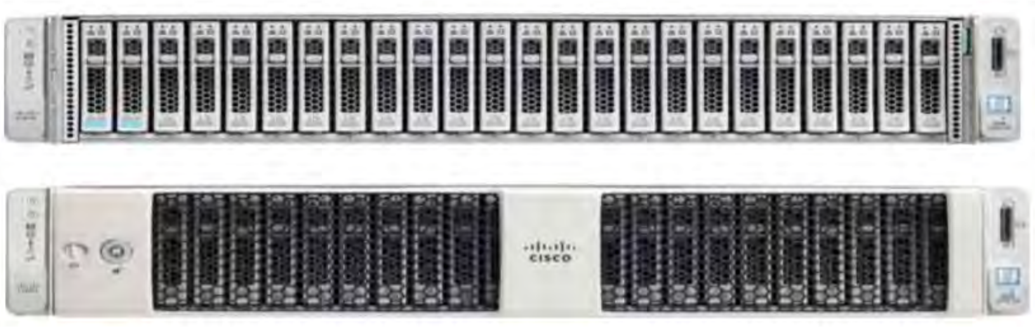


**Figure 5.** Cisco UCS C240 M5 Rack Server rear view

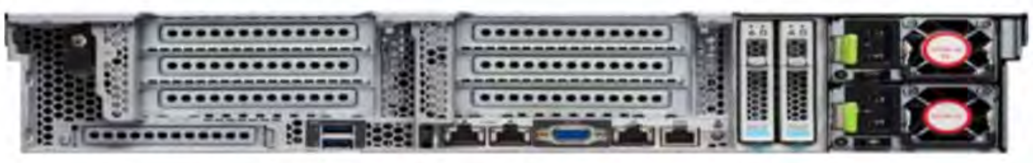


Table 1. Cisco UCS C240 M5 PCIe slots

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

### Cisco UCS Virtual Interface Card 1387

The Cisco UCS VIC 1387 (Figure 6) is a dual-port Enhanced Small Form-Factor Pluggable (SFP+) 40-Gbps Ethernet and FCoE-capable PCIe mLOM adapter installed in 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 CNA

## 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, from IT and web infrastructure through distributed database. The enterprise-class B200 M5 Blade Server 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 connectivity with throughput of up to 80 Gbps.

The Cisco UCS 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 additional connectivity bandwidth or with an NVIDIA P6 GPU. These hardware options enable an additional four ports of the VIC 1340, bringing the total capability of the VIC 1340 to a dual native 40-Gbps interface or a dual 4 x 10 Gigabit Ethernet port-channel interface, respectively. Alternatively, the same rear mezzanine adapter slot can be configured with an NVIDIA P6 GPU.

The B200 M5 also has one front mezzanine slot. The B200 M5 can be ordered with or without a front mezzanine card. The front mezzanine card can accommodate a storage controller or 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 Virtual Interface Card 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.

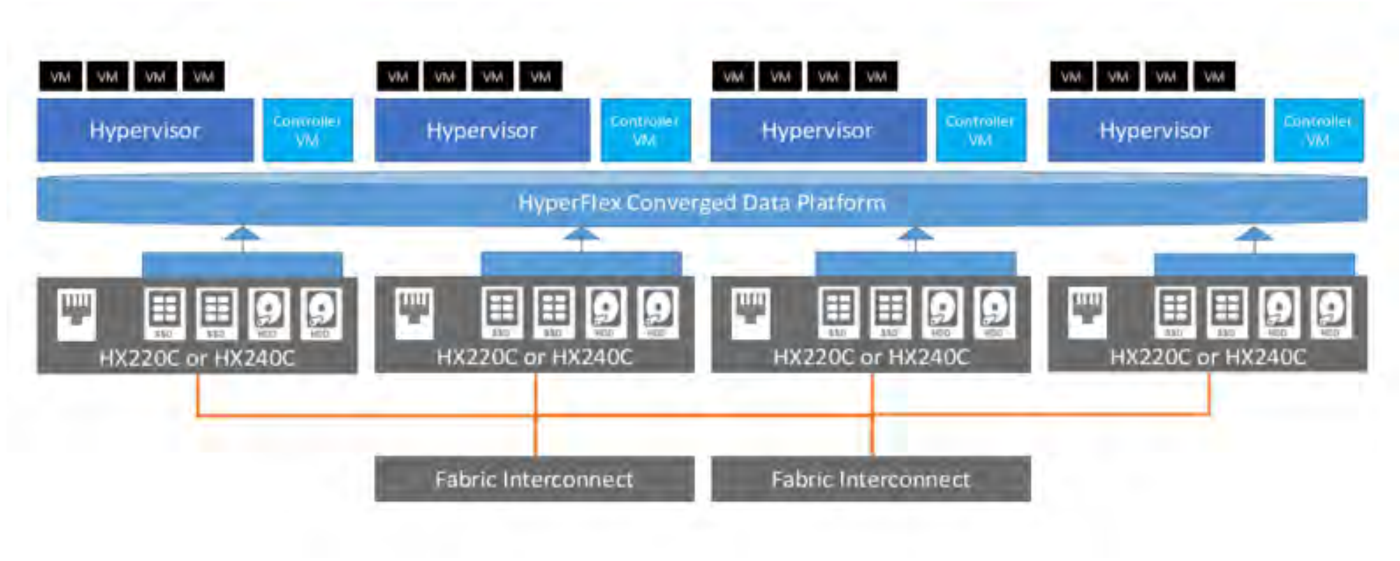
For more information, see https://www.cisco.com/c/en/us/products/collateral/interfaces-modules/ucs-virtual-interface-card-1340/datasheet-c78-732517.html.

**Figure 8.** Cisco UCS VIC 1340

## Cisco HyperFlex system

The Cisco HyperFlex system provides a fully contained virtual server platform. It includes computing and memory resources, integrated networking connectivity, a distributed high-performance log-structured file system for virtual machine storage, and hypervisor software for running the virtualized servers, all in a single Cisco UCS management domain (Figure 9).

**Figure 9.** Cisco HyperFlex system overview

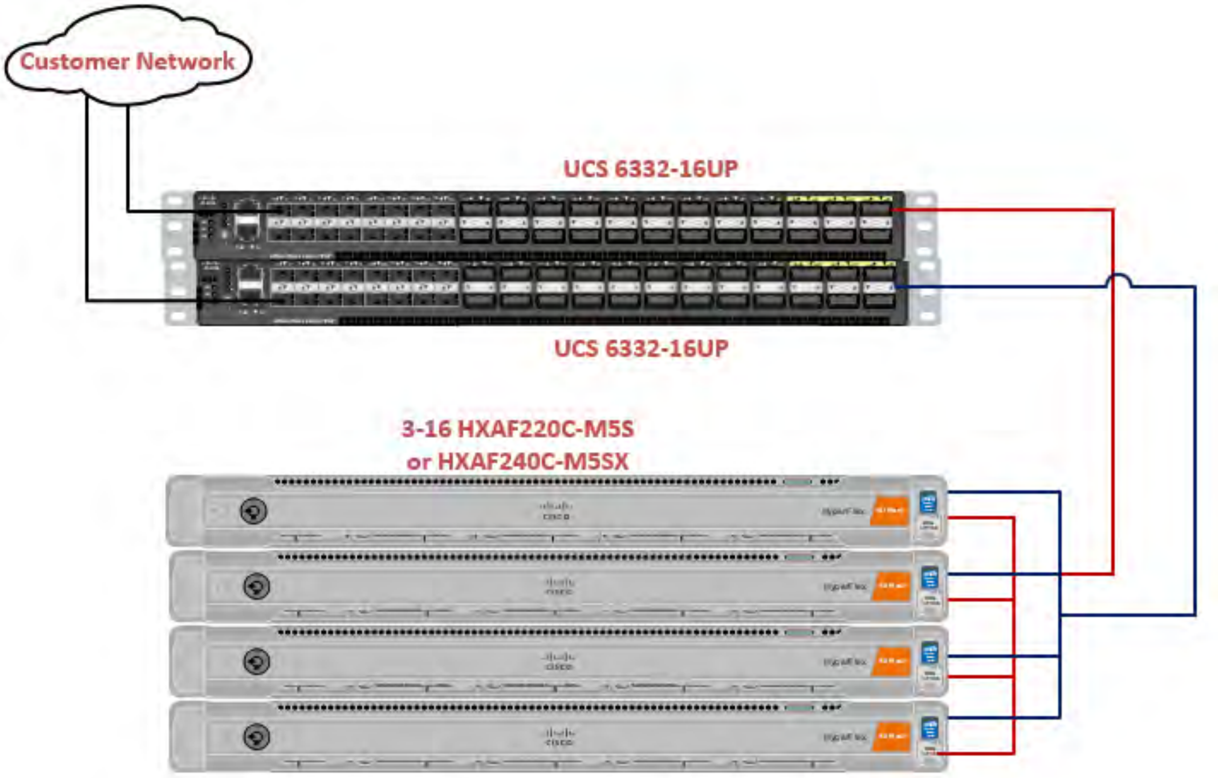


The Cisco HyperFlex system is composed of a pair of Cisco UCS 6200 or 6300 Series Fabric Interconnects, along with up to 16 Cisco HyperFlex HX-Series all-flash rack-mount servers per cluster. In addition, up to 16 computing-only servers can be added per cluster. Adding the Cisco UCS 5108 Blade Server Chassis allows the use of Cisco UCS B200 M5 Blade Servers for additional computing resources in a hybrid cluster design. Cisco UCS C240 and C220 servers can also be used for additional computing resources. Up to 8 separate HX-Series clusters can be installed under a single pair of fabric interconnects. The fabric interconnects connect both to every HX-Series rack mount server and to every Cisco UCS 5108 Blade Server Chassis. Upstream network connections, also referred to as northbound network connections, are made from the fabric interconnects to the customer data center network at the time of installation.

For the configuration used in this document, Cisco UCS 6332-16UP Fabric Interconnects are uplinked to Cisco Nexus® 9372PX Switches.

Figure 10 illustrates the hyperconverged topology used for this document.

**Figure 10.** Cisco HyperFlex standard topology



Additional graphics support for Cisco HyperFlex clusters can be achieved by adding Cisco UCS B200 M5 Blade Servers or Cisco UCS C240 M5 Rack Servers to the cluster as computing-only nodes.

To learn more about Cisco HyperFlex HX-Series servers, see http://www.cisco.com/c/en/us/products/hyperconverged-infrastructure/index.html.

## NVIDIA GRID

NVIDIA GRID is the industry's most advanced technology for sharing physical GPUs 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 and Pascal 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 cards

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

Table 2. Technical specifications for NVIDIA GRID cards

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

## 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 details 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 chipsets, 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 into 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 distributed virtual switch (DVS) 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 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 VMware 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:** VMware 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 Citrix Virtual GPU Solution.

### GPU acceleration for Microsoft Windows desktops

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

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

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

HDX 3D Pro offers the following features:

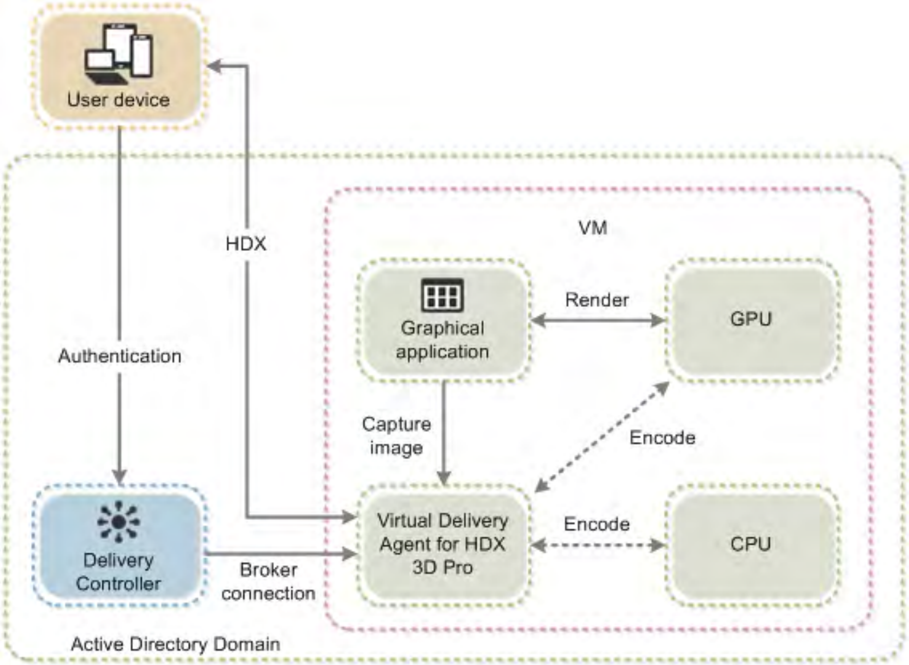
- Adaptive H.264-based deep compression for optimal WAN and wireless performance: HDX 3D Pro uses CPU-based full-screen H.264 compression as the default compression technique for encoding. Hardware encoding is used with NVIDIA cards that support NVIDIA NVENC.
- Lossless compression option for specialized use cases: HDX 3D Pro offers a CPU-based lossless codec to support applications that require pixel-perfect graphics, such as medical imaging. True lossless compression is recommended only for specialized use cases because it consumes significantly more network and processing resources.
    - When you use lossless compression:
        - The lossless indicator, a system-tray icon, shows the user whether the screen displayed is a lossy frame or a lossless frame. This information is helpful when the Visual Quality policy setting specifies a lossless build. The lossless indicator turns green when the frames sent are lossless.
        - The lossless switch enables the user to change to Always Lossless mode at any time in the session. To select or deselect Always Lossless at any time in a session, right-click the icon or use the shortcut Alt+Shift+1.
    - For lossless compression, HDX 3D Pro uses the lossless codec for compression regardless of the codec selected through policy.
    - For lossy compression, HDX 3D Pro uses the original codec: either the default or the one selected through policy.
    - Lossless switch settings are not retained for subsequent sessions. To use the lossless codec for every connection, select Always Lossless for the Visual Quality policy setting.
- Multiple and high-resolution monitor support: For Microsoft Windows 7 and 8 desktops, HDX 3D Pro supports user devices with up to four monitors. Users can arrange their monitors in any configuration and can mix monitors with different resolutions and orientations. The number of monitors is limited by the capabilities of the host computer GPU, the user device, and the available bandwidth. HDX 3D Pro supports all monitor resolutions and is limited only by the capabilities of the GPU on the host computer.
- Dynamic resolution: You can resize the virtual desktop or application window to any resolution.
- Support for NVIDIA Kepler architecture: HDX 3D Pro supports NVIDIA GRID K1 and K2 cards for GPU pass-through and GPU sharing. The GRID vGPU enables multiple virtual machines to have simultaneous, direct access to a single physical GPU, using the same NVIDIA graphics drivers as are deployed on nonvirtualized operating systems.
- Support for VMware vSphere and ESX using vDGA: You can use HDX 3D Pro with vDGA for both remote desktop service (RDS) and 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 11:

- 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 11.** 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 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:

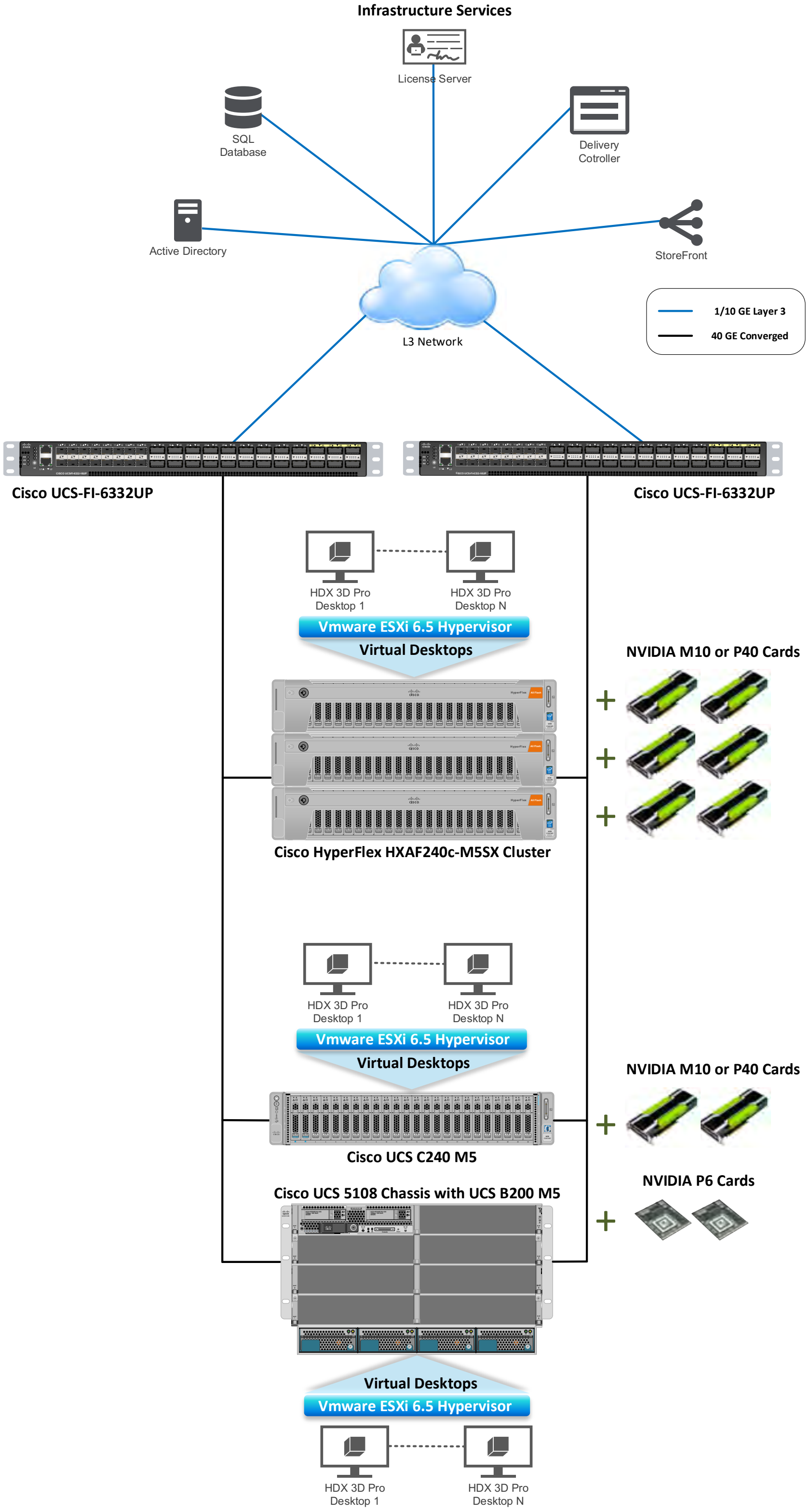
- HDX 3D Pro supports all monitor resolutions that are supported by the GPU on the host computer. However, for optimal performance with the minimum recommended user device and GPU specifications, Citrix recommends a maximum monitor resolution for user devices of 1920 x 1200 pixels for LAN connections, 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 optimal 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/7-15-ltsr/graphics/hdx-3d-pro.html and http://www.citrix.com/products/xenapp-xendesktop/hdx-3d-pro.html.

## Solution configuration

Figure 12 provides an overview of the solution configuration.

**Figure 12.** Reference architecture



The hardware components in the solution are:

- Cisco UCS C240 M5 Rack Server (two Intel Xeon Gold 5120 CPUs at 2.20 GHz) with 768 GB of memory (64 GB x 12 DIMMs at 2666 MHz)
- Cisco UCS B200 M5 Blade Server (two Intel Xeon Gold 5120 CPUs at 2.20 GHz) with 768 GB of memory (64 GB x 12 DIMMs at 2666 MHz)
- Cisco HyperFlex HX240c M5SX All Flash hyperconverged server (two Intel Xeon Gold 5120 CPUs at 2.20 GHz) with 768 GB of memory (64 GB x 12 DIMMs at 2666 MHz)
- Cisco UCS VIC 1387 mLOM (Cisco UCS C240 M5 Rack Server and Cisco HyperFlex HX240c M5S All Flash Node)
- Cisco UCS VIC 1340 mLOM (Cisco UCS B200 M5 Blade Server)
- Two Cisco UCS 6332 Fabric Interconnects (third-generation fabric interconnects)
- NVIDIA Tesla M10, P40, and P6 cards
- Two Cisco Nexus 9372 switches (optional access switches)

**Note:** For high-performance graphic applications, Intel Xeon processor 6154 CPUs at 3.0 GHz paired with NVIDIA Tesla P40 or P6 cards are recommended.

The software components of the solution are:

- Cisco UCS Firmware Release 3.2(2 c)
- Cisco HXDP 2.6(1a)
- VMware ESXi 6.5 (5969303) for VDI hosts
- Citrix XenApp and XenDesktop 7.15
- Microsoft Windows 10 64-bit
- Microsoft Server 2016
- Microsoft Office 2016
- NVIDIA GRID software and licenses:
- NVIDIA-VMware_ESXi_6.5_Host_Driver_384.99-1OEM.650.0.0.4598673
  - 385.90_grid_win10_server2016_64bit_international
  - vGPU License Server Version 5.0.0.22575570 with Quadro-Virtual-DWS licenses

## Configure Cisco UCS

This section describes the Cisco UCS configuration.

### Install NVIDIA Tesla GPU card on Cisco UCS C240 M5 and Cisco HyperFlex HX240c M5 All Flash server

Install the M10 or P40 GPU card on the Cisco UCS C240 M5 Rack Server and Cisco HyperFlex HX240c M5 All Flash Node.

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

Table 3. Minimum server firmware versions required for GPU cards

| Cisco Integrated Management Controller (IMC) | BIOS minimum version |
|---|---|
| NVIDIA Tesla M10 | Release 3.1(1) |
| NVIDIA Tesla P40 | Release 3.1(1) |

Mixing different brands or models of GPU cards in the server is not supported.

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 13 shows a one-GPU installation, and Figure 14 shows a two-GPU installation.

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

| Single GPU | Dual GPU |
|---|---|
| Riser 1, slot 2<br>or<br>Riser 2, slot 5<br>supported in all riser options | Riser 1, slot 2<br>and<br>Riser 2A/2B, slot 5 |

**Figure 13.** One-GPU scenario

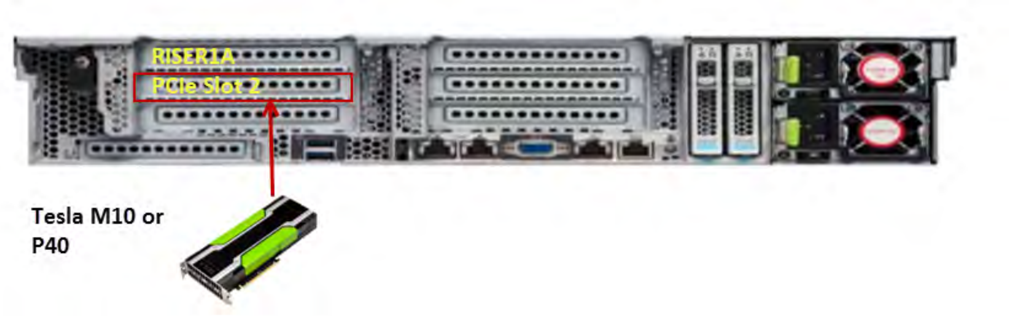
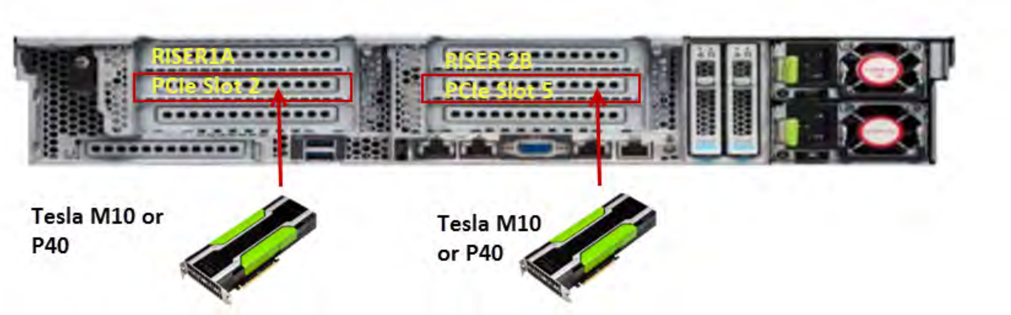


**Figure 14.** Two-GPU scenario



For more information, refer to these configuration documents:

- https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/c/hw/C240M5/install/C240M5.pdf
- https://www.cisco.com/c/dam/en/us/products/collateral/hyperconverged-infrastructure/hyperflex-hx-series/hxaf-240c-m5-specsheet.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 required for the GPU card. Figure 15 shows the card in the server.

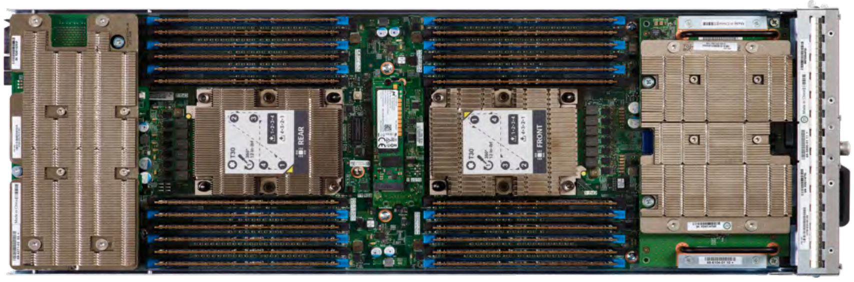
Table 5. Minimum server firmware versions required for GPU card

| Cisco Integrated Management Controller (IMC) | BIOS minimum version |
|---|---|
| NVIDIA Tesla M6 | 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 a 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 system 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 15.** Cisco UCS B200 M5 Blade Server with two NVIDIA GRID P6 GPU cards



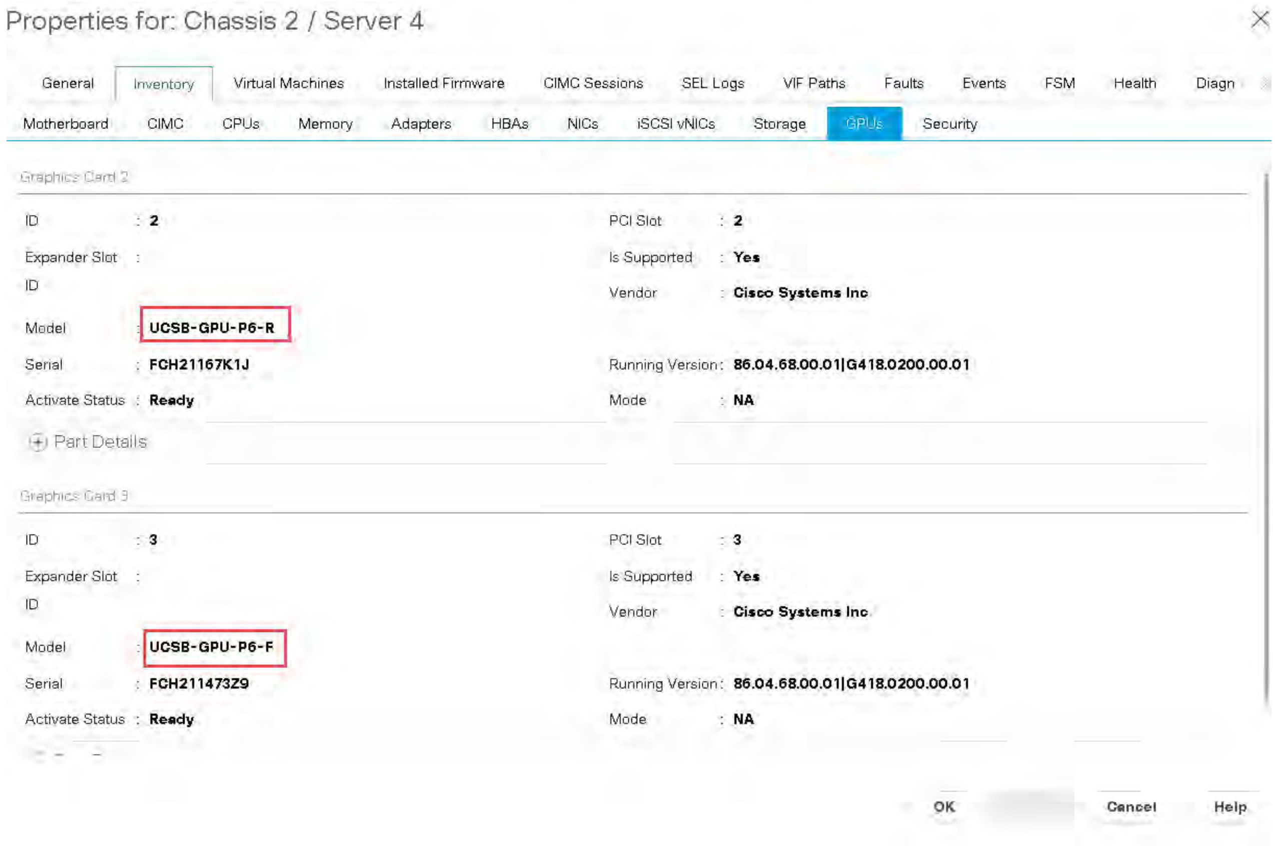
For more information, refer to this configuration document:
https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/hw/blade-servers/B200M5.pdf.

### Configure the GPU card

Follow these steps to configure the GPU card.

1. After the NVIDIA P6 GPU cards are 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 16, PCIe slots 2 and 3 are used with two GRID P6 cards.

**Figure 16.** NVIDIA GRID P6 card inventory displayed in Cisco UCS Manager



Properties for: Chassis 2 / Server 4

Graphics Card 2

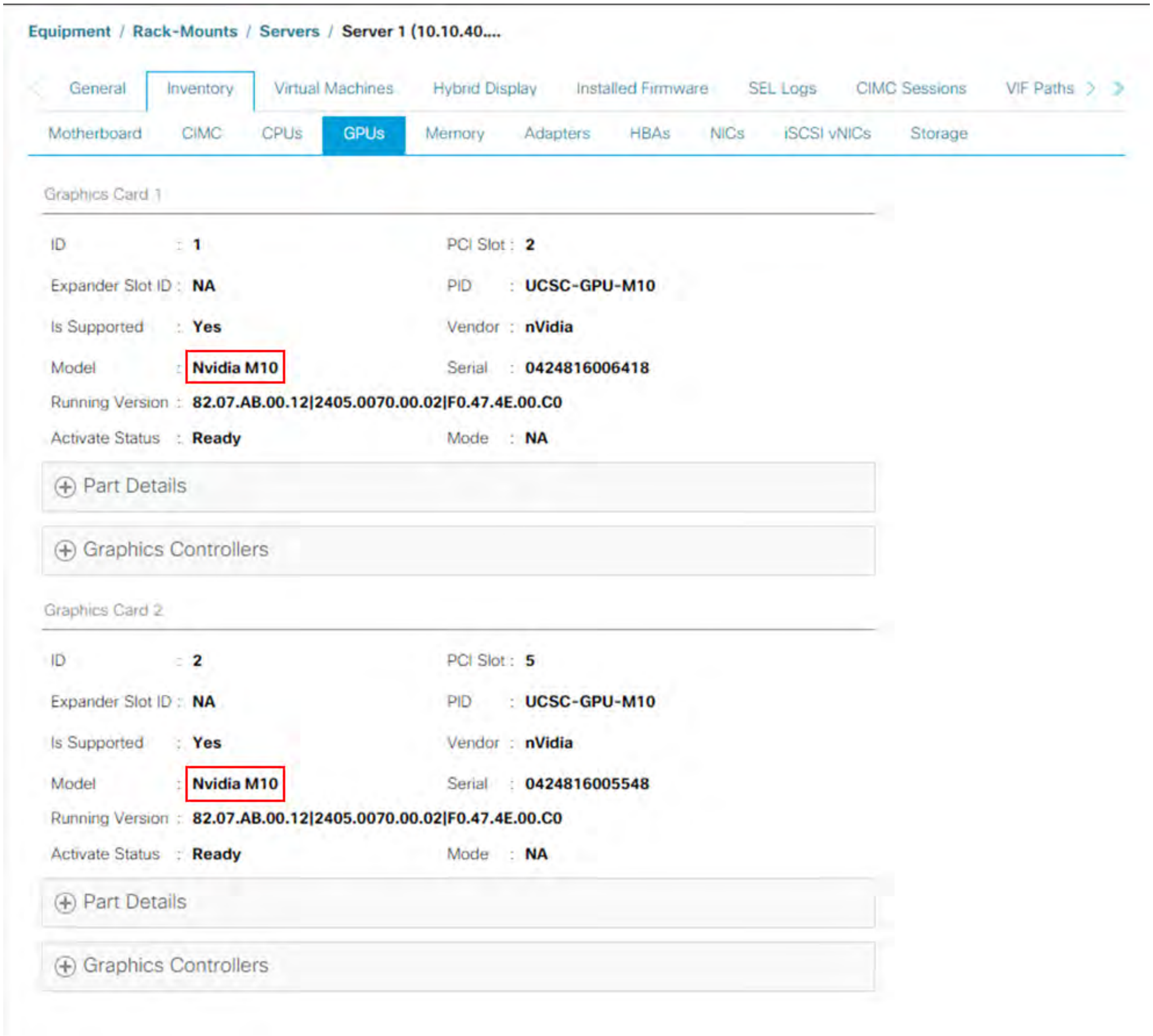
ID : 2
Expander Slot ID :
Model : UCSB-GPU-P6-R
Serial : FCH21167K1J
Activate Status : Ready
PCI Slot : 2
Is Supported : Yes
Vendor : Cisco Systems Inc
Running Version : 86.04.68.00.01|G418.0200.00.01
Mode : NA

Graphics Card 3

ID : 3
Expander Slot ID :
Model : UCSB-GPU-P6-F
Serial : FCH211473Z9
Activate Status : Ready
PCI Slot : 3
Is Supported : Yes
Vendor : Cisco Systems Inc
Running Version : 86.04.68.00.01|G418.0200.00.01
Mode : NA

2. After the NVIDIA M10 GPU cards are physically installed and the Cisco UCS C240 M5 and Cisco HyperFlex HX240c M5 All Flash 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 two GRID M10 cards.

**Figure 17.** NVIDIA GRID M10 card inventory displayed in Cisco UCS Manager



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 : 0424816006418  
Running Version : 82.07.AB.00.12|2405.0070.00.02|F0.47.4E.00.C0  
Activate Status : Ready  
Mode : NA

⊕ Part Details

⊕ Graphics Controllers

Graphics Card 2

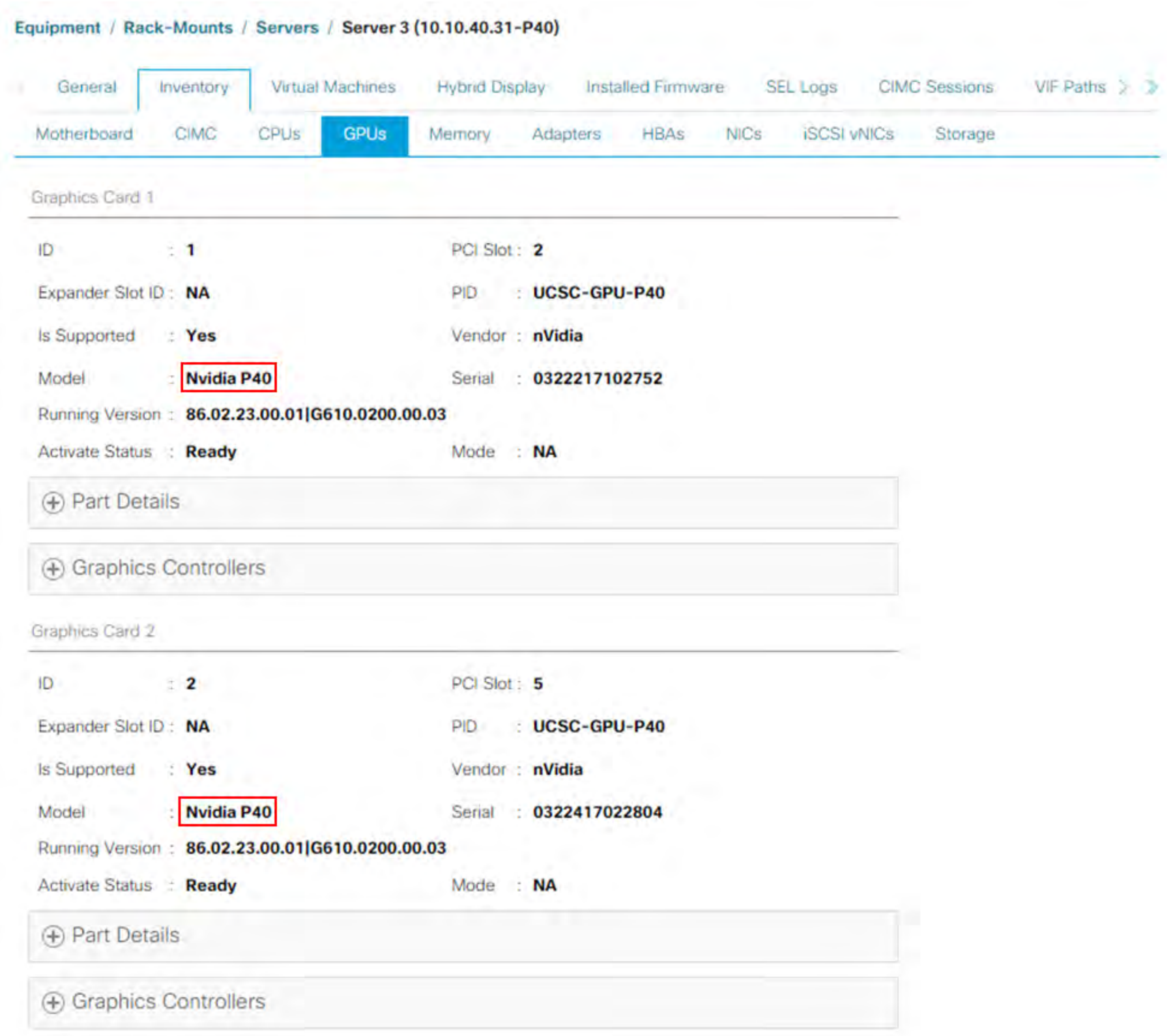
ID : 2  
PCI Slot : 5  
Expander Slot ID : NA  
PID : UCSC-GPU-M10  
Is Supported : Yes  
Vendor : nVidia  
Model : Nvidia M10  
Serial : 0424816005548  
Running Version : 82.07.AB.00.12|2405.0070.00.02|F0.47.4E.00.C0  
Activate Status : Ready  
Mode : NA

⊕ Part Details

⊕ Graphics Controllers

3. After the NVIDIA P40 GPU card is physically installed and the Cisco UCS C240 M5 and Cisco HyperFlex HX240c M5 All Flash server is discovered in Cisco UCS Manager, select the server and choose Inventory > GPUs. As shown in Figure 18, PCIe slots 2 and 5 are used with the two GRID P40 cards.

**Figure 18.** NVIDIA GRID P40 card inventory displayed in Cisco UCS Manager



Graphics Card 1

ID : 1
PCI Slot : 2
Expander Slot ID : NA
PID : UCSC-GPU-P40
Is Supported : Yes
Vendor : nVidia
Model : Nvidia P40
Serial : 0322217102752
Running Version : 86.02.23.00.01|G610.0200.00.03
Activate Status : Ready
Mode : NA

Graphics Card 2

ID : 2
PCI Slot : 5
Expander Slot ID : NA
PID : UCSC-GPU-P40
Is Supported : Yes
Vendor : nVidia
Model : Nvidia P40
Serial : 0322417022804
Running Version : 86.02.23.00.01|G610.0200.00.03
Activate Status : Ready
Mode : NA

You can use Cisco UCS Manager to perform firmware upgrades to the NVIDIA GPU cards in managed Cisco UCS C240 M5 and Cisco HyperFlex HX240c M5 All Flash servers.

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

## Install the NVIDIA GRID license server

This section summarizes the installation and configuration process for the GRID 5.0 license server.

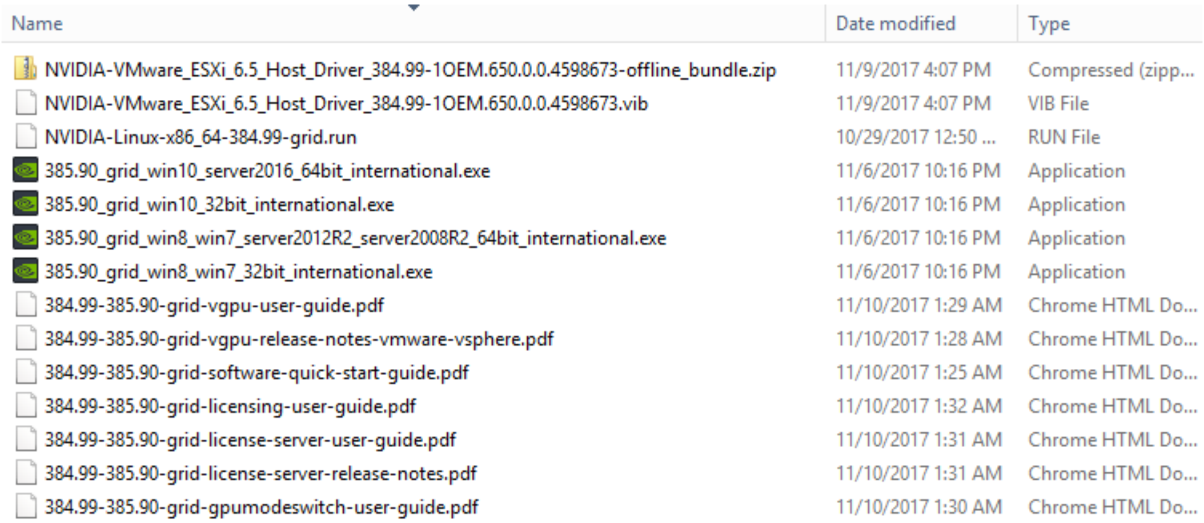
The NVIDIA GRID vGPU is a licensed feature on Tesla P6, P40, and M10 cards. A software license is required to use the full vGPU feature set on a guest virtual machine. An NVIDIA license server with the appropriate licenses is required.

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.

Three packages are required for VMware ESXi host setup, as shown in Figure 19:

- 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 19.** Software required for NVIDIA GRID 5.0 setup on the VMware ESXi host



| Name | Date modified | Type |
|---|---|---|
| NVIDIA-VMware_ESXi_6.5_Host_Driver_384.99-1OEM.650.0.0.4598673-offline_bundle.zip | 11/9/2017 4:07 PM | Compressed (zipp... |
| NVIDIA-VMware_ESXi_6.5_Host_Driver_384.99-1OEM.650.0.0.4598673.vib | 11/9/2017 4:07 PM | VIB File |
| NVIDIA-Linux-x86_64-384.99-grid.run | 10/29/2017 12:50 ... | RUN File |
| 385.90_grid_win10_server2016_64bit_international.exe | 11/6/2017 10:16 PM | Application |
| 385.90_grid_win10_32bit_international.exe | 11/6/2017 10:16 PM | Application |
| 385.90_grid_win8_win7_server2012R2_server2008R2_64bit_international.exe | 11/6/2017 10:16 PM | Application |
| 385.90_grid_win8_win7_32bit_international.exe | 11/6/2017 10:16 PM | Application |
| 384.99-385.90-grid-vgpu-user-guide.pdf | 11/10/2017 1:29 AM | Chrome HTML Do... |
| 384.99-385.90-grid-vgpu-release-notes-vmware-vsphere.pdf | 11/10/2017 1:28 AM | Chrome HTML Do... |
| 384.99-385.90-grid-software-quick-start-guide.pdf | 11/10/2017 1:25 AM | Chrome HTML Do... |
| 384.99-385.90-grid-licensing-user-guide.pdf | 11/10/2017 1:32 AM | Chrome HTML Do... |
| 384.99-385.90-grid-license-server-user-guide.pdf | 11/10/2017 1:31 AM | Chrome HTML Do... |
| 384.99-385.90-grid-license-server-release-notes.pdf | 11/10/2017 1:31 AM | Chrome HTML Do... |
| 384.99-385.90-grid-gpumodeswitch-user-guide.pdf | 11/10/2017 1:30 AM | Chrome HTML Do... |

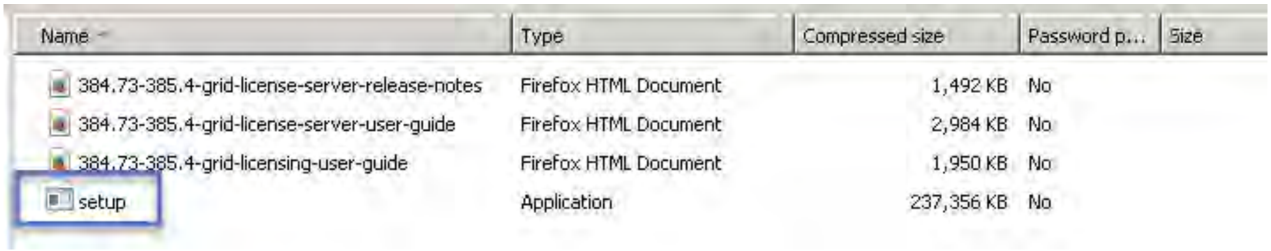
## Install the NVIDIA GRID 5.0 license server

The steps shown here use the Microsoft Windows version of the license server installed on Windows Server 2016. 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 and install the latest version.

1. Extract and open the NVIDIA-ls-windows-2017.08-0001 folder. Run setup.exe (Figure 20).

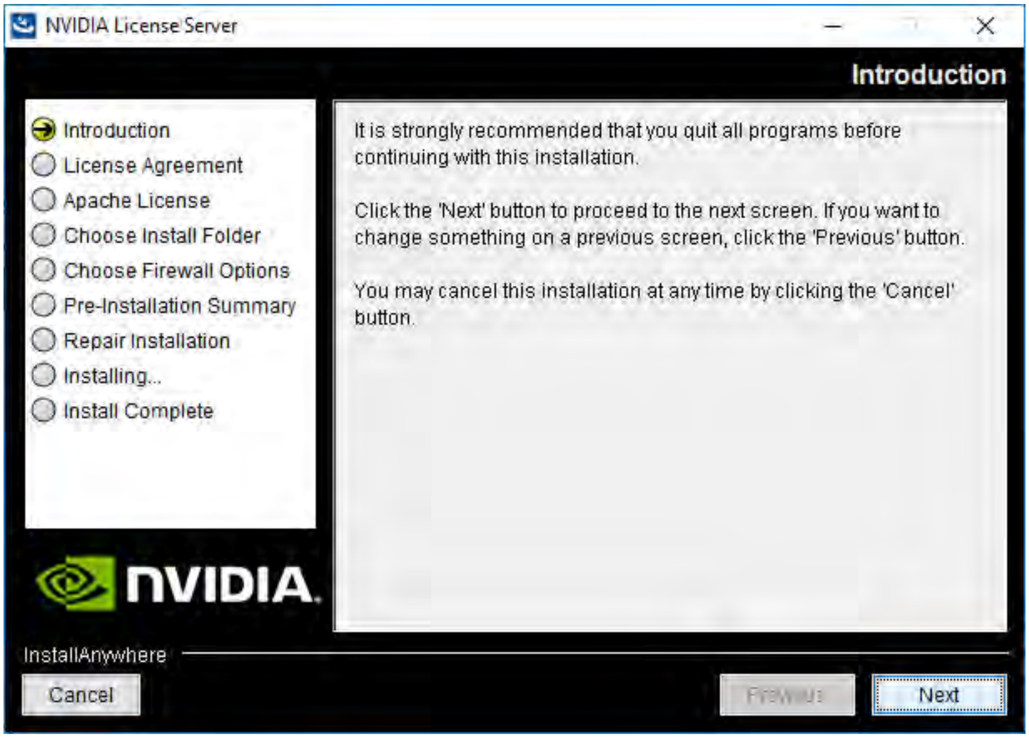
**Figure 20.** Run setup.exe



| 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 21).

**Figure 21.** NVIDIA License Server page



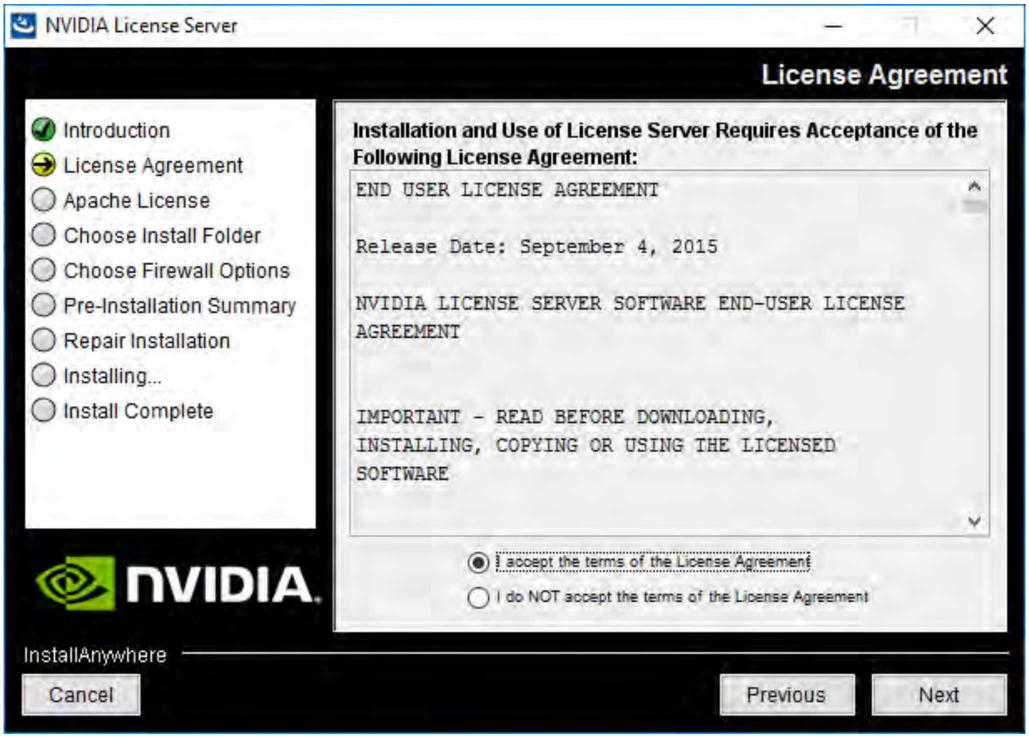
It is strongly recommended that you quit all programs before continuing with this installation.

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

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

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

**Figure 22.** NVIDIA License Agreement page



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

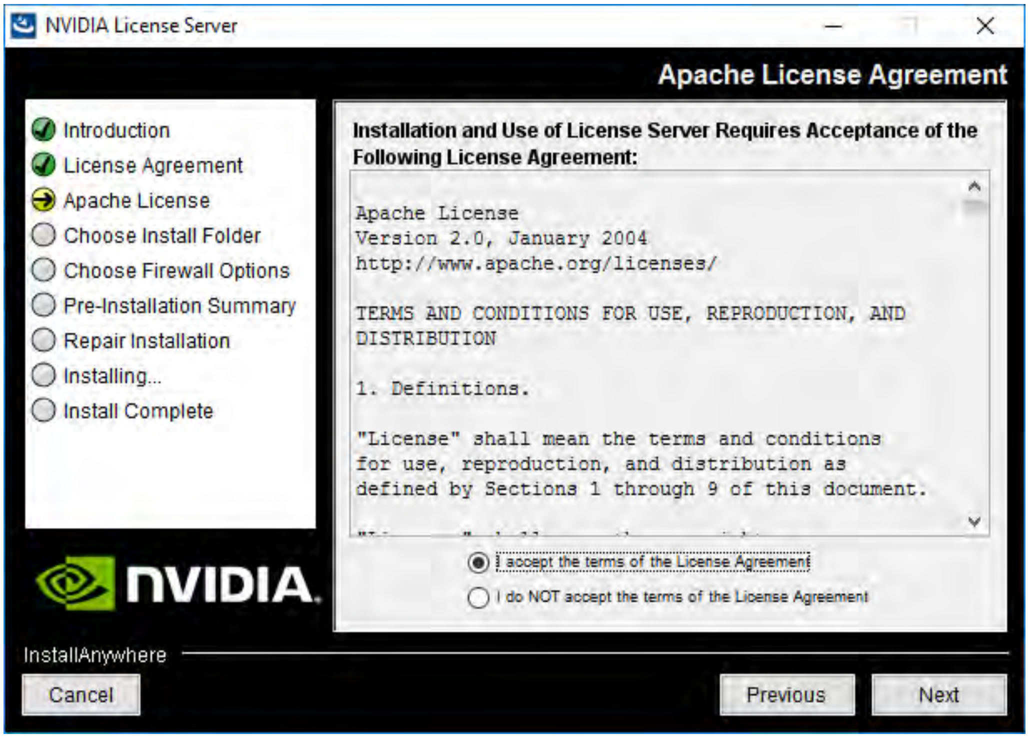
```
END USER LICENSE AGREEMENT

Release Date: September 4, 2015

NVIDIA LICENSE SERVER SOFTWARE END-USER LICENSE
AGREEMENT


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

● I accept the terms of the License Agreement

○ I do NOT accept the terms of the License Agreement

4. Accept the Apache license agreement and click Next (Figure 23).

**Figure 23.** Apache License Agreement page



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

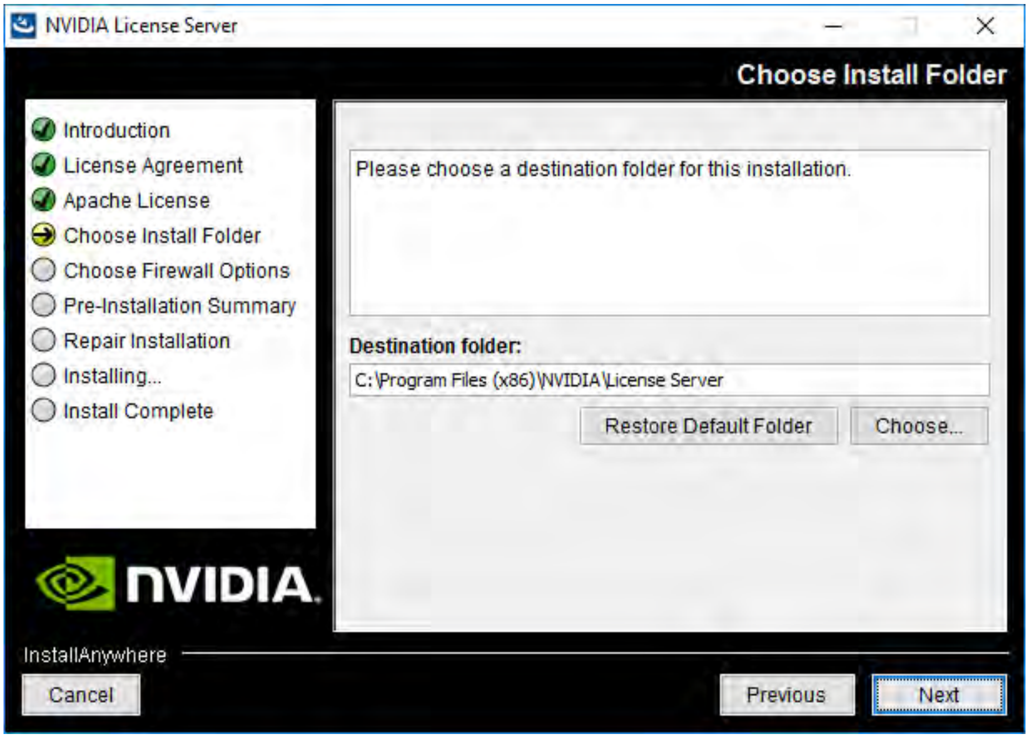
```
Apache License
Version 2.0, January 2004
http://www.apache.org/licenses/

TERMS AND CONDITIONS FOR USE, REPRODUCTION, AND
DISTRIBUTION

1. Definitions.

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

● I accept the terms of the License Agreement

○ I do NOT accept the terms of the License Agreement

5. Choose the desired installation folder and click Next (Figure 24).

**Figure 24.** Choosing a destination folder

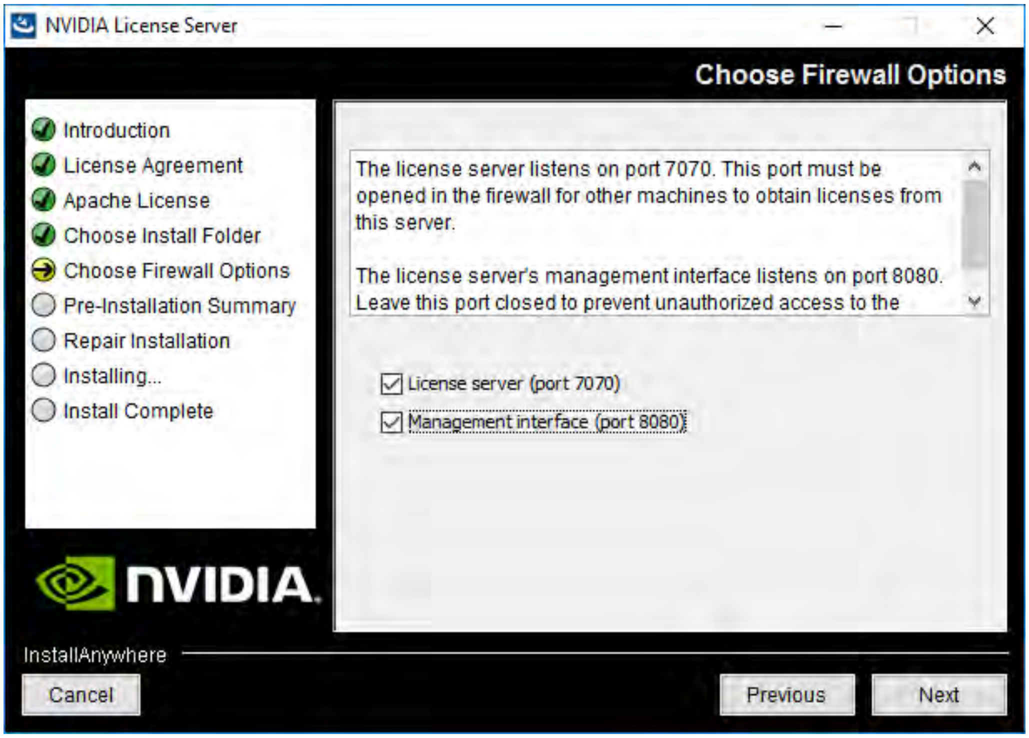


Please choose a destination folder for this installation.

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

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 25).

**Figure 25.** Setting firewall options



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

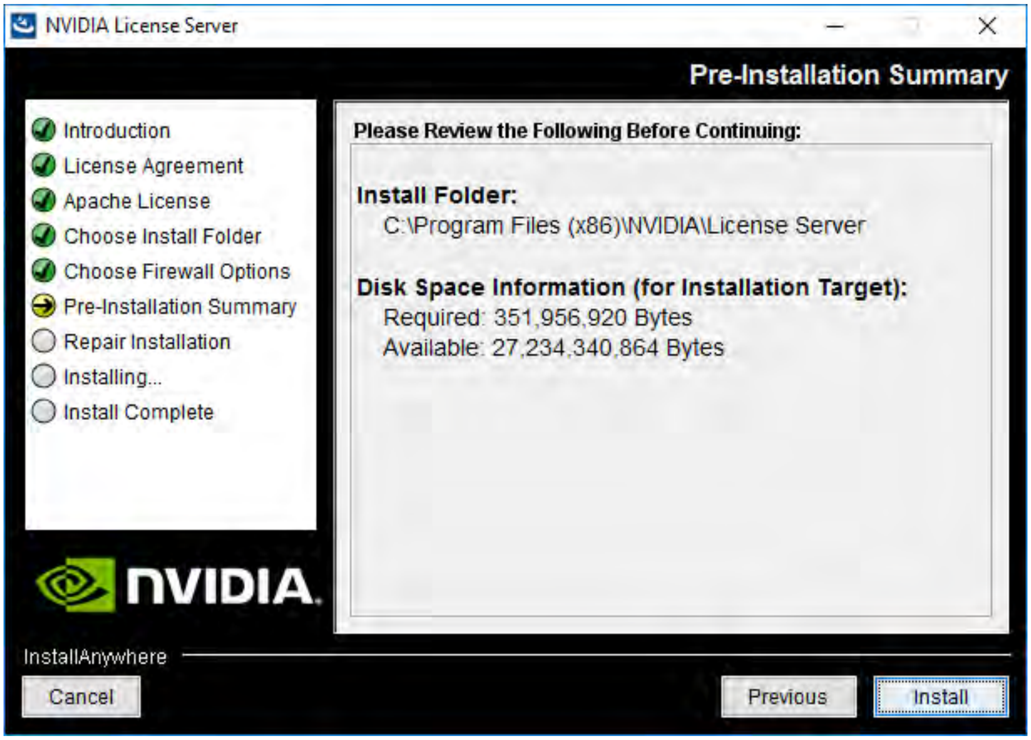
The license server's management interface listens on port 8080. Leave this port closed to prevent unauthorized access to the

☑ License server (port 7070)

☑ Management interface (port 8080)

9. On the Pre-installation Summary page, click Install (Figure 26). Installation will automatically progress without user input (Figure 27).

**Figure 26.** Pre-Installation Summary page

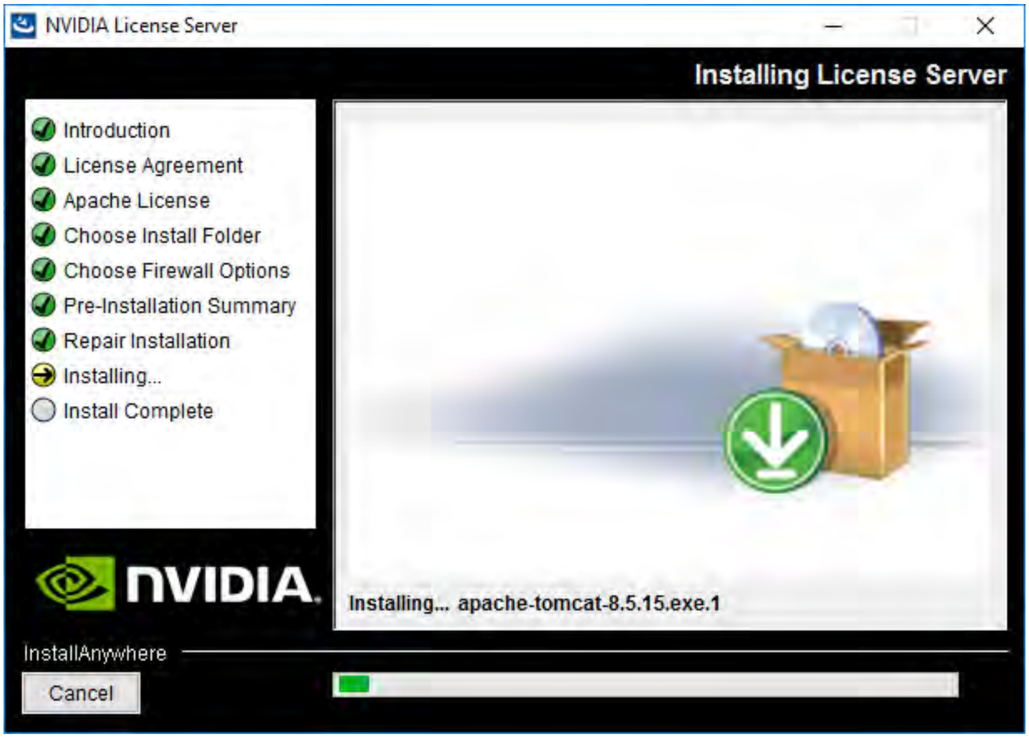


Please Review the Following Before Continuing:

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

Disk Space Information (for Installation Target):
Required: 351,956,920 Bytes
Available: 27,234,340,864 Bytes

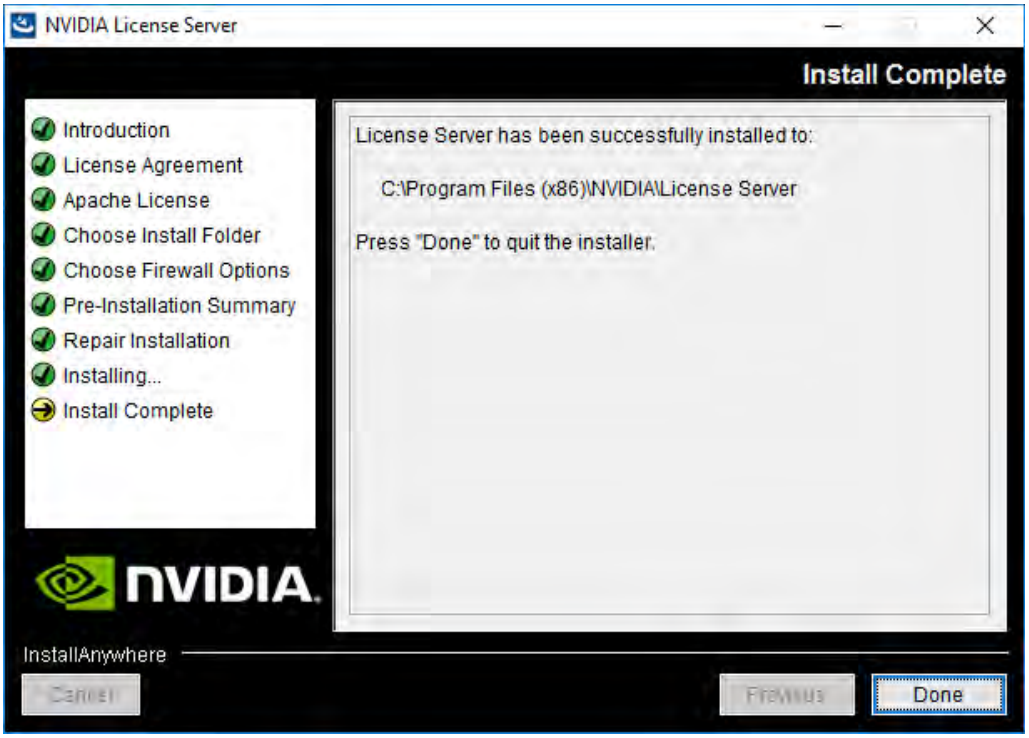
**Figure 27.** Installing the license server



Installing... apache-tomcat-8.5.15.exe.1

10. When the installation process is complete, click Done (Figure 28).

**Figure 28.** Installation Complete page



License Server has been successfully installed to:

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

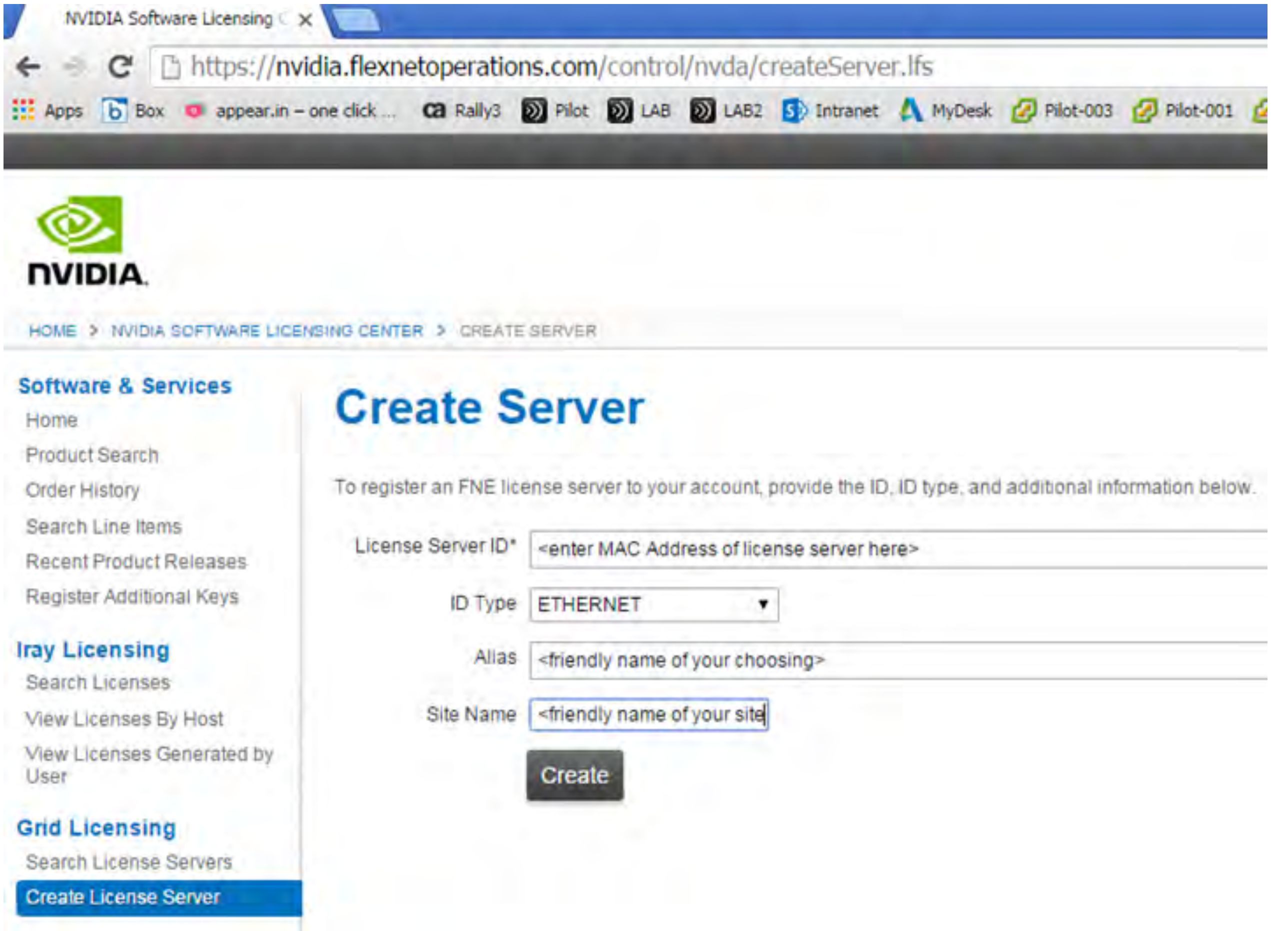
Press "Done" to quit the installer.

### 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. 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 29. 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 29.** Creating the 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*: <enter MAC Address of license server here>

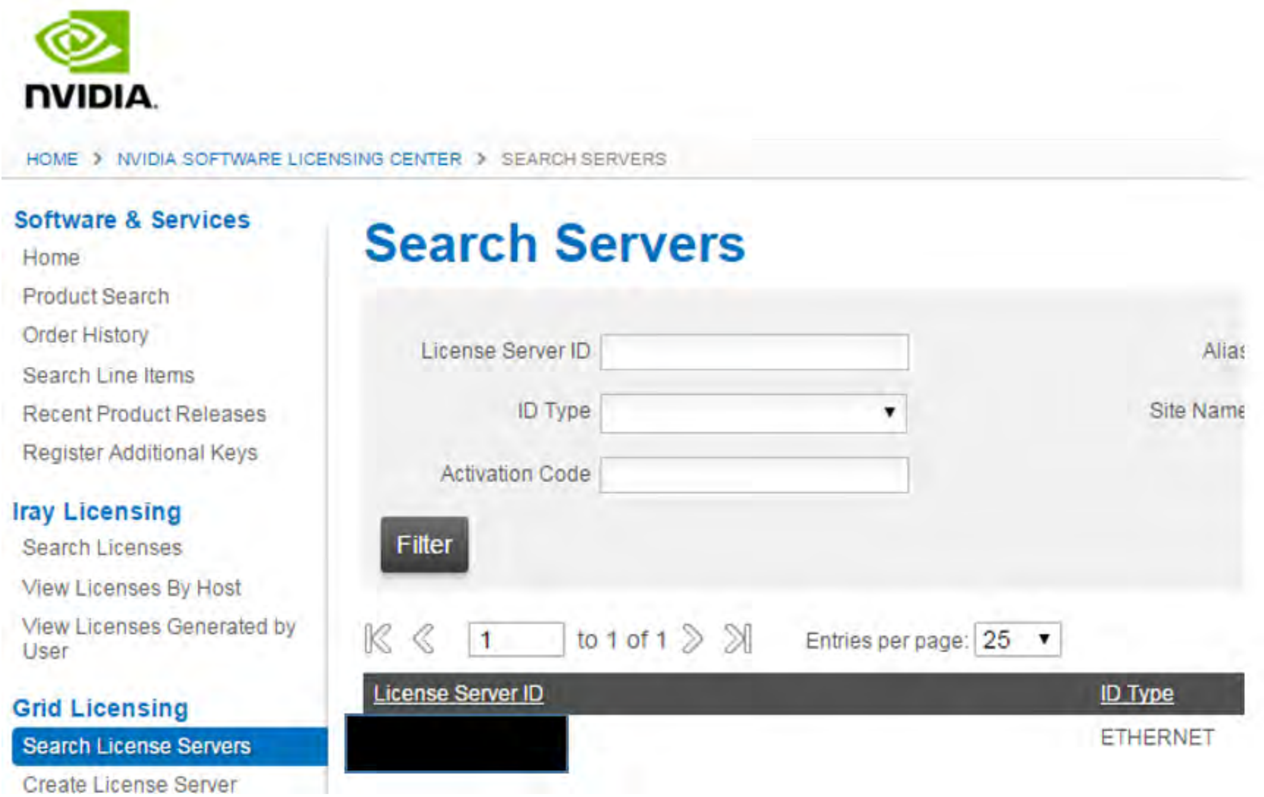
ID Type: ETHERNET

Alias: <friendly name of your choosing>

Site Name: <friendly name of your site

4. Click the Search License Servers node.
5. Click your license server ID (Figure 30).

**Figure 30.** Selecting the license server ID



NVIDIA.

HOME > NVIDIA SOFTWARE LICENSING CENTER > SEARCH SERVERS

**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

## Search Servers

License Server ID

ID Type ▼

Activation Code

Alias

Site Name

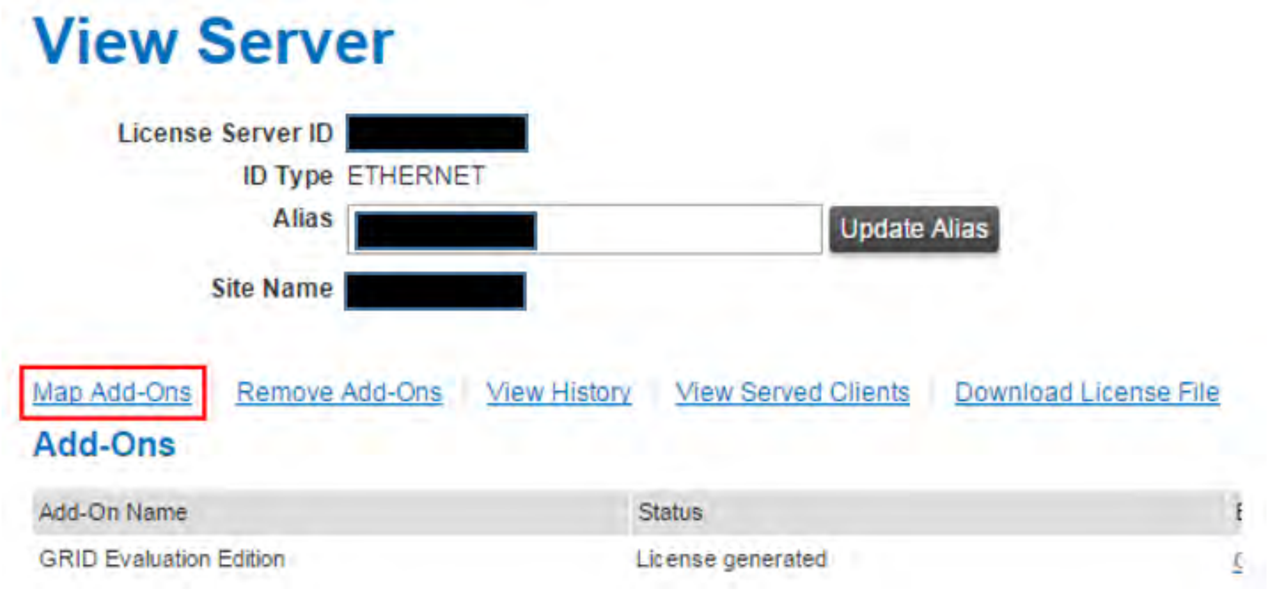
Filter

1 to 1 of 1 — Entries per page: 25 ▼

| License Server ID | ID Type |
|---|---|
| [redacted] | ETHERNET |

6. Click Map Add-Ons and choose the number of license units out of your total pool to allocate to this license server (Figure 31).

**Figure 31.** Choosing the number of license units from the pool



## View Server

License Server ID [redacted]

ID Type ETHERNET

Alias [redacted] Update Alias

Site Name [redacted]

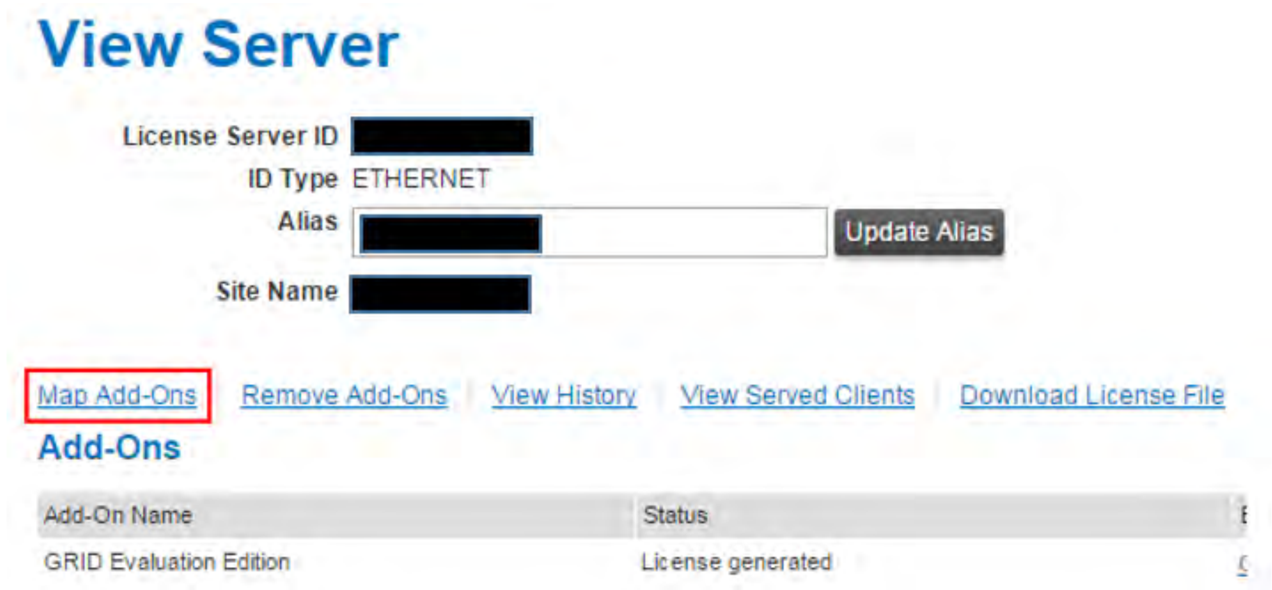
Map Add-Ons | Remove Add-Ons | View History | View Served Clients | Download License File

### Add-Ons

| Add-On Name | Status |
|---|---|
| GRID Evaluation Edition | License generated |

After the add-ons are mapped, the interface will look like Figure 32, showing 128 units mapped, for example.

**Figure 32.** View Server page after the add-ons are mapped



View Server

License Server ID [redacted]

ID Type ETHERNET

Alias [redacted]

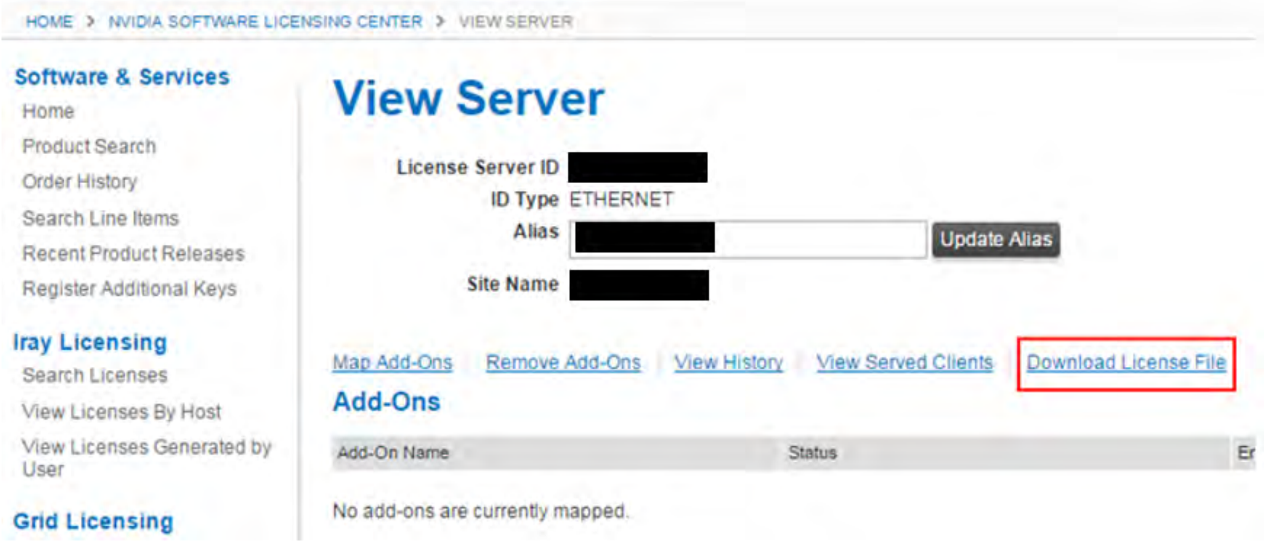
Site Name [redacted]

| Add-On Name | Status |
|---|---|
| GRID Evaluation Edition | License generated |

7. Click Download License File and save the .bin file to your license server (Figure 33).

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

**Figure 33.** Saving the .bin file



View Server

License Server ID [redacted]

ID Type ETHERNET

Alias [redacted]

Site Name [redacted]

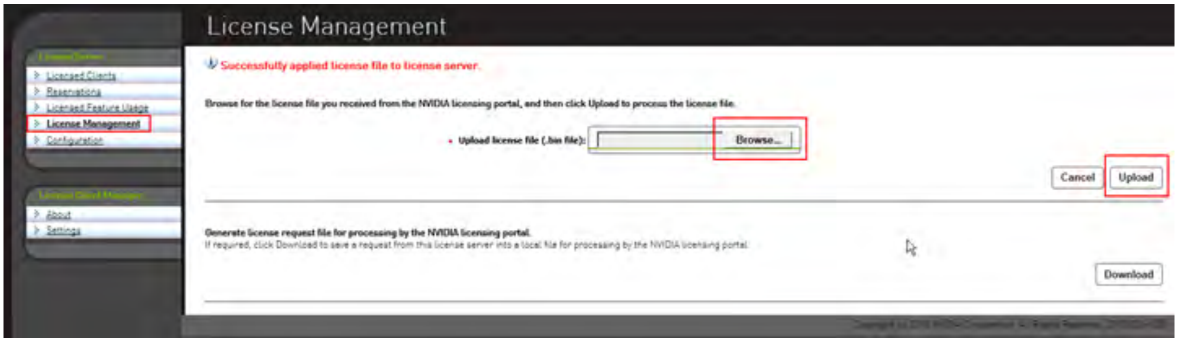
| Add-On Name | Status | Er |
|---|---|---|

No add-ons are currently mapped.

8. On the local license server, browse to http://<FQDN>:8080/licserver to display the License Server Configuration page.
9. Click License Management in the left pane.
10. 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 34). The features are available (Figure 35).

**Figure 34.** License file successfully applied



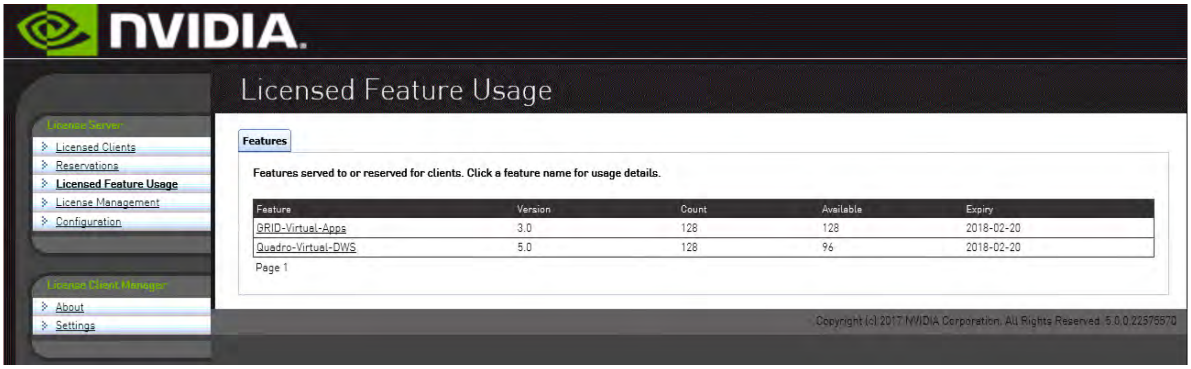
Successfully applied license file to license server.

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

- Upload license file (.bin file):

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

**Figure 35.** NVIDIA license server with features available for use



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

| Feature | Version | Count | Available | Expiry |
|---|---|---|---|---|
| GRID-Virtual-Apps | 3.0 | 128 | 128 | 2018-02-20 |
| Quadro-Virtual-DWS | 5.0 | 128 | 96 | 2018-02-20 |

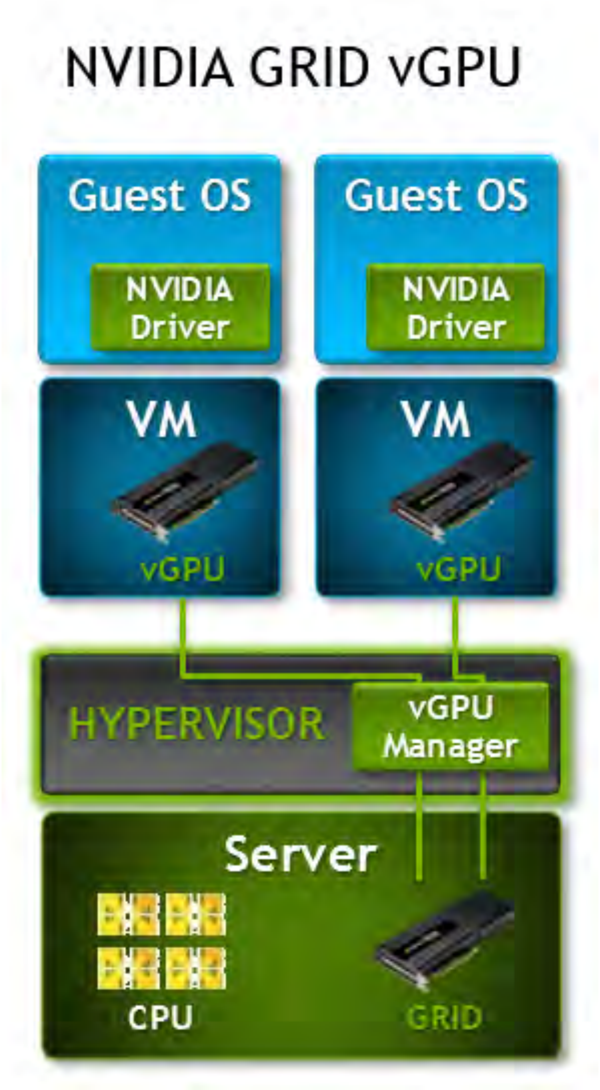
Page 1

Copyright (c) 2017 NVIDIA Corporation. All Rights Reserved. 5.0.0.22575570

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

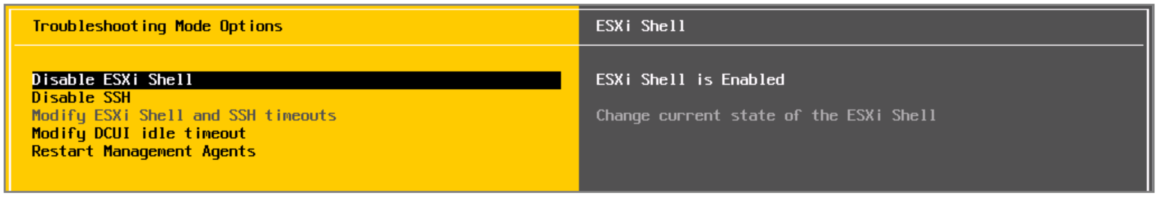
This section summarizes the installation process for configuring an ESXi host and virtual machine for vGPU support. Figure 36 shows the components used for vGPU support.

**Figure 36.** NVIDIA GRID vGPU components



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 37).

**Figure 37.** VMware ESXi configuration console



```
Troubleshooting Mode Options                ESXi Shell

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

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.

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

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

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

The command should return output similar to that shown here:

```
# esxcli software vib install --no-sig-check -v /tmp/NVIDIA-VMware_ESXi_6.5_Host_Driver_384.99-
1OEM.650.0.0.4598673.vib
Installation Result
   Message: Operation finished successfully.
   Reboot Required: false
   VIBs Installed: NVIDIA_bootbank_NVIDIA-VMware_ESXi_6.5_Host_Driver_384.99-
1OEM.650.0.0.4598673
   VIBs Removed:
   VIBs Skipped:
```

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

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

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

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

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

The command should return output similar to that shown here:

```
# esxcli software vib list | grep -i nvidia
NVIDIA-VMware_ESXi_6.5_Host_Driver  384.99-1OEM.650.0.0.4598673           NVIDIA
VMwareAccepted    2017-11-27
```

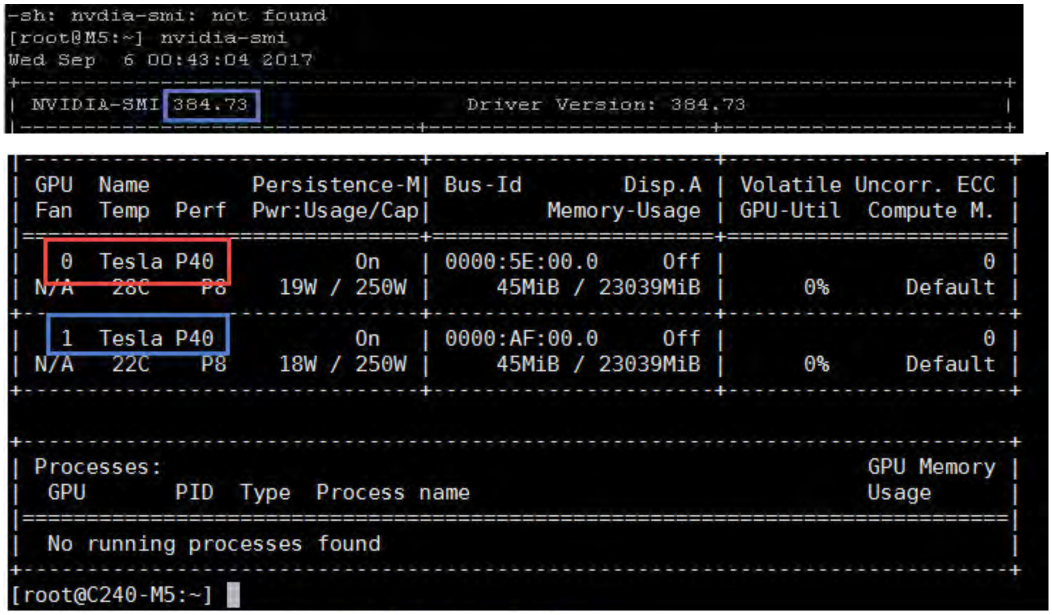
**Note:** See the VMware knowledge base article for information about removing any existing NVIDIA drivers before installing new drivers: http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2033434.

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

```
nvidia-smi
```

The command should return output similar to that shown in Figure 38, 39, or 40, depending on the card used in your environment.

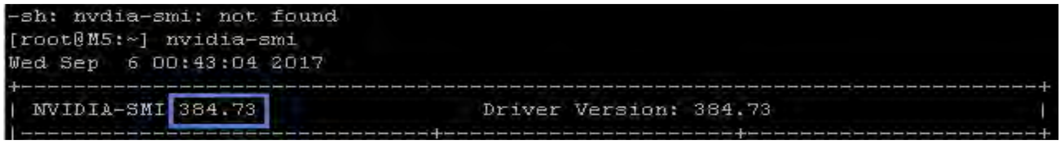
**Figure 38.** VMware ESX SSH console report for GPU P40 card detection on Cisco UCS C240 M5 Rack Server



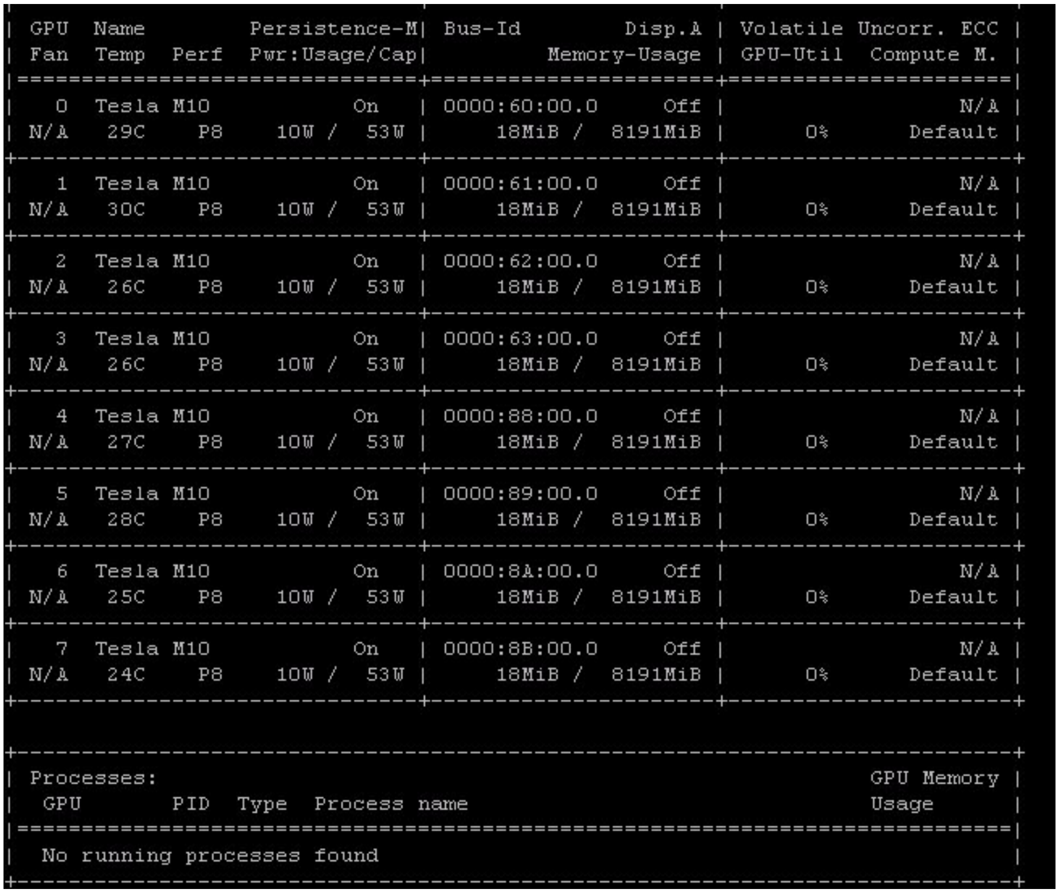
```
-sh: nvdia-smi: not found
[root@M5:~] nvidia-smi
Wed Sep  6 00:43:04 2017
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 384.73                 Driver Version: 384.73                    |
|-------------------------------+----------------------+----------------------+
```

```
|-------------------------------+----------------------+----------------------+
| 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 |
| N/A   28C    P8    19W / 250W |     45MiB / 23039MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   1  Tesla P40           On   | 0000:AF:00.0     Off |                    0 |
| N/A   22C    P8    18W / 250W |     45MiB / 23039MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
[root@C240-M5:~]
```

**Figure 39.** VMware ESX SSH console report for GPU M10 card detection on Cisco UCS C240 M5 Rack Server

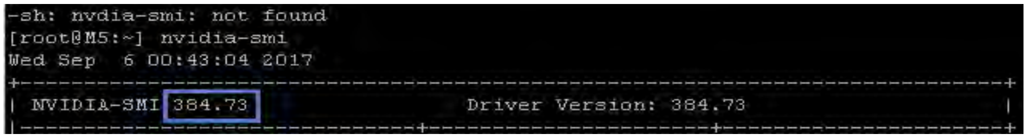


```
-sh: nvdia-smi: not found
[root@M5:~] nvidia-smi
Wed Sep  6 00:43:04 2017
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 384.73                 Driver Version: 384.73                    |
|-------------------------------+----------------------+----------------------+
```

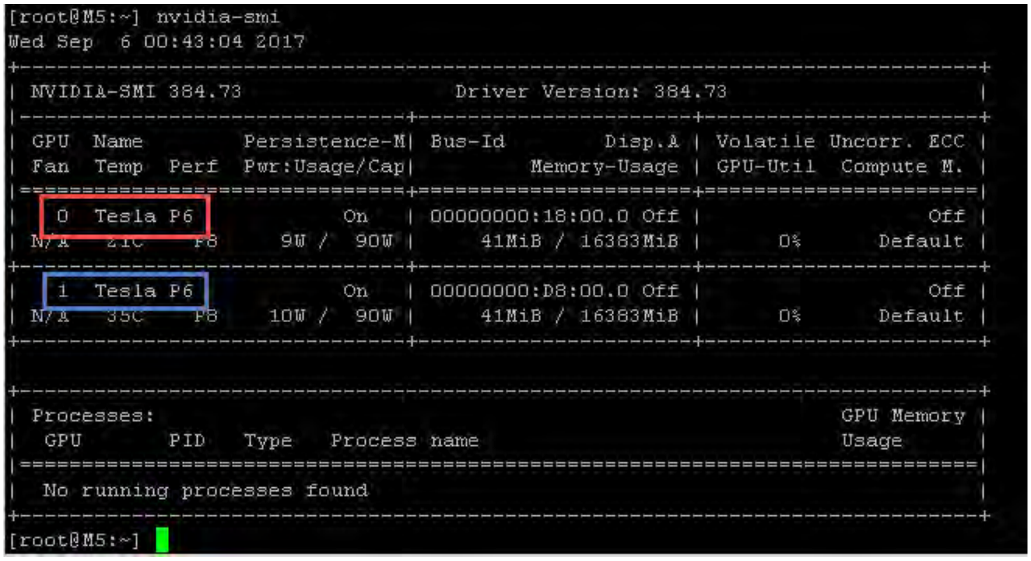


```
|                               |                      |                      |
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   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 40.** VMware ESX SSH console report for GPU P6 card detection on Cisco UCs B200 M5 Blade Server



```
-sh: nvdia-smi: not found
[root@M5:~] nvidia-smi
Wed Sep  6 00:43:04 2017
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 384.73                 Driver Version: 384.73                    |
|-------------------------------+----------------------+----------------------+
```



```
[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: nvidia-smi –l (this command adds a loop, automatically refreshing the display).

### NVIDIA Tesla P6, P40, and M10 profile specifications

The Tesla P6 and P40 cards each have a single physical GPU, and the Tesla M10 card has multiple physical GPUs. Each physical GPU can support several different types of vGPU. Each type of vGPU has a fixed amount of frame buffer space, a fixed number of supported display heads, and a fixed maximum resolution, and each is targeted at a different class of workload. Table 6 lists the vGPU types supported by GRID GPUs.

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

Table 6. NVIDIA GRID 5 user profile specifications for NVIDIA Tesla cards

| End-user profile | GRID<br>Virtual App (vApp) profiles | GRID<br>Virtual PC (vPC)profiles | Quadro<br>Virtual Datacenter Workstation (vDWS) profiles |
|---|---|---|---|
| 1 GB | • P6-1A<br>• M10-1A<br>• P40-1A | P6-1B<br>M10-1B<br>P40-1B | • P6-1Q<br>• M10-1Q<br>• P40-1Q |
| 2 GB | • P6-2A<br>• M10-2A<br>• P40-2A | – | • P6-2Q<br>• M10-2Q<br>• P40-2Q |
| 3 GB | P40-3A | – | P40-3Q |
| 4 GB | • P6-4A<br>• M10-4A<br>• P40-4A | – | P6-4Q<br>M10-4Q<br>P40-4Q |
| 6 GB | P40-6A | – | P40-6Q |
| 8 GB | • P6-8A<br>• M10-8A<br>• P40-8A | – | P6-8Q<br>M10-8Q<br>P40-8Q |

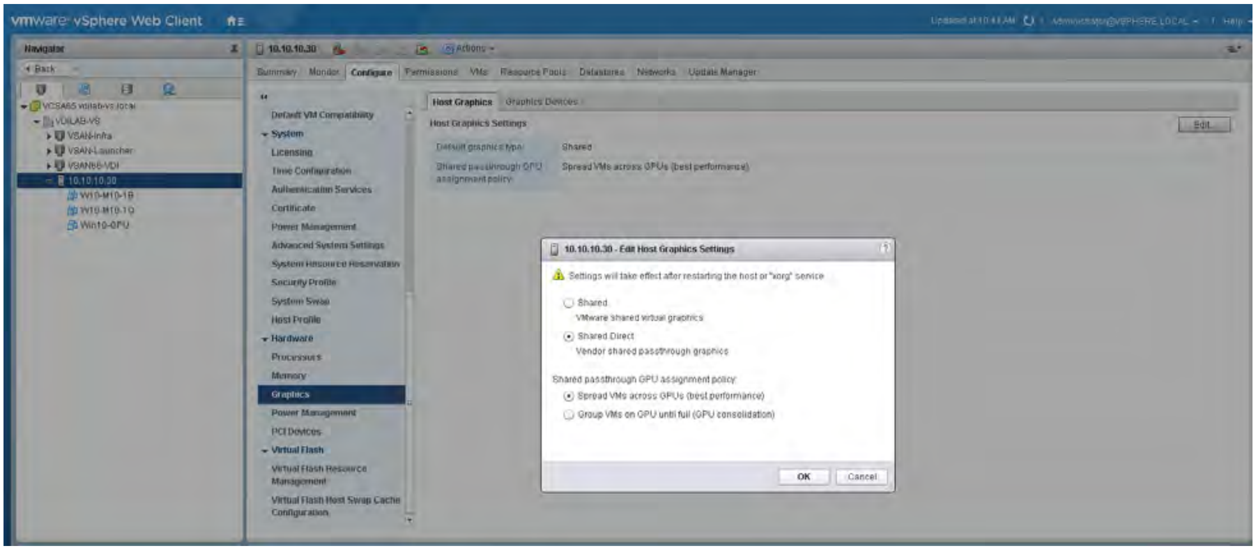
| End-user profile | GRID<br>Virtual App (vApp) profiles | GRID<br>Virtual PC (vPC)profiles | Quadro<br>Virtual Datacenter Workstation (vDWS) profiles |
|---|---|---|---|
| 12 GB | P40-12A | – | P40-12Q |
| 16 GB | P6-16A | – | P6-16Q |
| 24 GB | P40-24A | – | P40-24Q |
| Pass-through | – | – | – |

### Prepare a virtual machine for vGPU support

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

1. Select the ESXi host and click the Configure tab. From the list of options at the left, choose Graphics > Edit Host Graphics Settings. Select Shared Direct "Vendor shared passthrough graphics" (Figure 41). Reboot the system to make the changes effective.

**Figure 41.** Edit Host Graphics Settings window



Host Graphics Settings

Default graphics type: Shared

Shared passthrough GPU assignment policy: Spread VMs across GPUs (best performance)

10.10.10.30 - Edit Host Graphics Settings

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

○ Shared
VMware shared virtual graphics

● Shared Direct
Vendor shared passthrough graphics

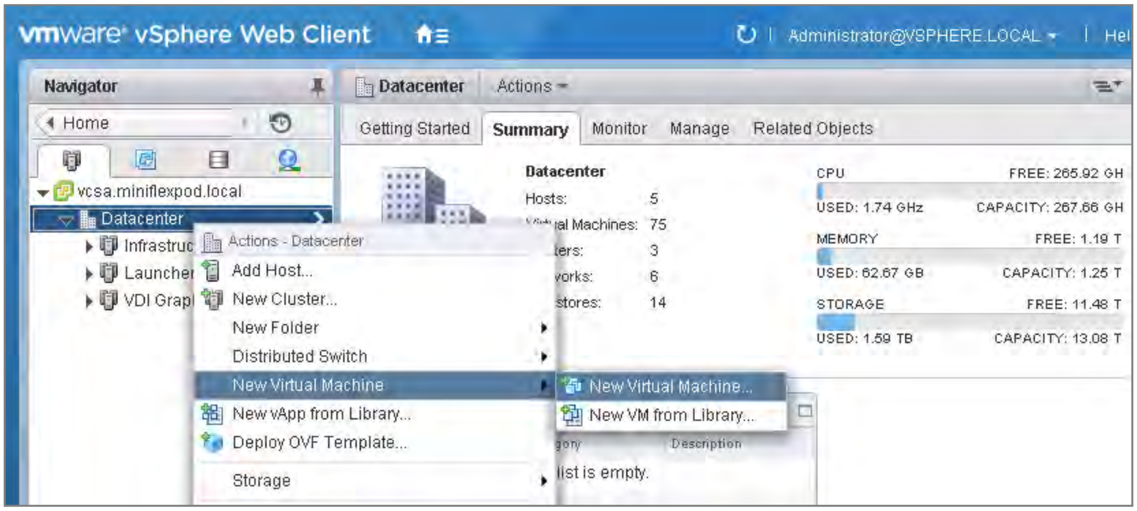
Shared passthrough GPU assignment policy:

● Spread VMs across GPUs (best performance)

○ Group VMs on GPU until full (GPU consolidation)

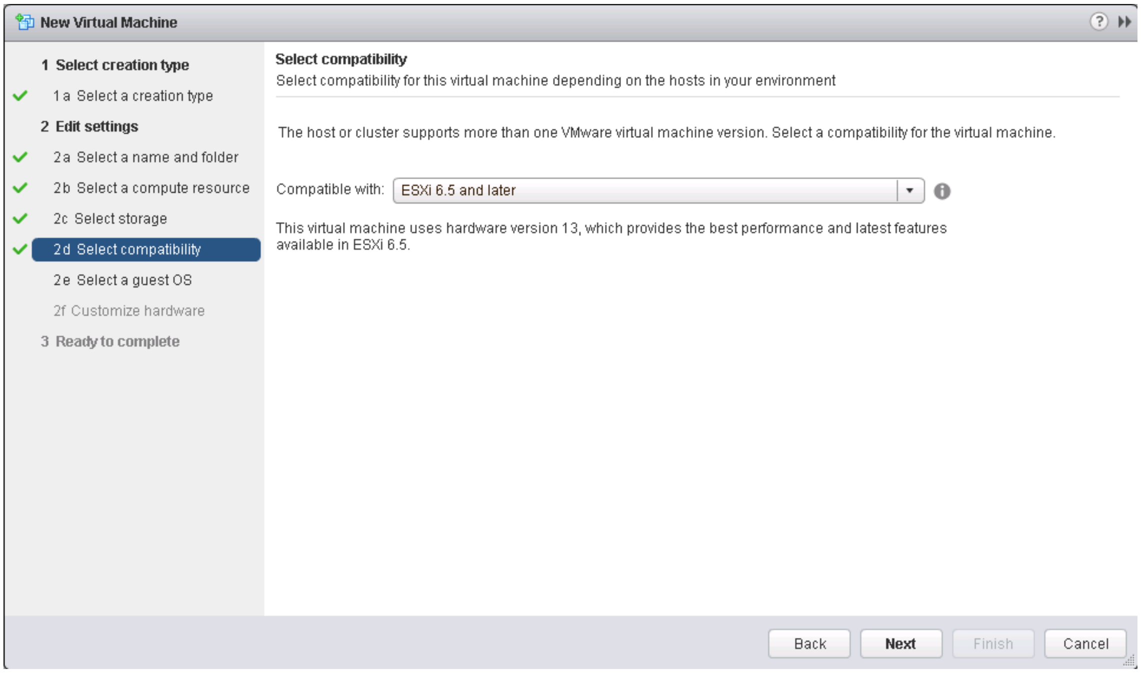
2. Using the vSphere Web Client, create a new virtual machine. To do this, right-click a host or cluster and choose New Virtual Machine. Work through the New Virtual Machine wizard. Unless another configuration is specified, select the configuration settings appropriate for your environment (Figure 42).

**Figure 42.** Creating a new virtual machine in VMware vSphere Web Client



3. Choose "ESXi 6.0 and later" from the "Compatible with" drop-down menu to use the latest features, including the mapping of shared PCI devices, which is required for the vGPU feature (Figure 43). 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 43.** Selecting virtual machine hardware Version 11 or later



**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.

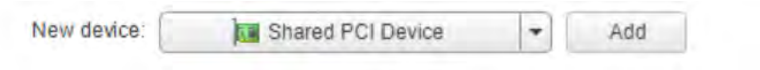
Compatible with: ESXi 6.5 and later

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

4. To customize 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 44 and 45).

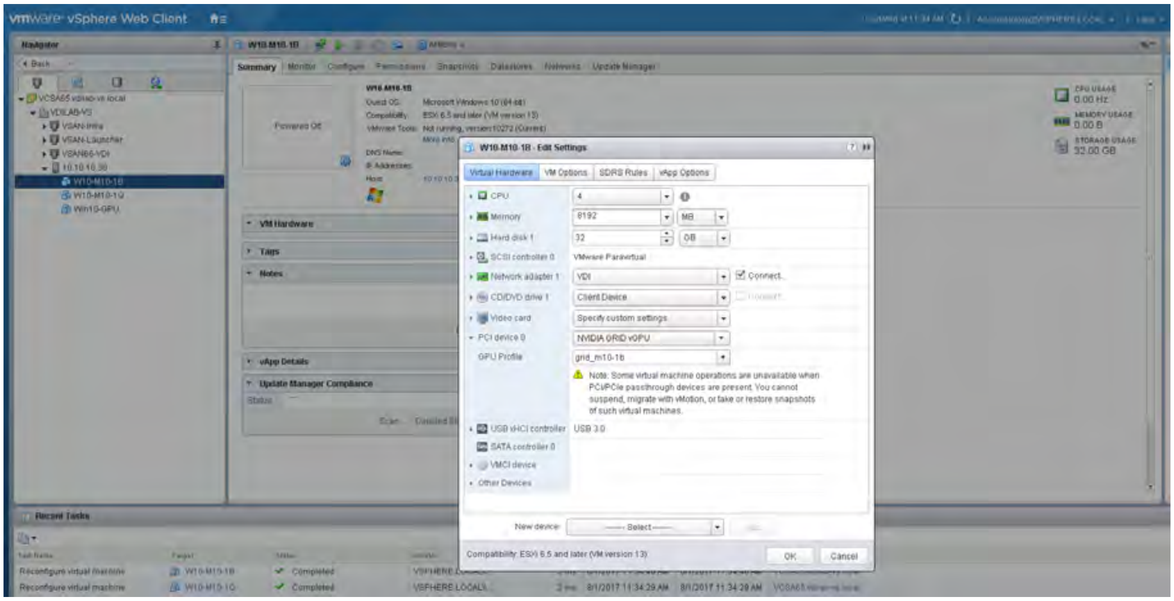
**Note:** If you are creating a new virtual machine and using the vSphere Web Client's virtual machine console functions, the mouse will not be usable in the virtual machine until after both the operating system and VMware Tools have been installed. If you cannot use the traditional vSphere Web Client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

**Figure 44.** Adding a shared PCI device to the virtual machine to attach the GPU profile



New device: Shared PCI Device

**Figure 45.** Attaching the GPU profile to a shared PCI device



W10-M10-1B - Edit Settings

CPU: 4
Memory: 8192 MB
Hard disk 1: 32 GB
SCSI controller 0: VMware Paravirtual
Network adapter 1: VDI ☑ Connect...
CD/DVD drive 1: Client Device ☐ Connect...
Video card: Specify custom settings
PCI device 0: NVIDIA GRID vGPU
GPU Profile: grid_m10-1b

Note: Some virtual machine operations are unavailable when PCI/PCIe passthrough devices are present. You cannot suspend, migrate with vMotion, or take or restore snapshots of such virtual machines.

USB xHCI controller: USB 3.0
SATA controller 0
VMCI device
Other Devices

New device: ------- Select -------

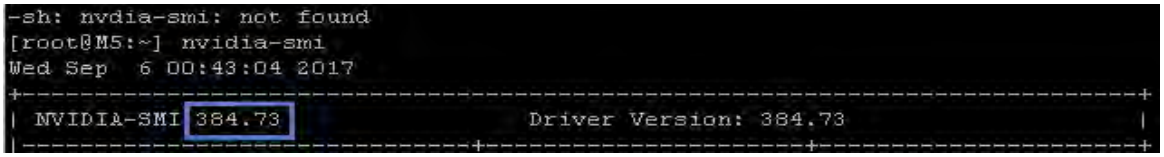
Compatibility: ESXi 6.5 and later (VM version 13)

5. A virtual machine with a vGPU assigned will not start if ECC is enabled. If this is the case, as a workaround disable ECC by entering the following commands (Figure 46):

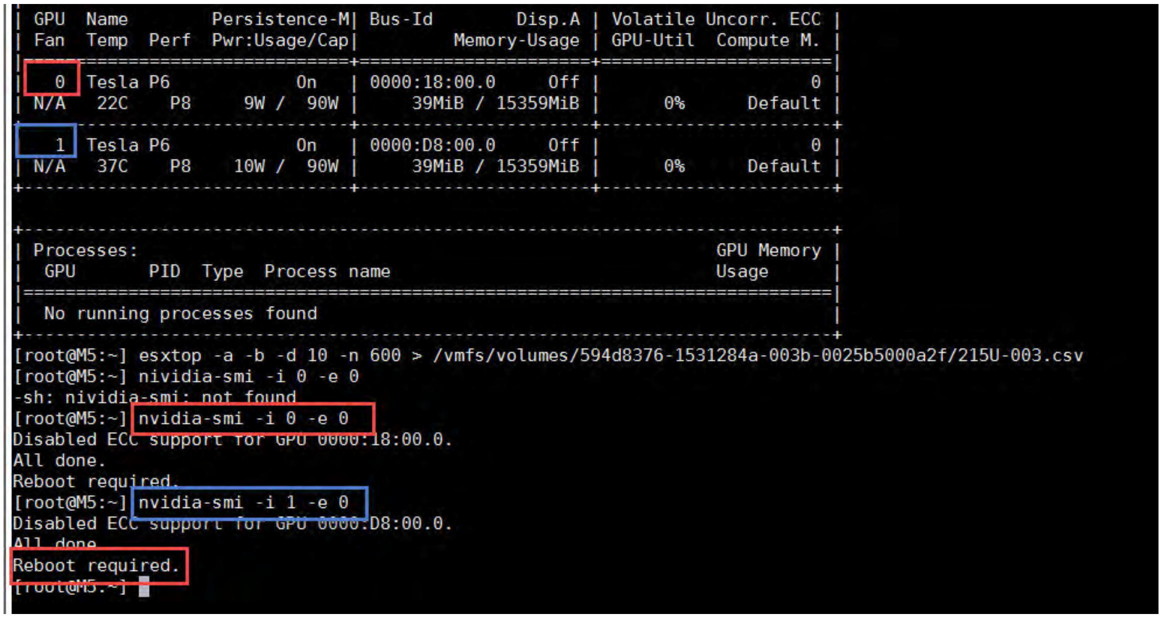
```
# nvidia-smi -i 0 -e 0
# nvidia-smi -i 1 -e 0
```

**Note:** Use -i to target a specific GPU. If two cards are installed in a server, run the command twice as shown in the example here, where 0 and 1 each specify a GPU card.

**Figure 46.** Disabling ECC



```
-sh: nvdia-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:~] nividia-smi -i 0 -e 0
-sh: nividia-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:~]
```

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

   a. Configure the virtual machine with the appropriate amount of vCPU and RAM according to the GPU profile selected.

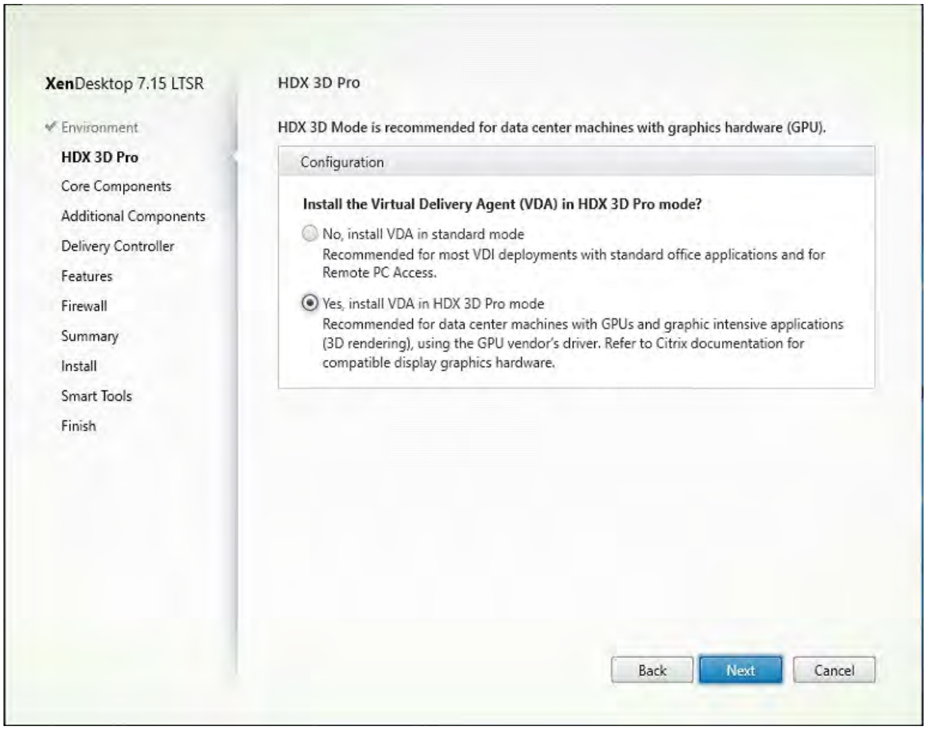
   b. Install VMware Tools.

   c. Join the virtual machine to the Microsoft Active Directory domain.

   d. Install or upgrade Citrix HDX 3D Pro Virtual Desktop Agent.

   - When you use the installer's GUI to install a VDA for a Windows desktop, select Yes on the HDX 3D Pro page (Figure 47).

**Figure 47.** Selecting HDX 3D Pro during VDA installation



HDX 3D Pro

HDX 3D Mode is recommended for data center machines with graphics hardware (GPU).

Configuration

**Install the Virtual Delivery Agent (VDA) in HDX 3D Pro mode?**

○ No, install VDA in standard mode
Recommended for most VDI deployments with standard office applications and for Remote PC Access.

● Yes, install VDA in HDX 3D Pro mode
Recommended for data center machines with GPUs and graphic intensive applications (3D rendering), using the GPU vendor's driver. Refer to Citrix documentation for compatible display graphics hardware.

When you use the command-line interface (CLI) to install the VDA, 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.

**Note:** VMware OSOT b1090 with the Windows 10 – LoginVSI template was used for the purposes of this document.

Table 7 shows differences in applied optimizations in the master image used for vGPU-enabled desktops.

Table 7. Optimization differences

| Optimization | Description | No GPU | vGPU |
|---|---|---|---|
| Software Rendering Internet Explorer | Uses software rendering instead of hardware rendering. Enable this option if no vGPU is present. | Applied | Not applied |
| Disable Hardware Acceleration Office 14 | Uses software rendering instead of hardware rendering. Enable this option if no vGPU is present. | Applied | Not applied |
| Disable Hardware Acceleration Office 15 | Uses software rendering instead of hardware rendering. Enable this option if no vGPU is present. | Applied | Not applied |

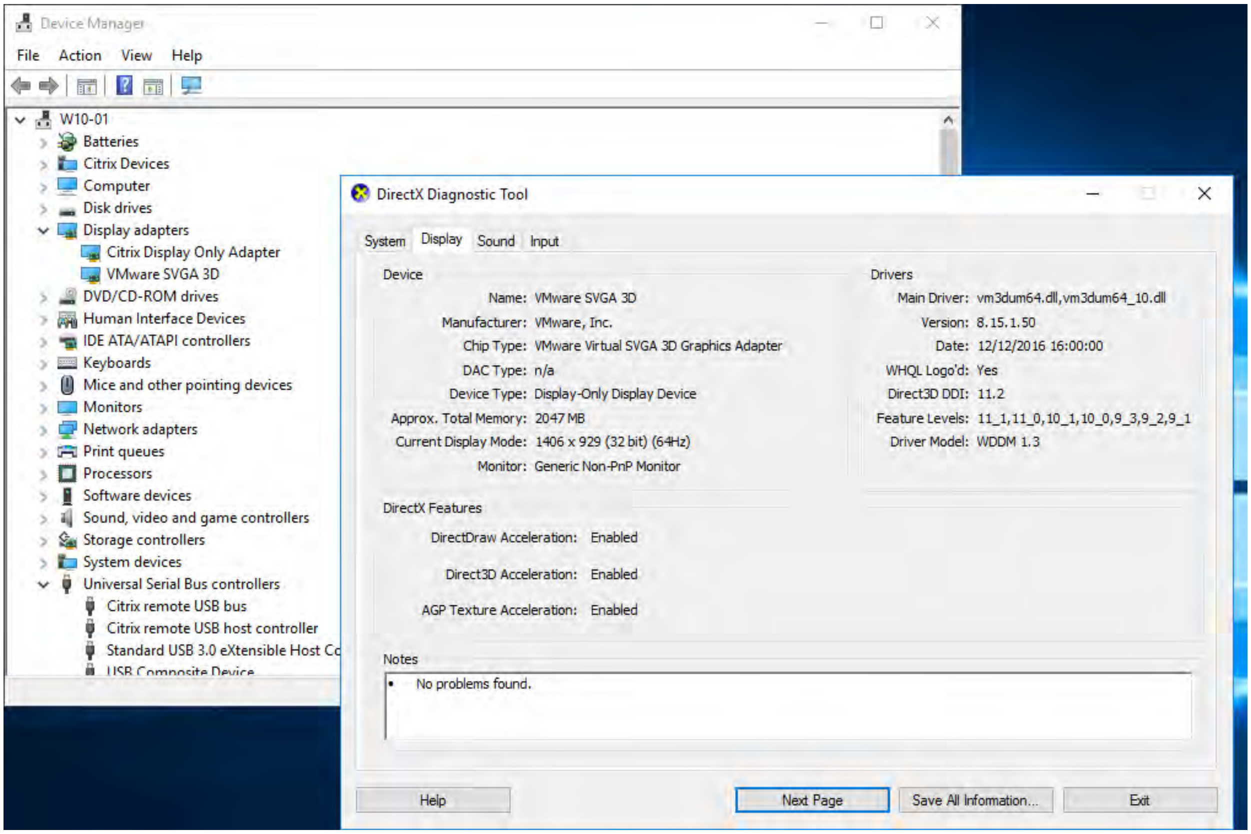
| Optimization | Description | No GPU | vGPU |
|---|---|---|---|
| Disable Animations Office 15 | Uses software rendering instead of hardware rendering. Enable this option if no vGPU is present. | Applied | Not applied |
| Disable Hardware Acceleration Office 16 | Uses software rendering instead of hardware rendering. Enable this option if no vGPU is present. | Applied | Not applied |
| Disable Animations Office 16 | Uses software rendering instead of hardware rendering. Enable this option if no vGPU is present. | Applied | Not applied |

### Install the NVIDIA vGPU software driver

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

Before the NVIDIA driver is installed on the guest virtual machine, the Device Manager shows the standard VGA graphics adapter (Figure 48).

**Figure 48.** Device Manager before the NVIDIA driver is installed



W10-01
- Batteries
- Citrix Devices
- Computer
- Disk drives
- Display adapters
  - Citrix Display Only Adapter
  - VMware SVGA 3D
- DVD/CD-ROM drives
- Human Interface Devices
- IDE ATA/ATAPI controllers
- Keyboards
- Mice and other pointing devices
- Monitors
- Network adapters
- Print queues
- Processors
- Software devices
- Sound, video and game controllers
- Storage controllers
- System devices
- Universal Serial Bus controllers
  - Citrix remote USB bus
  - Citrix remote USB host controller
  - Standard USB 3.0 eXtensible Host Co
  - USB Composite Device

Device

Name: VMware SVGA 3D
Manufacturer: VMware, Inc.
Chip Type: VMware Virtual SVGA 3D Graphics Adapter
DAC Type: n/a
Device Type: Display-Only Display Device
Approx. Total Memory: 2047 MB
Current Display Mode: 1406 x 929 (32 bit) (64Hz)
Monitor: Generic Non-PnP Monitor

Drivers

Main Driver: vm3dum64.dll,vm3dum64_10.dll
Version: 8.15.1.50
Date: 12/12/2016 16:00:00
WHQL Logo'd: Yes
Direct3D DDI: 11.2
Feature Levels: 11_1,11_0,10_1,10_0,9_3,9_2,9_1
Driver Model: WDDM 1.3

DirectX Features

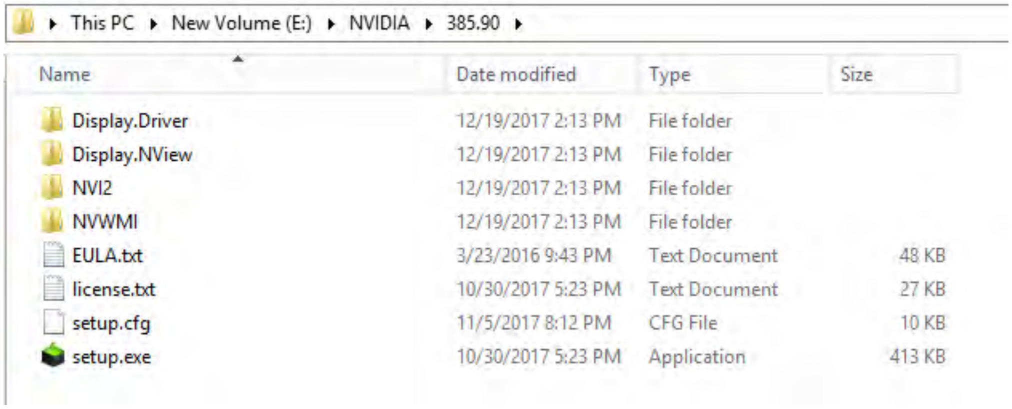
DirectDraw Acceleration: Enabled
Direct3D Acceleration: Enabled
AGP Texture Acceleration: Enabled

Notes

- No problems found.

1. Copy the Microsoft 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 49).

**Figure 49.** NVIDIA driver pack

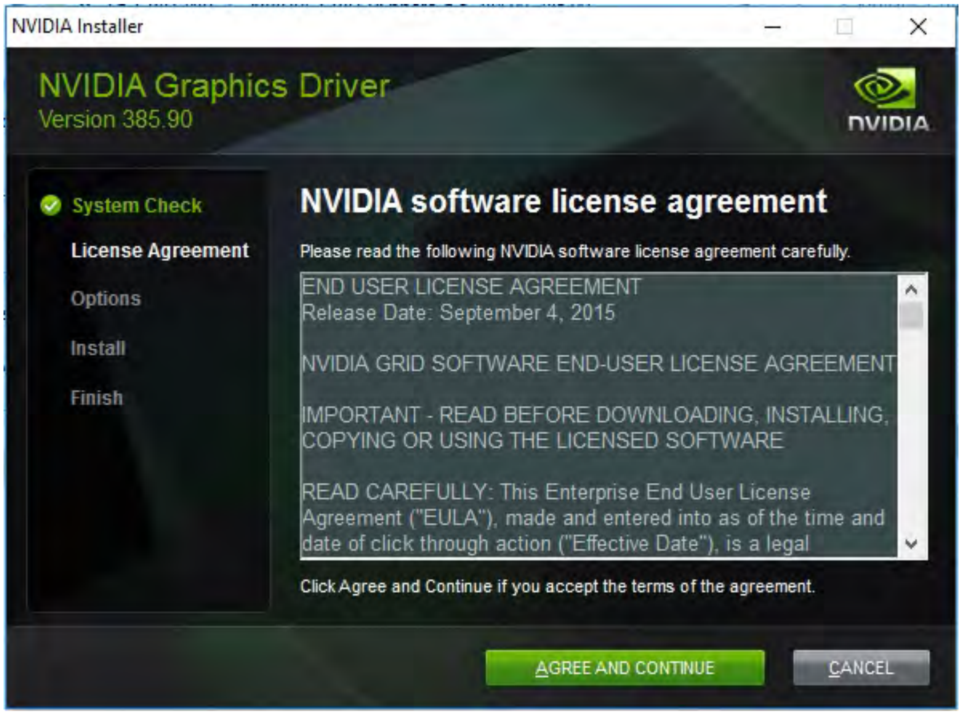


| Name | Date modified | Type | Size |
|---|---|---|---|
| Display.Driver | 12/19/2017 2:13 PM | File folder | |
| Display.NView | 12/19/2017 2:13 PM | File folder | |
| NVI2 | 12/19/2017 2:13 PM | File folder | |
| NVWMI | 12/19/2017 2:13 PM | File folder | |
| EULA.txt | 3/23/2016 9:43 PM | Text Document | 48 KB |
| license.txt | 10/30/2017 5:23 PM | Text Document | 27 KB |
| setup.cfg | 11/5/2017 8:12 PM | CFG File | 10 KB |
| setup.exe | 10/30/2017 5:23 PM | Application | 413 KB |

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

3. Agree to the NVIDIA software license (Figure 50).

**Figure 50.** Agreeing to the NVIDIA software license



NVIDIA Graphics Driver
Version 385.90

NVIDIA software license agreement

Please read the following NVIDIA software license agreement carefully.

END USER LICENSE AGREEMENT
Release Date: September 4, 2015

NVIDIA GRID SOFTWARE END-USER LICENSE AGREEMENT

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

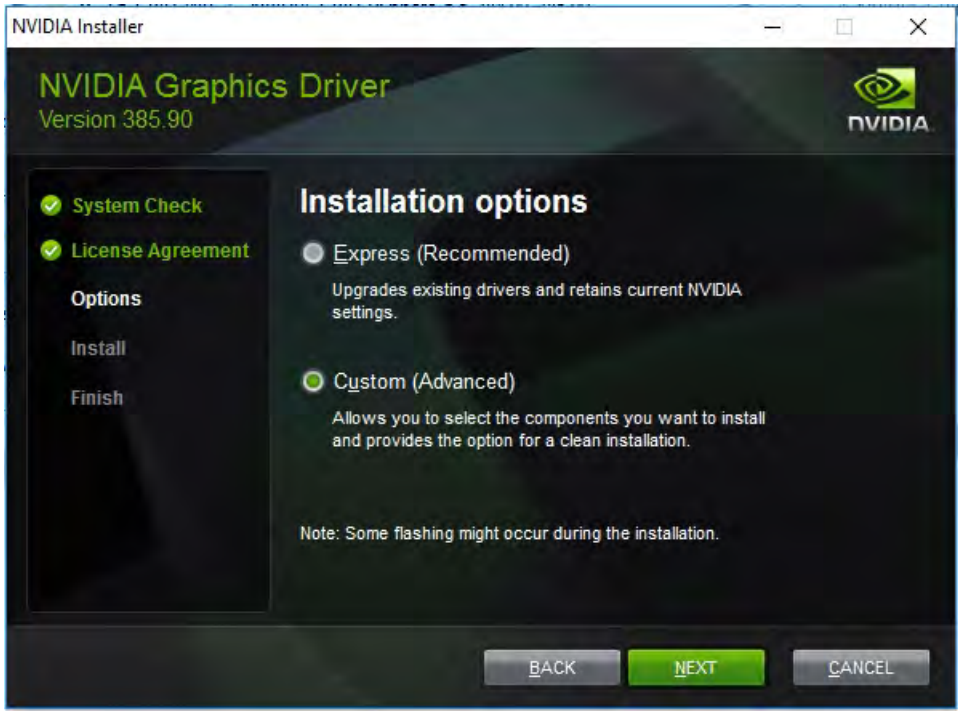
READ CAREFULLY: This Enterprise End User License Agreement ("EULA"), made and entered into as of the time and date of click through action ("Effective Date"), is a legal

Click Agree and Continue if you accept the terms of the agreement.

4. Install the graphics drivers using the Express or Custom option (Figures 51 and 52). After the installation has completed successfully, restart the virtual machine (Figure 53).

**Note:** Be sure that remote desktop connections are enabled. After this step, console access may not be available for the virtual machine when you connect from a vSphere Client.

**Figure 51.** Selecting the Express or Custom installation option



Installation options

○ Express (Recommended)

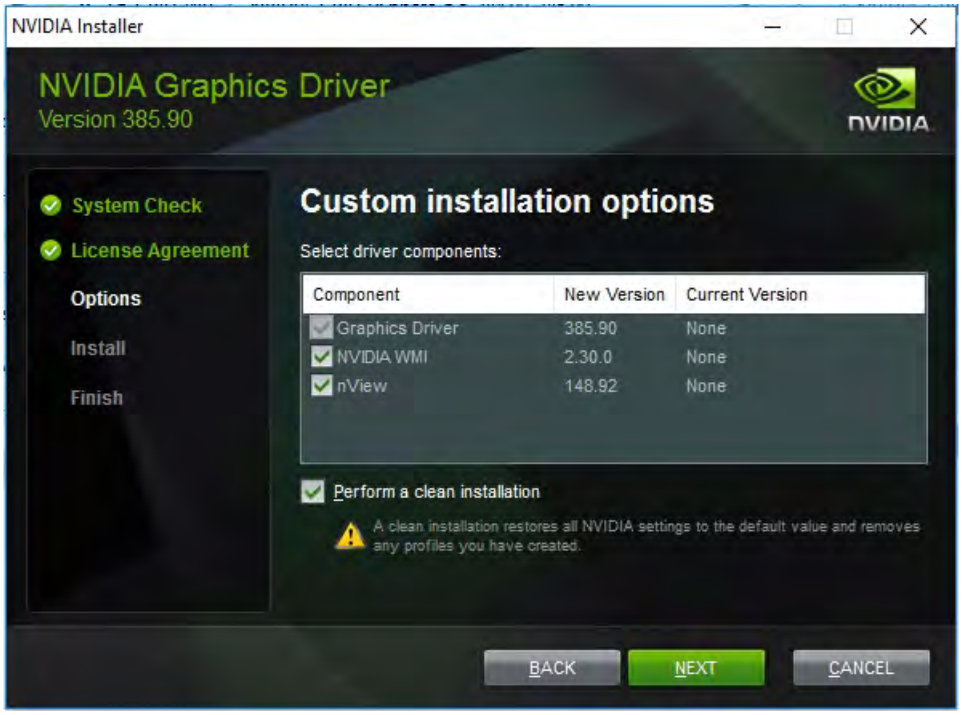
Upgrades existing drivers and retains current NVIDIA settings.

● Custom (Advanced)

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

Note: Some flashing might occur during the installation.

**Figure 52.** Components to be installed during NVIDIA graphics driver custom installation process



Custom installation options

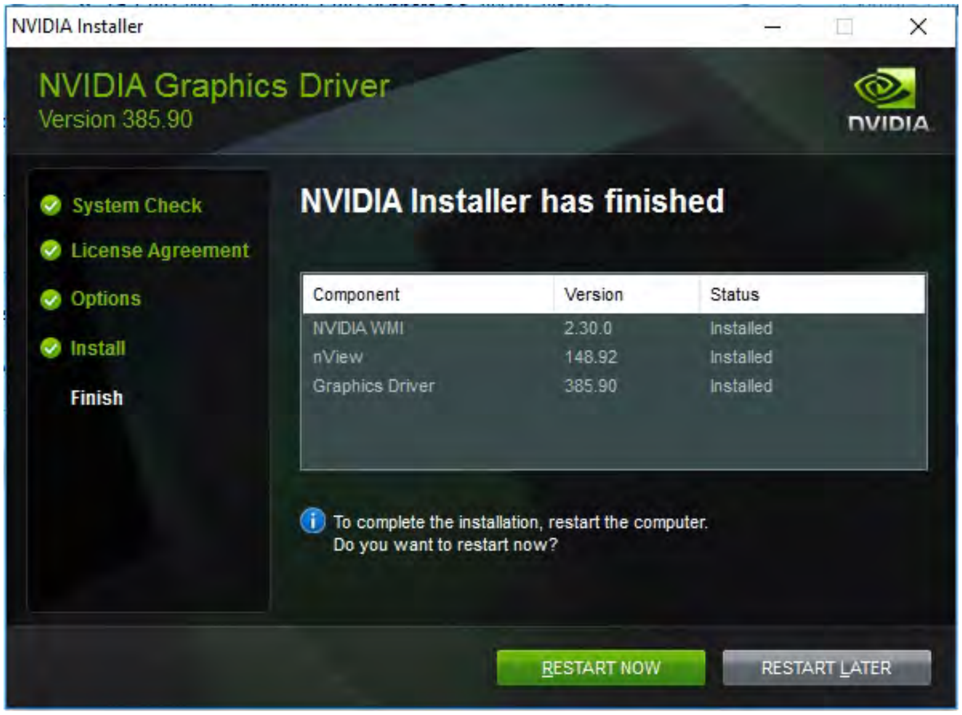
Select driver components:

| Component | New Version | Current Version |
|---|---|---|
| ☑ Graphics Driver | 385.90 | None |
| ☑ NVIDIA WMI | 2.30.0 | None |
| ☑ nView | 148.92 | None |

☑ Perform a clean installation

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

**Figure 53.** Resarting the virtual machine



NVIDIA Installer

NVIDIA Graphics Driver
Version 385.90

NVIDIA Installer has finished

| Component | Version | Status |
|---|---|---|
| NVIDIA WMI | 2.30.0 | Installed |
| nView | 148.92 | Installed |
| Graphics Driver | 385.90 | Installed |

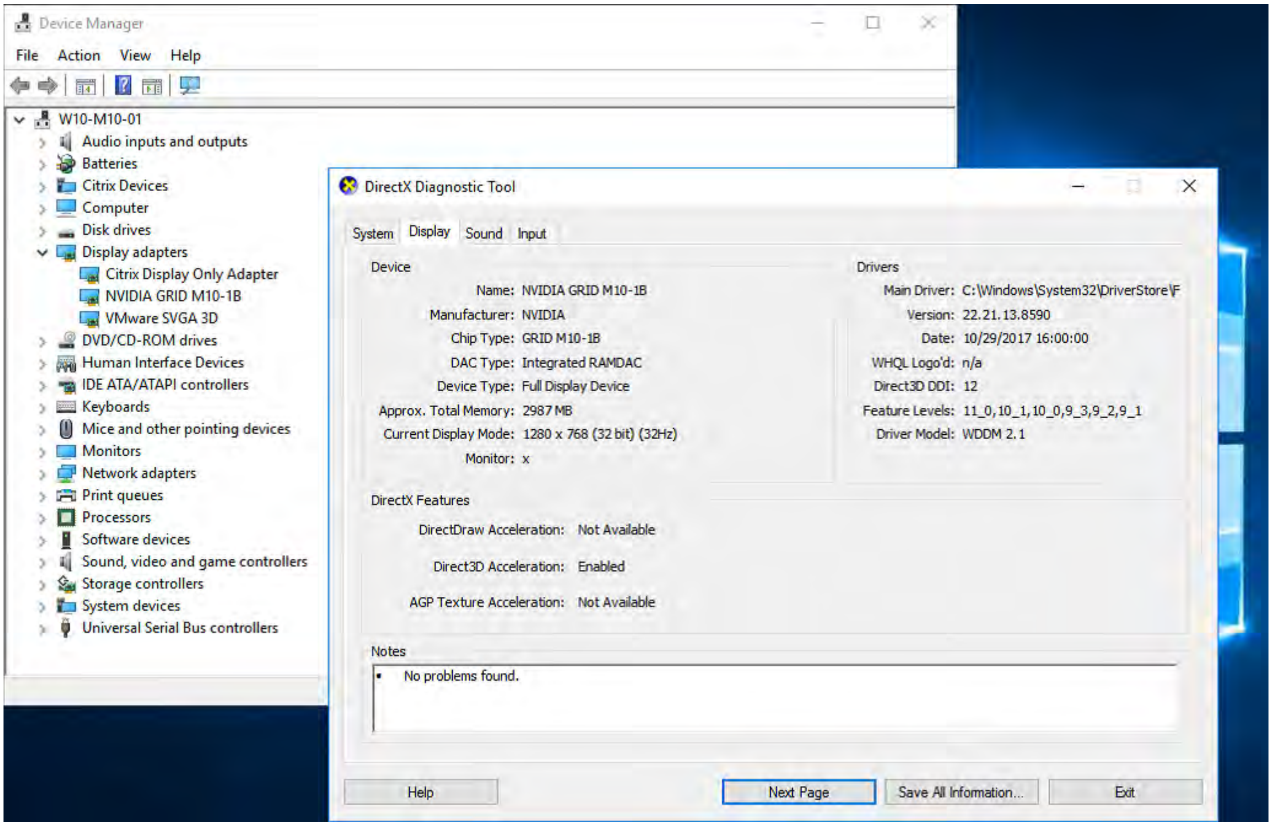
To complete the installation, restart the computer.
Do you want to restart now?

## Verify that applications are ready to support the vGPU

Verify 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 your chosen profile (Figure 54).

**Figure 54.** Verifying the driver installation



W10-M10-01
- Audio inputs and outputs
- Batteries
- Citrix Devices
- Computer
- Disk drives
- Display adapters
  - Citrix Display Only Adapter
  - NVIDIA GRID M10-1B
  - VMware SVGA 3D
- DVD/CD-ROM drives
- Human Interface Devices
- IDE ATA/ATAPI controllers
- Keyboards
- Mice and other pointing devices
- Monitors
- Network adapters
- Print queues
- Processors
- Software devices
- Sound, video and game controllers
- Storage controllers
- System devices
- Universal Serial Bus controllers

Device

Name: NVIDIA GRID M10-1B
Manufacturer: NVIDIA
Chip Type: GRID M10-1B
DAC Type: Integrated RAMDAC
Device Type: Full Display Device
Approx. Total Memory: 2987 MB
Current Display Mode: 1280 x 768 (32 bit) (32Hz)
Monitor: x

Drivers

Main Driver: C:\Windows\System32\DriverStore\F
Version: 22.21.13.8590
Date: 10/29/2017 16:00:00
WHQL Logo'd: n/a
Direct3D DDI: 12
Feature Levels: 11_0,10_1,10_0,9_3,9_2,9_1
Driver Model: WDDM 2.1

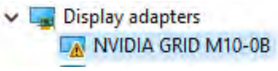
DirectX Features

DirectDraw Acceleration: Not Available
Direct3D Acceleration: Enabled
AGP Texture Acceleration: Not Available

Notes

- No problems found.

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



- Display adapters
  - NVIDIA GRID M10-0B

The following are the most likely the reasons:

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

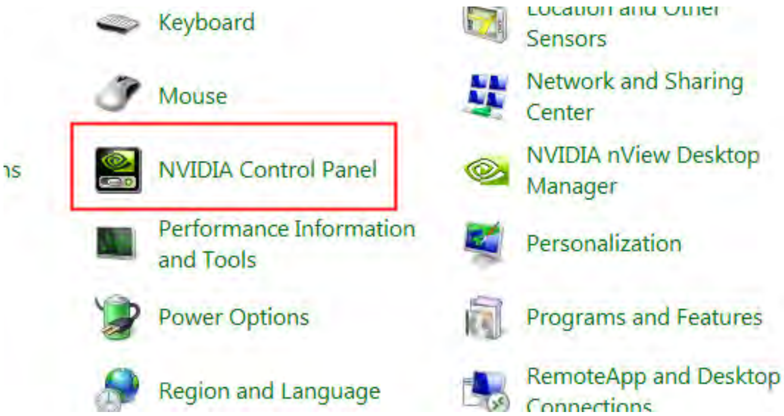
## Configure the virtual machine for an NVIDIA GRID vGPU license

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 registry keys.

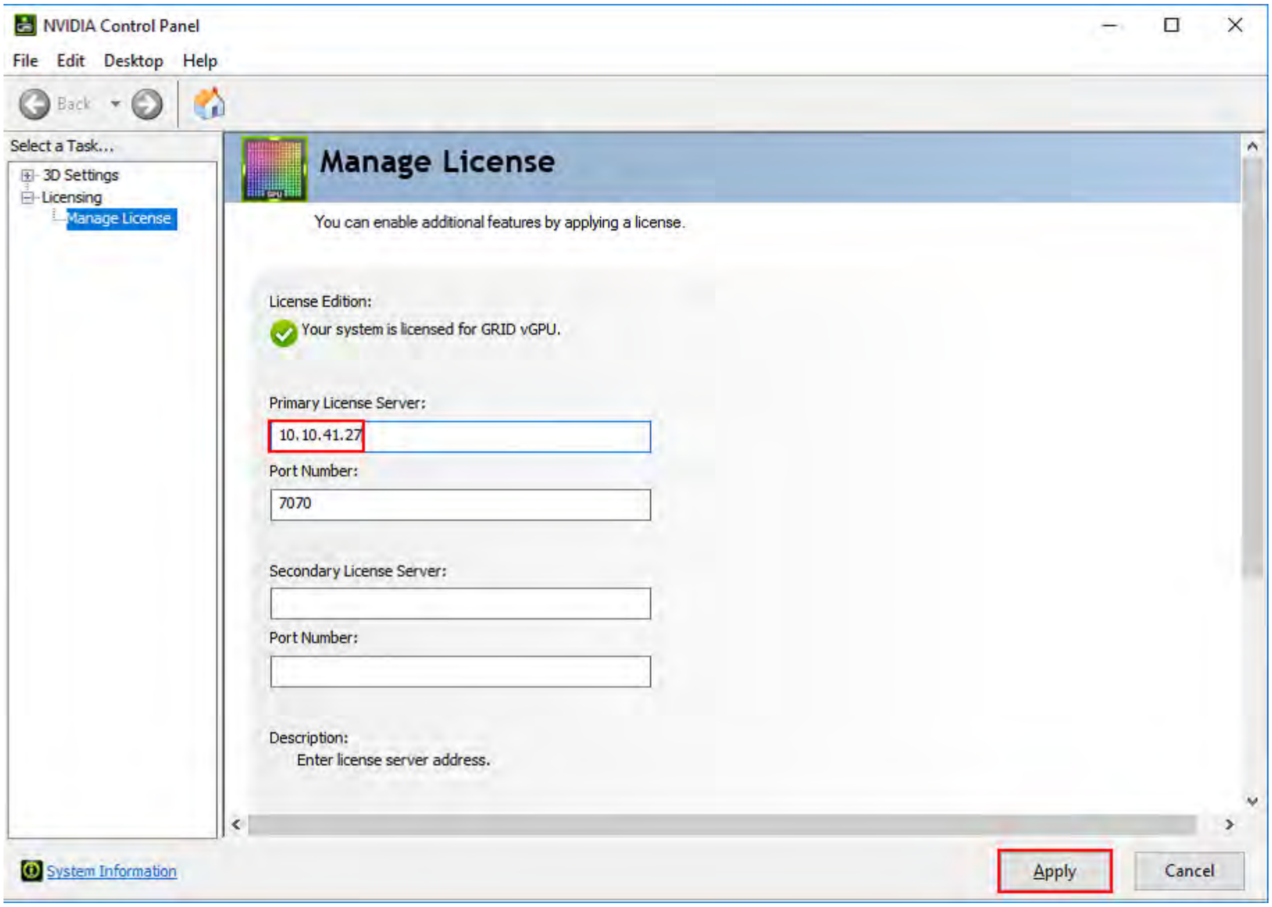
1. In the Microsoft Windows Control Panel, double-click NVIDIA Control Panel (Figure 55).

**Figure 55.** Choosing NVIDIA Control Panel


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

2. Select Manage License from the left pane and enter your license server address and port. Click Apply (Figure 56).

**Figure 56.** Managing your license.



Manage License

You can enable additional features by applying a license.

License Edition:
Your system is licensed for GRID vGPU.

Primary License Server: 10.10.41.27

Port Number: 7070

Secondary License Server:

Port Number:

Description:
Enter license server address.

## Verify vGPU deployment

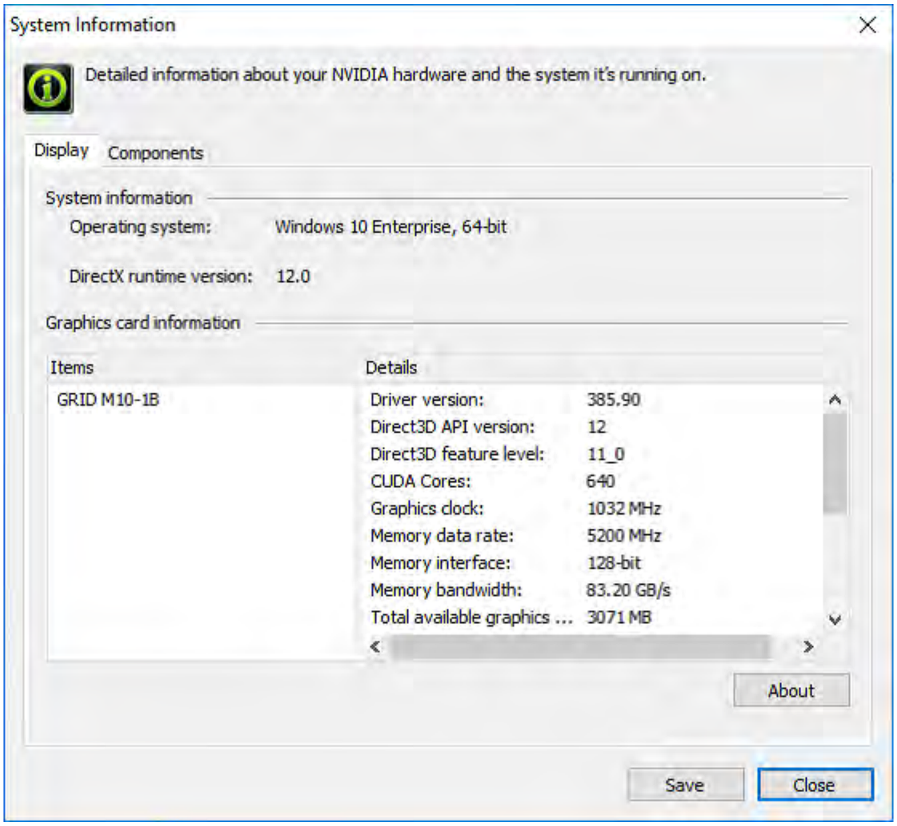
After the desktops are provisioned, use the following steps to verify vGPU deployment in the Citrix XenDesktop environment.

### Verify that the NVIDIA driver is running on the desktop

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

1. Right-click the desktop. In the menu, choose NVIDIA Control Panel to open the control panel.
2. In the control panel, select System Information to see the vGPU that the virtual machine is using, the vGPU's capabilities, and the NVIDIA driver version that is loaded (Figure 57).

**Figure 57.** NVIDIA Control Panel System Information window



System Information

Detailed information about your NVIDIA hardware and the system it's running on.

System information

Operating system: Windows 10 Enterprise, 64-bit

DirectX runtime version: 12.0

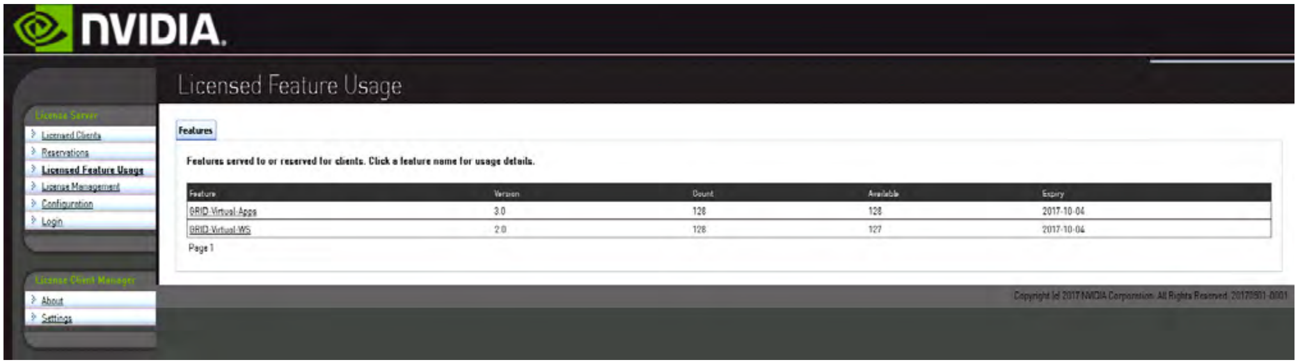
Graphics card information

| Items | Details | |
|---|---|---|
| GRID M10-1B | Driver version: | 385.90 |
| | Direct3D API version: | 12 |
| | Direct3D feature level: | 11_0 |
| | CUDA Cores: | 640 |
| | Graphics clock: | 1032 MHz |
| | Memory data rate: | 5200 MHz |
| | Memory interface: | 128-bit |
| | Memory bandwidth: | 83.20 GB/s |
| | Total available graphics ... | 3071 MB |

### 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 58).

**Figure 58.** NVIDIA License Server: Licensed Feature Usage window



Licensed Feature Usage

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

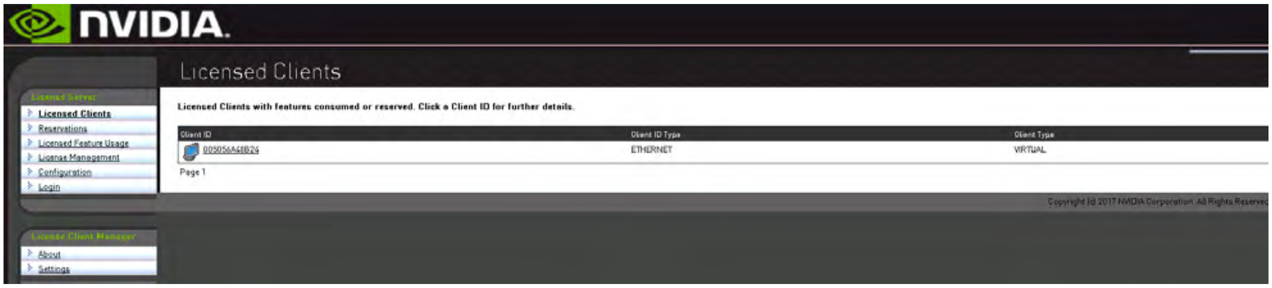
| Feature | Version | Count | Available | Expiry |
|---|---|---|---|---|
| GRID-Virtual-Apps | 3.0 | 128 | 128 | 2017-10-04 |
| GRID-Virtual-WS | 2.0 | 128 | 127 | 2017-10-04 |

Page 1

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

To view the details, select Licensed Clients in the left pane (Figure 59).

**Figure 59.** NVIDIA License Server: Licensed Clients window



Licensed Clients

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

| Client ID | Client ID Type | Client Type |
|---|---|---|
| 005056A48B24 | ETHERNET | VIRTUAL |

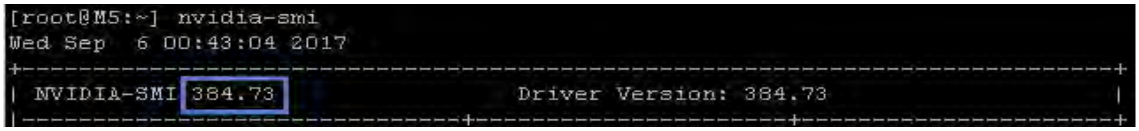
Page 1

Copyright (c) 2017 NVIDIA Corporation. All Rights Reserved

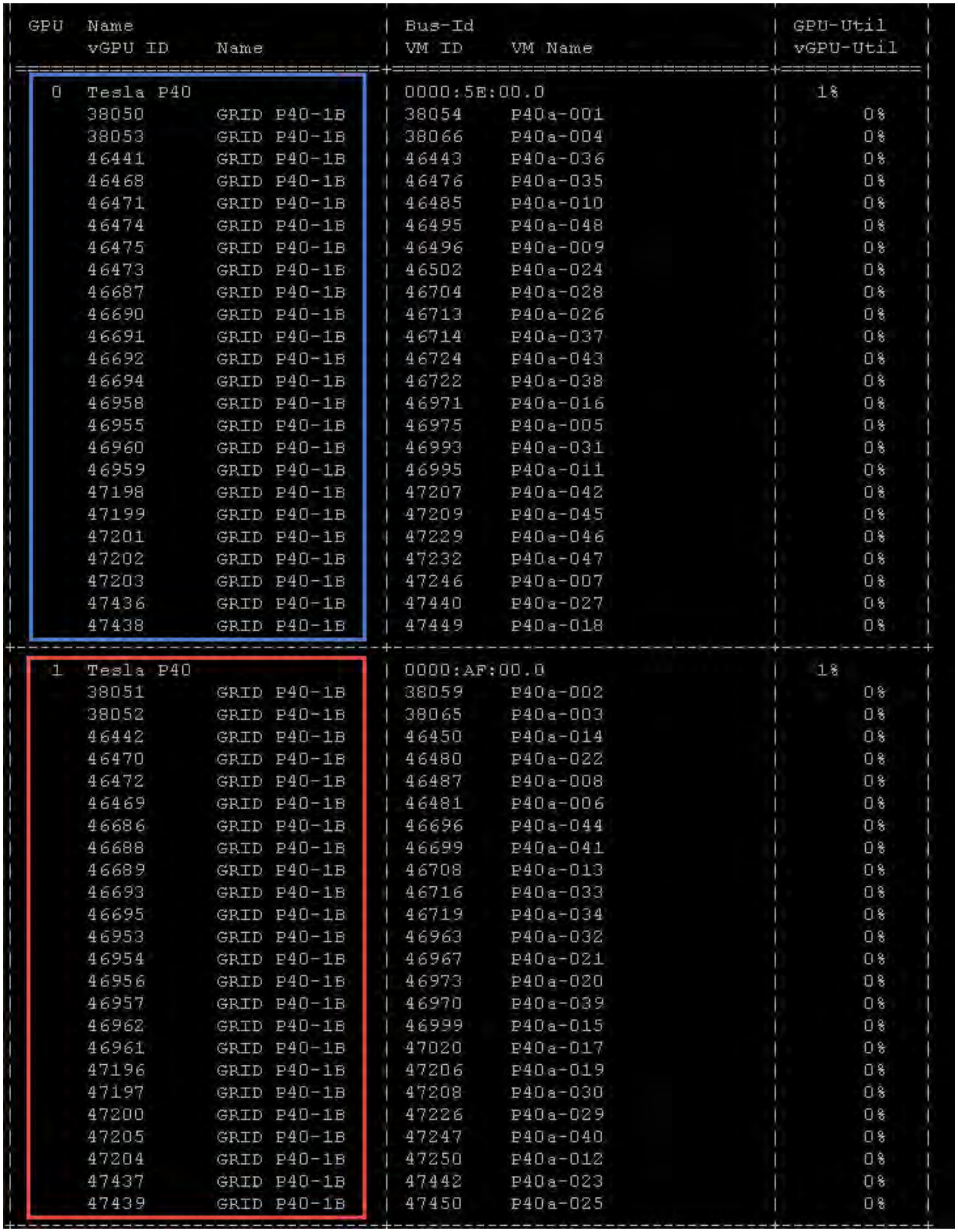
## Verify the NVIDIA configuration on the host

To obtain a hostwide overview of the NVIDIA GPUs, enter the nvidia-smi command without any arguments (Figures 60, 61, and 62).

**Figure 60.** The nvidia-smi command output from the host with two NVIDIA P40 cards and 48 Microsoft Windows 10 desktops with P40-1B vGPU profile

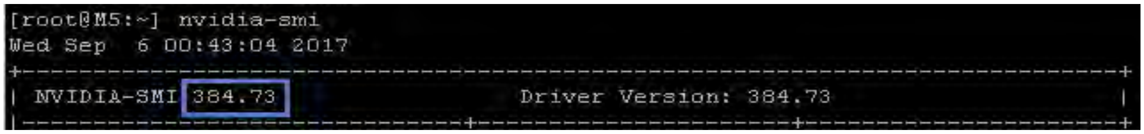


```
[root@M5:~] nvidia-smi
Wed Sep  6 00:43:04 2017
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 384.73                 Driver Version: 384.73                    |
|-------------------------------+----------------------+----------------------+
```

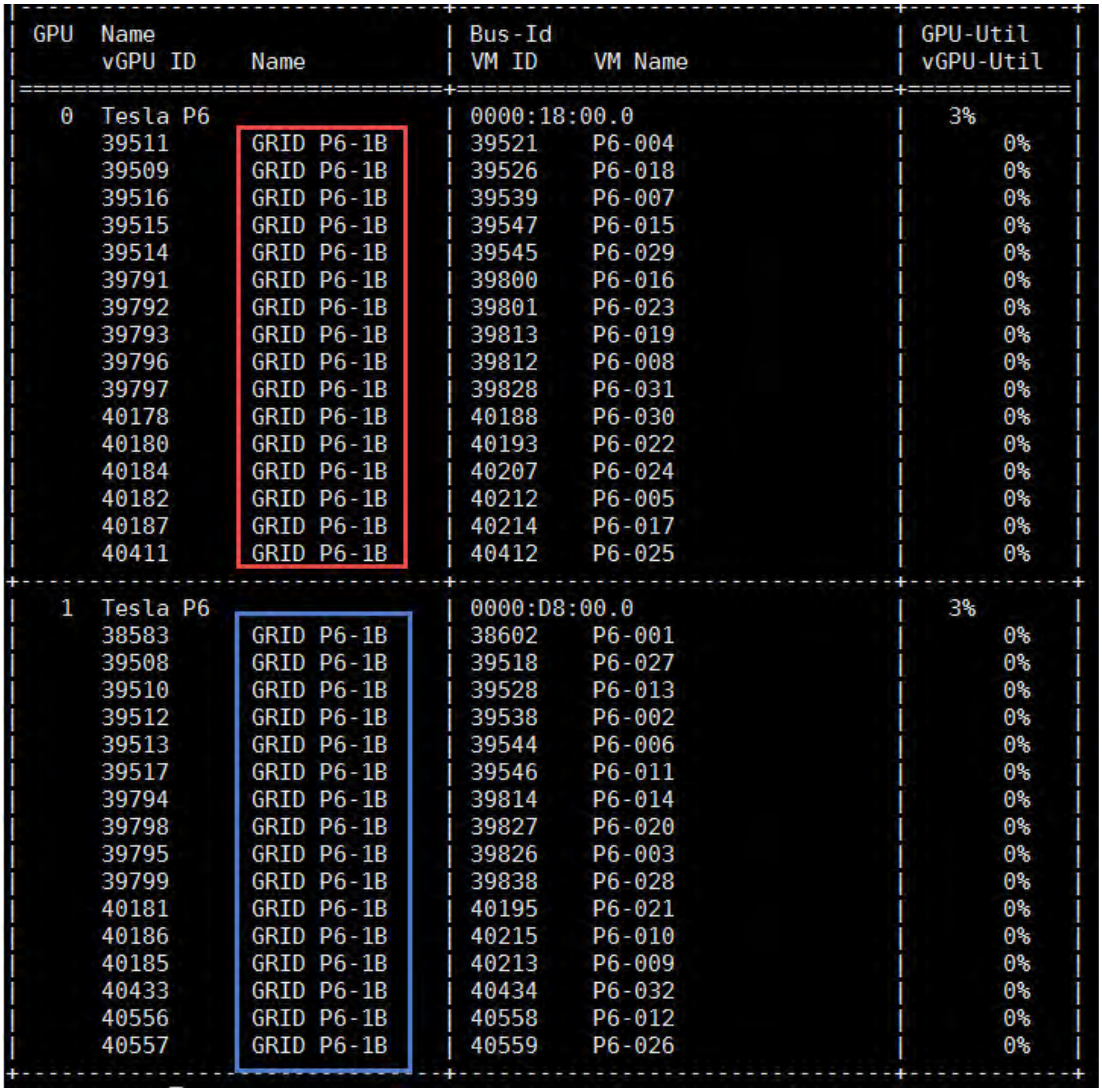


```
| GPU  Name                            | Bus-Id                              | GPU-Util    |
|      vGPU ID     Name                | VM ID     VM Name                   | vGPU-Util   |
|======================================+=====================================+=============|
|   0  Tesla P40                       | 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 61.** The nvidia-smi command output from the host with two NVIDIA P6 cards and 32 Microsoft Windows 10 desktops with P6-1B vGPU profile

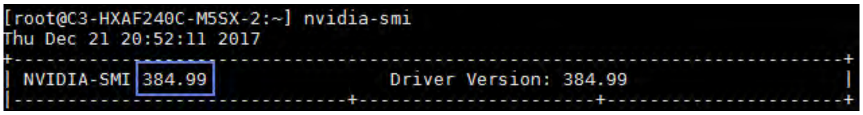


```
[root@M5:~] nvidia-smi
Wed Sep  6 00:43:04 2017
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 384.73                 Driver Version: 384.73                    |
|-------------------------------+----------------------+----------------------+
```

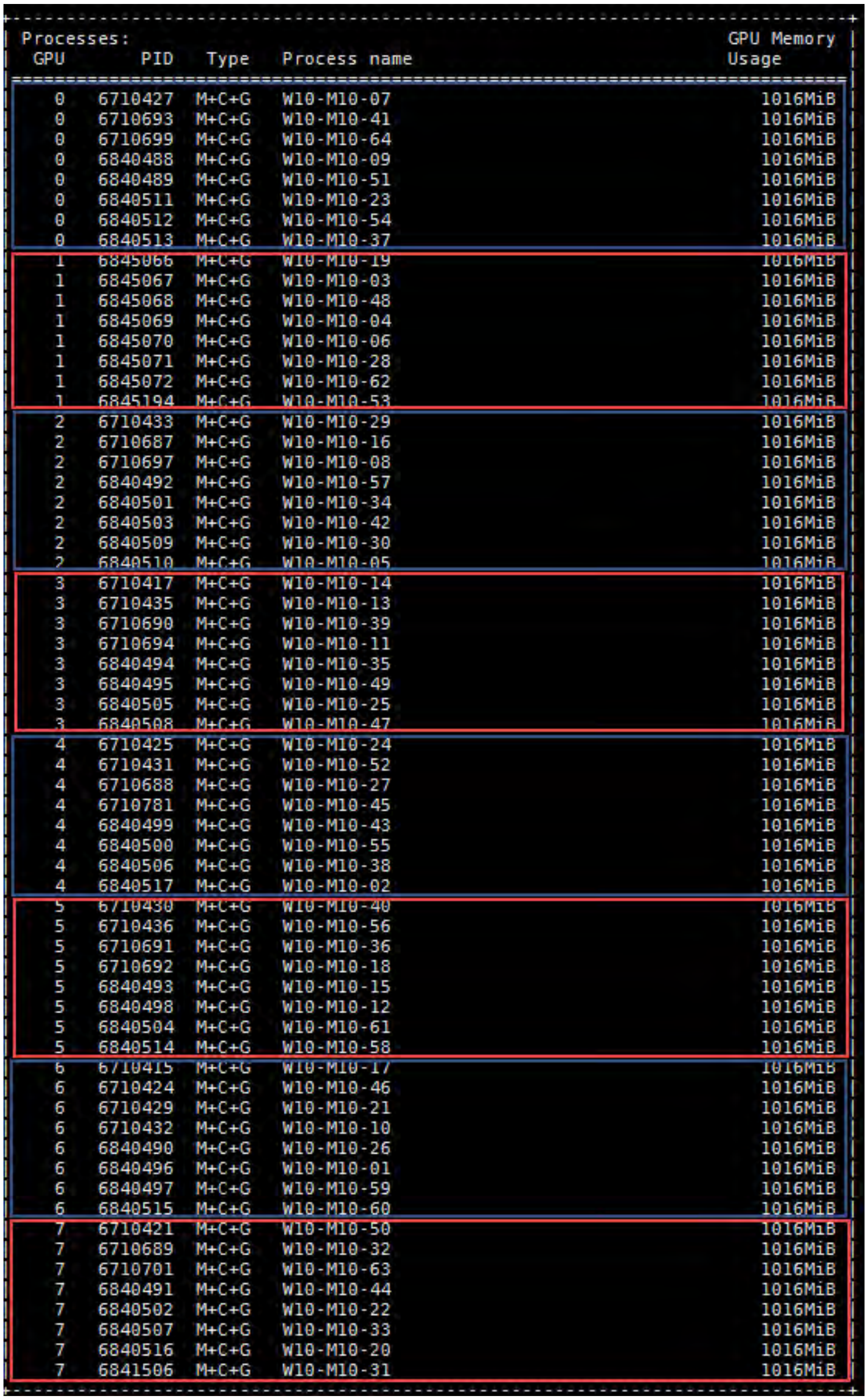


```
+-------------------------------+--------------------------------+------------+
| GPU  Name                     | Bus-Id                         | GPU-Util   |
|      vGPU ID     Name         | VM ID     VM Name              | vGPU-Util  |
|===============================+================================+============|
|   0  Tesla P6                 | 0000:18:00.0                   |   3%       |
|      39511       GRID P6-1B   | 39521     P6-004               |      0%    |
|      39509       GRID P6-1B   | 39526     P6-018               |      0%    |
|      39516       GRID P6-1B   | 39539     P6-007               |      0%    |
|      39515       GRID P6-1B   | 39547     P6-015               |      0%    |
|      39514       GRID P6-1B   | 39545     P6-029               |      0%    |
|      39791       GRID P6-1B   | 39800     P6-016               |      0%    |
|      39792       GRID P6-1B   | 39801     P6-023               |      0%    |
|      39793       GRID P6-1B   | 39813     P6-019               |      0%    |
|      39796       GRID P6-1B   | 39812     P6-008               |      0%    |
|      39797       GRID P6-1B   | 39828     P6-031               |      0%    |
|      40178       GRID P6-1B   | 40188     P6-030               |      0%    |
|      40180       GRID P6-1B   | 40193     P6-022               |      0%    |
|      40184       GRID P6-1B   | 40207     P6-024               |      0%    |
|      40182       GRID P6-1B   | 40212     P6-005               |      0%    |
|      40187       GRID P6-1B   | 40214     P6-017               |      0%    |
|      40411       GRID P6-1B   | 40412     P6-025               |      0%    |
+-------------------------------+--------------------------------+------------+
|   1  Tesla P6                 | 0000:D8:00.0                   |   3%       |
|      38583       GRID P6-1B   | 38602     P6-001               |      0%    |
|      39508       GRID P6-1B   | 39518     P6-027               |      0%    |
|      39510       GRID P6-1B   | 39528     P6-013               |      0%    |
|      39512       GRID P6-1B   | 39538     P6-002               |      0%    |
|      39513       GRID P6-1B   | 39544     P6-006               |      0%    |
|      39517       GRID P6-1B   | 39546     P6-011               |      0%    |
|      39794       GRID P6-1B   | 39814     P6-014               |      0%    |
|      39798       GRID P6-1B   | 39827     P6-020               |      0%    |
|      39795       GRID P6-1B   | 39826     P6-003               |      0%    |
|      39799       GRID P6-1B   | 39838     P6-028               |      0%    |
|      40181       GRID P6-1B   | 40195     P6-021               |      0%    |
|      40186       GRID P6-1B   | 40215     P6-010               |      0%    |
|      40185       GRID P6-1B   | 40213     P6-009               |      0%    |
|      40433       GRID P6-1B   | 40434     P6-032               |      0%    |
|      40556       GRID P6-1B   | 40558     P6-012               |      0%    |
|      40557       GRID P6-1B   | 40559     P6-026               |      0%    |
+-------------------------------+--------------------------------+------------+
```

**Figure 62.** The nvidia-smi command output from the host with two NVIDIA M10 cards and 64 Microsoft Windows 10 desktops with M10-1B vGPU profile



```
[root@C3-HXAF240C-M5SX-2:~] nvidia-smi
Thu Dec 21 20:52:11 2017
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 384.99                 Driver Version: 384.99                    |
|-------------------------------+----------------------+----------------------+
```



```
+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|    0   6710427  M+C+G   W10-M10-07                                  1016MiB |
|    0   6710693  M+C+G   W10-M10-41                                  1016MiB |
|    0   6710699  M+C+G   W10-M10-64                                  1016MiB |
|    0   6840488  M+C+G   W10-M10-09                                  1016MiB |
|    0   6840489  M+C+G   W10-M10-51                                  1016MiB |
|    0   6840511  M+C+G   W10-M10-23                                  1016MiB |
|    0   6840512  M+C+G   W10-M10-54                                  1016MiB |
|    0   6840513  M+C+G   W10-M10-37                                  1016MiB |
|    1   6845066  M+C+G   W10-M10-19                                  1016MiB |
|    1   6845067  M+C+G   W10-M10-03                                  1016MiB |
|    1   6845068  M+C+G   W10-M10-48                                  1016MiB |
|    1   6845069  M+C+G   W10-M10-04                                  1016MiB |
|    1   6845070  M+C+G   W10-M10-06                                  1016MiB |
|    1   6845071  M+C+G   W10-M10-28                                  1016MiB |
|    1   6845072  M+C+G   W10-M10-62                                  1016MiB |
|    1   6845194  M+C+G   W10-M10-53                                  1016MiB |
|    2   6710433  M+C+G   W10-M10-29                                  1016MiB |
|    2   6710687  M+C+G   W10-M10-16                                  1016MiB |
|    2   6710697  M+C+G   W10-M10-08                                  1016MiB |
|    2   6840492  M+C+G   W10-M10-57                                  1016MiB |
|    2   6840501  M+C+G   W10-M10-34                                  1016MiB |
|    2   6840503  M+C+G   W10-M10-42                                  1016MiB |
|    2   6840509  M+C+G   W10-M10-30                                  1016MiB |
|    2   6840510  M+C+G   W10-M10-05                                  1016MiB |
|    3   6710417  M+C+G   W10-M10-14                                  1016MiB |
|    3   6710435  M+C+G   W10-M10-13                                  1016MiB |
|    3   6710690  M+C+G   W10-M10-39                                  1016MiB |
|    3   6710694  M+C+G   W10-M10-11                                  1016MiB |
|    3   6840494  M+C+G   W10-M10-35                                  1016MiB |
|    3   6840495  M+C+G   W10-M10-49                                  1016MiB |
|    3   6840505  M+C+G   W10-M10-25                                  1016MiB |
|    3   6840508  M+C+G   W10-M10-47                                  1016MiB |
|    4   6710425  M+C+G   W10-M10-24                                  1016MiB |
|    4   6710431  M+C+G   W10-M10-52                                  1016MiB |
|    4   6710688  M+C+G   W10-M10-27                                  1016MiB |
|    4   6710781  M+C+G   W10-M10-45                                  1016MiB |
|    4   6840499  M+C+G   W10-M10-43                                  1016MiB |
|    4   6840500  M+C+G   W10-M10-55                                  1016MiB |
|    4   6840506  M+C+G   W10-M10-38                                  1016MiB |
|    4   6840517  M+C+G   W10-M10-02                                  1016MiB |
|    5   6710430  M+C+G   W10-M10-40                                  1016MiB |
|    5   6710436  M+C+G   W10-M10-56                                  1016MiB |
|    5   6710691  M+C+G   W10-M10-36                                  1016MiB |
|    5   6710692  M+C+G   W10-M10-18                                  1016MiB |
|    5   6840493  M+C+G   W10-M10-15                                  1016MiB |
|    5   6840498  M+C+G   W10-M10-12                                  1016MiB |
|    5   6840504  M+C+G   W10-M10-61                                  1016MiB |
|    5   6840514  M+C+G   W10-M10-58                                  1016MiB |
|    6   6710415  M+C+G   W10-M10-17                                  1016MiB |
|    6   6710424  M+C+G   W10-M10-46                                  1016MiB |
|    6   6710429  M+C+G   W10-M10-21                                  1016MiB |
|    6   6710432  M+C+G   W10-M10-10                                  1016MiB |
|    6   6840490  M+C+G   W10-M10-26                                  1016MiB |
|    6   6840496  M+C+G   W10-M10-01                                  1016MiB |
|    6   6840497  M+C+G   W10-M10-59                                  1016MiB |
|    6   6840515  M+C+G   W10-M10-60                                  1016MiB |
|    7   6710421  M+C+G   W10-M10-50                                  1016MiB |
|    7   6710689  M+C+G   W10-M10-32                                  1016MiB |
|    7   6710701  M+C+G   W10-M10-63                                  1016MiB |
|    7   6840491  M+C+G   W10-M10-44                                  1016MiB |
|    7   6840502  M+C+G   W10-M10-22                                  1016MiB |
|    7   6840507  M+C+G   W10-M10-33                                  1016MiB |
|    7   6840516  M+C+G   W10-M10-20                                  1016MiB |
|    7   6841506  M+C+G   W10-M10-31                                  1016MiB |
+-----------------------------------------------------------------------------+
```

## 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 graphical application session is running. After closing the application session, a user can change the resolution of the Desktop Viewer window in Citrix Receiver Desktop Viewer Preferences.

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

For example, if 20 users share a 60-Mbps connection, the bandwidth available to each user can vary between 3 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 OpenGL Software Accelerator

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, OpenGL Software Accelerator can eliminate the need to use graphics cards to deliver a good user experience with OpenGL applications.

**Note:** 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 OpenGL Software Accelerator works with your applications, you can use it to avoid the cost of GPU hardware.

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

Try OpenGL Software Accelerator in the following cases:

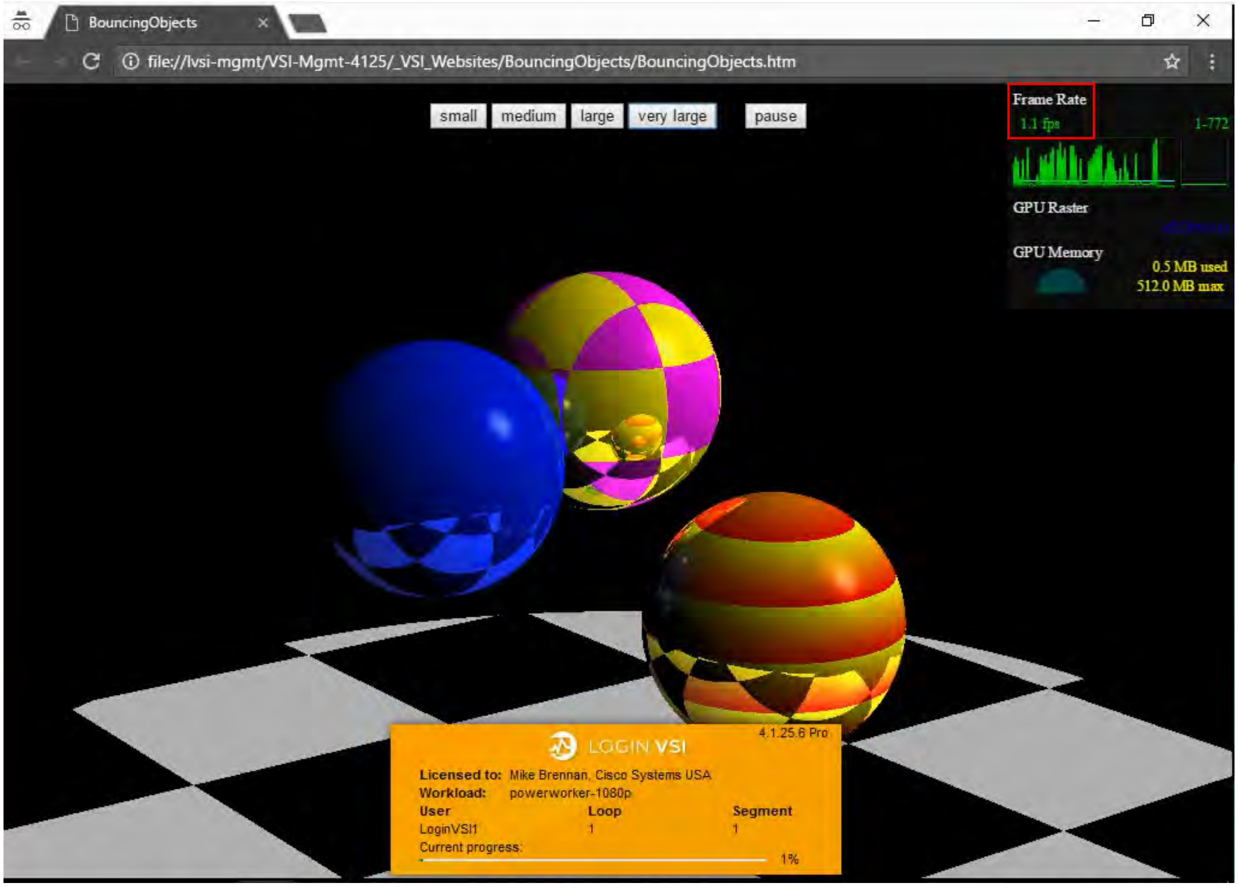
- If the performance of OpenGL applications running on virtual machines is a concern, try using the OpenGL accelerator. For some applications, the accelerator outperforms the Microsoft OpenGL software rasterizer that is included with Windows because the OpenGL accelerator uses SSE4.1 and AVX. The OpenGL 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 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.

## Test and evaluation notes

The Login VSI test framework with a custom power user workload was used to simulate users running graphics-intensive workloads and high-definition video content.

Figures 63 and 64 show differences in frames-per-second (FPS) rates during tests between desktops configured with and without a vGPU.

**Figure 63.** BouncingObjects.htm FPS on desktop without vGPU



file://lvsi-mgmt/VSI-Mgmt-4125/_VSI_Websites/BouncingObjects/BouncingObjects.htm

Frame Rate
1.1 fps
1-772

GPU Memory
0.5 MB used
512.0 MB max

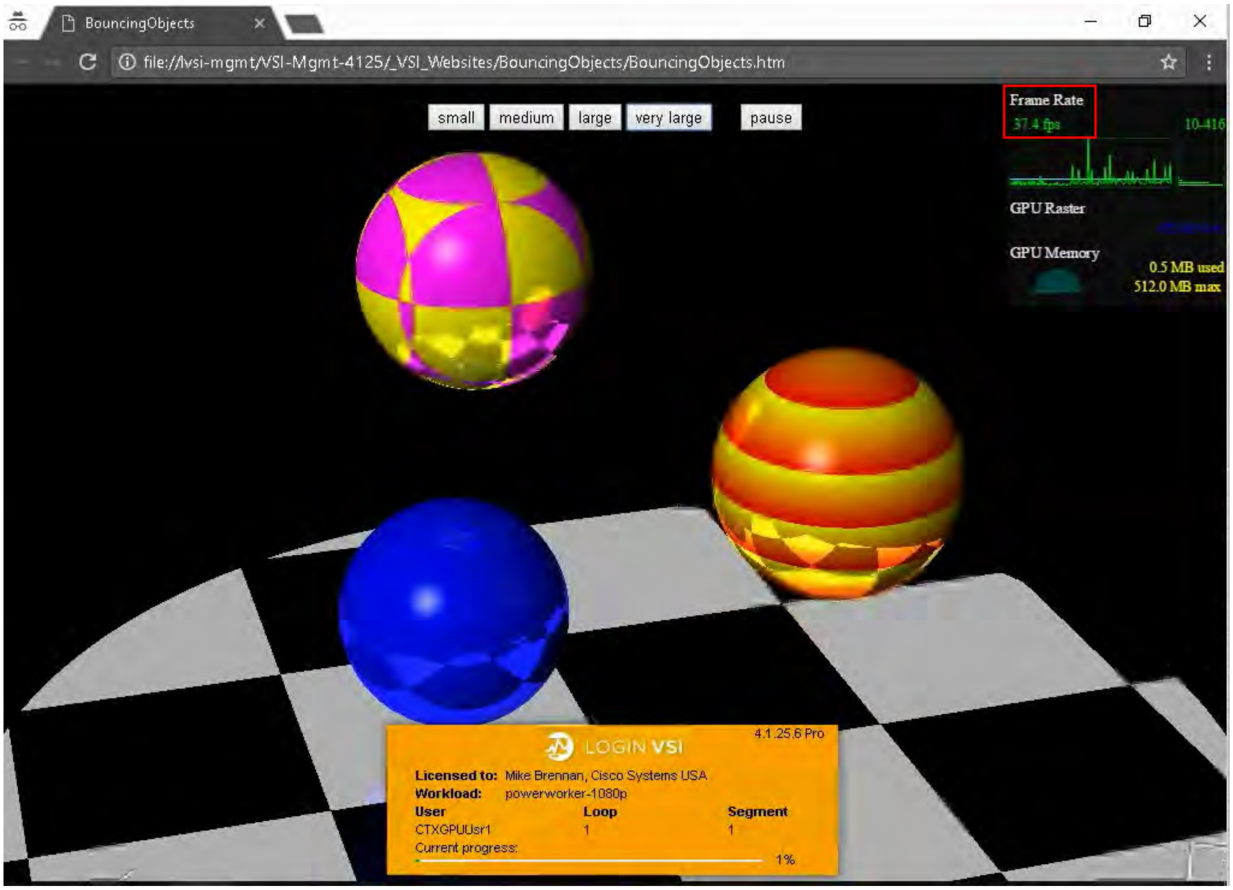
LOGIN VSI 4.1.25.6 Pro

Licensed to: Mike Brennan, Cisco Systems USA
Workload: powerworker-1080p

| User | Loop | Segment |
|---|---|---|
| LoginVSI1 | 1 | 1 |

Current progress: 1%

**Figure 64.** BouncingObjects.htm FPS on desktop with vGPU



file://lvsi-mgmt/VSI-Mgmt-4125/_VSI_Websites/BouncingObjects/BouncingObjects.htm

37.4 fps 10-416

0.5 MB used
512.0 MB max

4.1.25.6 Pro

**Licensed to:** Mike Brennan, Cisco Systems USA
**Workload:** powerworker-1080p

| User | Loop | Segment |
|---|---|---|
| CTXGPUUsr1 | 1 | 1 |

Current progress: 1%

Figures 65 and 66 provide examples of the CPU utilization seen during tests from the hosts running the same number of Microsoft Windows 10 desktops with and without a vGPU.

**Figure 65.** CPU utilization from the host with two NVIDIA P40 cards running 48 Microsoft Windows 10 desktops with P40-1B vGPU profile and the host without GPU running 48 Microsoft Windows 10 desktops

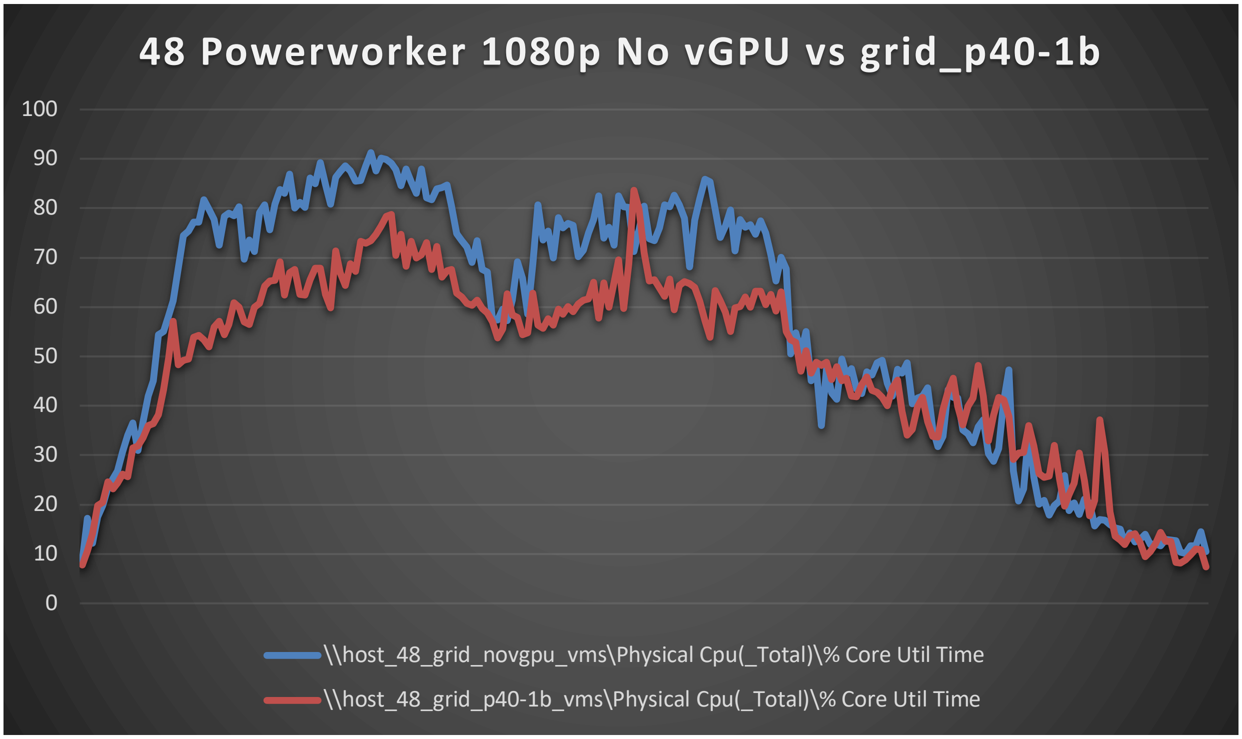
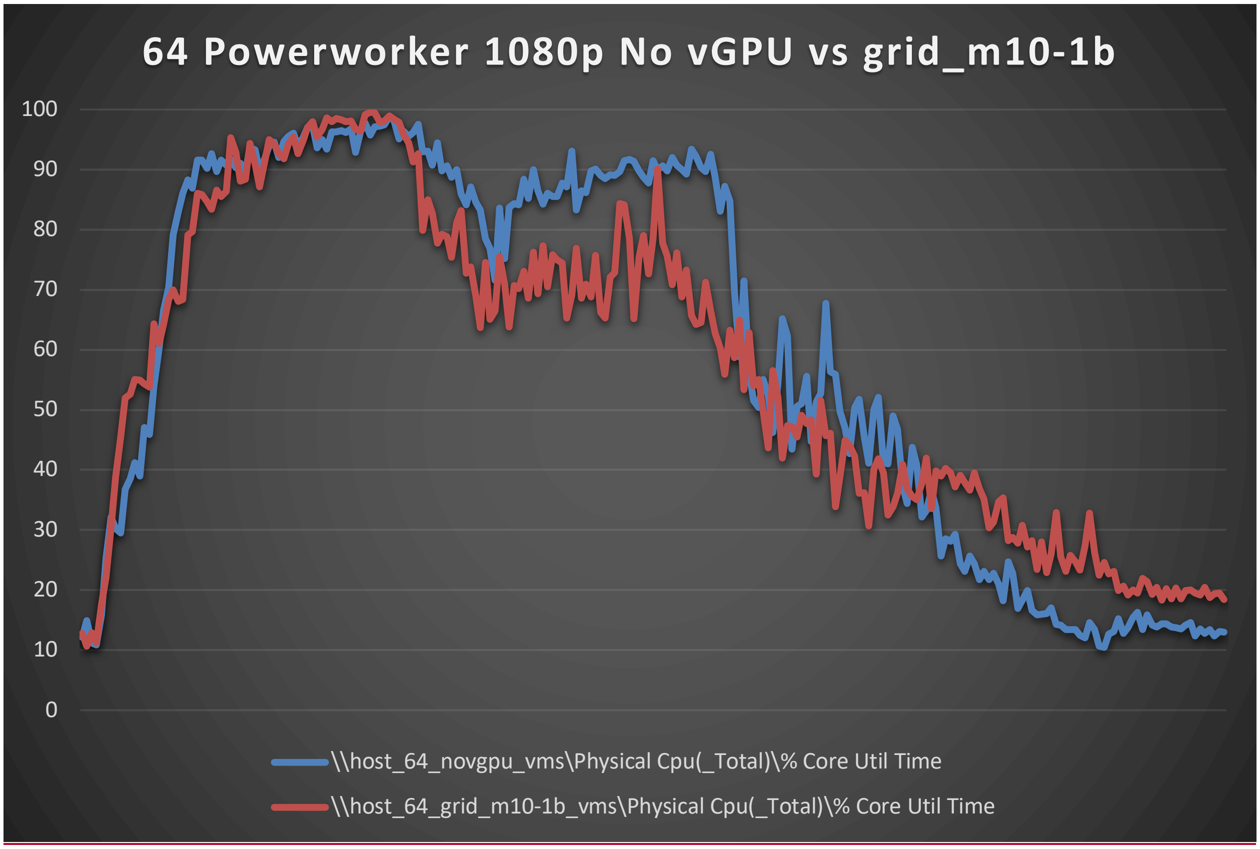


**Figure 66.** CPU utilization from the host with two NVIDIA M10 cards running 64 Microsoft Windows 10 desktops with M10-1B vGPU profile and the host without GPU running 64 Microsoft Windows 10 desktops



## Conclusion

The combination of Cisco UCS Manager, Cisco HyperFlex 2.6, Cisco UCS C240 M5 Rack Servers and B200 M5 Blade Servers, and NVIDIA Tesla cards running VMware vSphere 6.5 and Citrix XenDesktop 7.15 provides a high-performance platform for virtualizing graphics-intensive workloads.

By following the configuration 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/
- Cisco HyperFlex hyperconverged servers:
  - https://www.cisco.com/c/en/us/products/hyperconverged-infrastructure/hyperflex-hx-series/index.html
- NVIDIA:
  - http://www.nvidia.com/object/grid-technology.html
  - http://docs.nvidia.com/grid/latest/pdf/grid-software-quick-start-guide.pdf
  - http://docs.nvidia.com/grid/latest/pdf/grid-vgpu-release-notes-vmware-vsphere.pdf

- Citrix XenApp and XenDesktop 7.15:
    - https://docs.citrix.com/en-us/xenapp-and-xendesktop/7-15-ltsr.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
    - 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.

Printed in USA C11-740243-00 02/18