White Paper

# FlexPod for 3D Graphics Visualization with Citrix and NVIDIA

## Understand the performance of the Cisco UCS B200 M5 Blade Server and C240 M5 Rack Server solution with NetApp AFF A300 system on Citrix XenServer 7.5 and XenDesktop 7.15 LTSR

# Contents

## What you will learn

Using the increased processing power of today's Cisco UCS® B-Series Blade Servers and C-Series Rack 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 client computing (VCC), Cisco offers support for the NVIDIA GRID P6, P40, and P4 cards in the Cisco Unified Computing System™ (Cisco UCS) portfolio of PCI Express (PCIe) and mezzanine form-factor cards for the B-Series Blade Servers and C-Series Rack Servers.

With the availability of these 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. These new graphics capabilities help enable organizations to centralize their graphics workloads and data in the data center, facilitating collaboration across geographical boundaries.

A major focus of this document is the FlexPod Datacenter infrastructure and Citrix support for the NVIDIA GRID virtual graphics processing unit (vGPU), including the capability of Citrix XenServer, through Citrix XenMotion, to move vGPU-enabled virtual machines, reducing user downtime.

The purpose of this document is to help our partners and customers integrate NVIDIA GRID 6.2 software and NVIDIA Tesla graphics cards and Cisco UCS B200 M5 Blade Servers and C240 M5 Rack Servers with the Citrix XenServer 7.5 hypervisor and Citrix XenDesktop 7.15 desktop virtualization software using Microsoft Windows 10 virtual machines in vGPU mode. This document is an extension of the Cisco Validated Design FlexPod Datacenter with Citrix XenDesktop and XenApp 7.15 and VMware vSphere 6.5 Update 1 for 6000 Seats and focuses on NVIDIA GPU testing.

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

This document describes in detail how to integrate Cisco FlexPod Datacenter architecture using NVIDIA Tesla P6, P40, and P4 cards with Citrix products so that the servers, hypervisor, and virtual desktops are ready for installation of high-performance graphics applications.

For the first time, we are using SPECviewperf 13 to provide relative performance information for NVIDIA Tesla graphics cards for the eight high-performance applications included in the tool. We also measured the impact of various frame buffer sizes (profiles) on the same card set. In all cases except one, the testing was performed in benchmark mode. The goal is to give readers a starting point to help them select the right card for their application environments.

## vGPU profiles

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

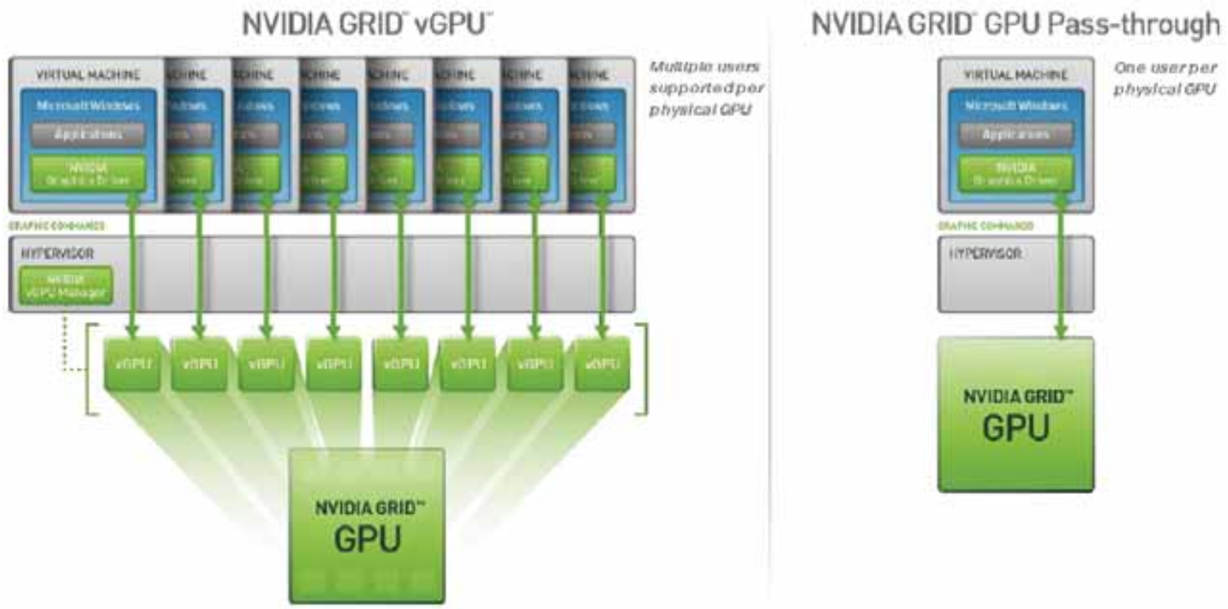
Although the needs of end users can be diverse, for simplicity users can be grouped into the following categories: knowledge workers, designers, and power users.

- 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.

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

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



## Cisco Unified Computing System

The main components of Cisco UCS are:

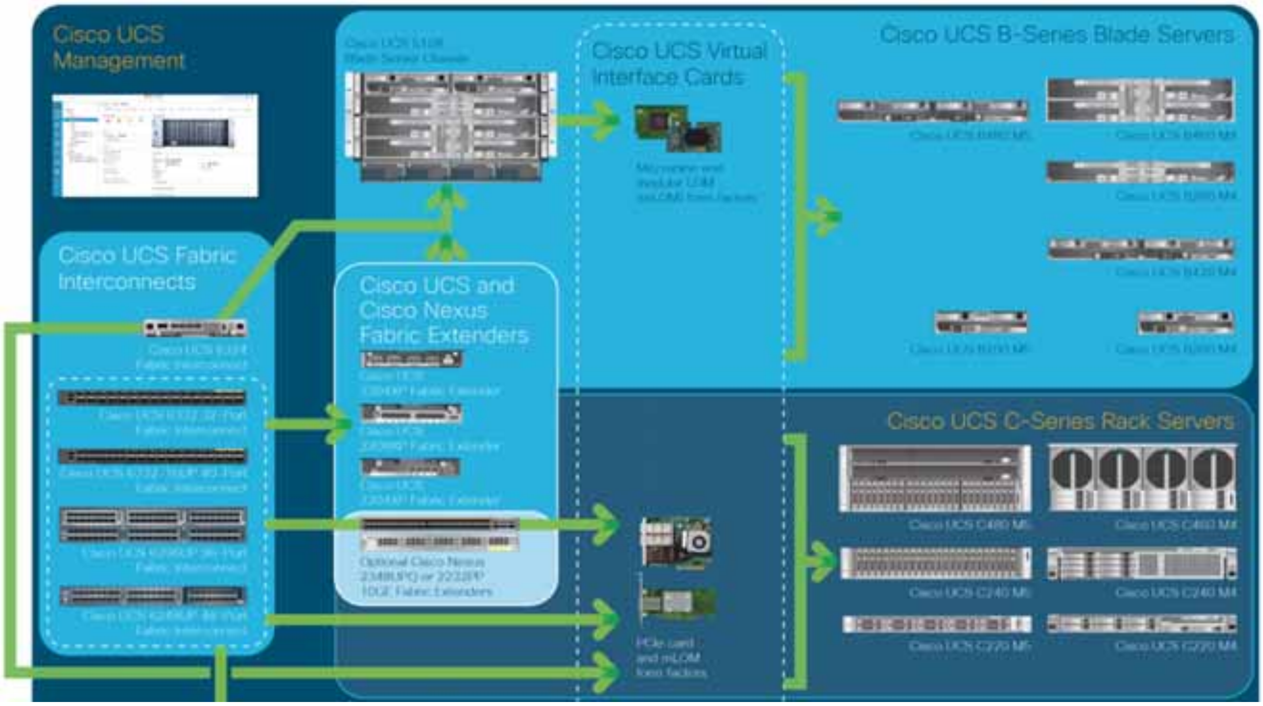
Compute: The system is based on an entirely new class of computing system that incorporates blade servers based on Intel® Xeon® processor E5-2600 and 4600 v3 and E7-2800 v3 family CPUs.

- **Network:** The system is integrated on 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 needed and by decreasing the 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.

Figure 2 provides an overview of the Cisco® data center with Cisco UCS.

**Figure 2.** Cisco Data center overview


Cisco UCS Management
Cisco UCS 5108 Blade Server Chassis
Cisco UCS Virtual Interface Cards
Cisco UCS B-Series Blade Servers
Cisco UCS Fabric Interconnects
Cisco UCS and Cisco Nexus Fabric Extenders
Optional Cisco Nexus 2348UPQ or 2232PP 10GE Fabric Extenders
PCIe card and mLOM form factors
Cisco UCS C-Series Rack Servers
Cisco UCS C480 M5
Cisco UCS C460 M4
Cisco UCS C240 M5
Cisco UCS C240 M4
Cisco UCS C220 M5
Cisco UCS C220 M4

Cisco UCS is designed to deliver:

- Reduced total cost of ownership (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 provides unified, embedded management of all software and hardware components of the Cisco Unified Computing System across multiple chassis, rack servers, and thousands of virtual machines. Cisco UCS Manager manages Cisco UCS as a single entity through an intuitive GUI, a CLI, or an XML API for comprehensive access to all Cisco UCS Manager functions.

### Cisco UCS Manager

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

### Cisco UCS 6332 Fabric Interconnect

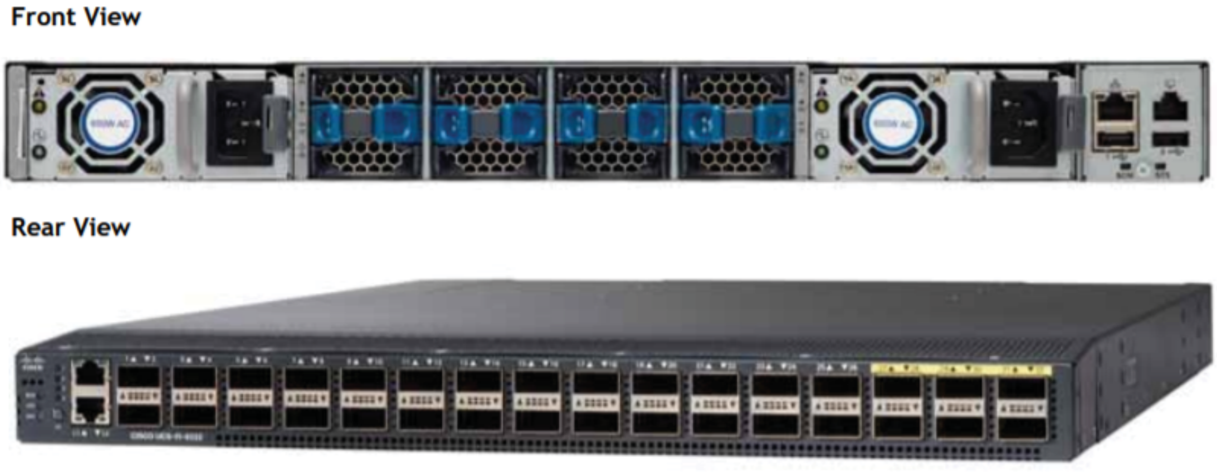
The Cisco UCS 6332 Fabric Interconnect (Figure 3) is the management and communication backbone for Cisco UCS B-Series Blade Servers, C-Series Rack Servers, and 5100 Series Blade Server Chassis. All servers attached to 6332 Fabric Interconnects become part of one highly available management domain.

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.

Features and capabilities include:

- Bandwidth of up to 2.56-Tbps full-duplex throughput
- Thirty-two 40-Gbps QSFP+ ports in one 1 rack unit (RU)
- 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



### Cisco UCS C-Series Rack Servers

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

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

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

### Cisco UCS C240 M5 Rack Server

The Cisco UCS C240 M5 Rack Server (Figure 4, Figure 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 Cisco UCS 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 family processors, 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-Gbps SAS storage controller card. The C240 M5 server includes a dedicated internal modular LAN on motherboard (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 2 x 10GBASE-T Intel x550 LOM ports (embedded on the motherboard).

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 of up to 2666 MHz for improved performance and lower power consumption
- One or two Intel Xeon processor scalable family CPUs
- Up to 6 PCIe 3.0 slots (4 full-height, full-length for GPU)
- Six hot-swappable fans for front-to-rear cooling
- 24 SFF front-facing SAS/SATA hard disk drives (HDDs) or SAS/SATA solid state disks (SSDs)
- Optionally, up to two front-facing SFF Non-Volatile Memory Express (NVMe) PCIe SSDs (replacing SAS/SATA drives); these drives must be placed in front drive bays 1 and 2 only and are controlled from Riser 2 option C
- Optionally, up to two SFF, rear-facing SAS/SATA HDDs or SSDs, or up to two rear-facing SFF NVMe PCIe SSDs, with rear-facing SFF NVMe drives connected from Riser 2, Option B or C; 12-Gbps SAS drives are also supported
- The dedicated mLOM slot on the motherboard can flexibly accommodate the following cards:
  - Cisco VICs
  - Quad-port Intel i350 1GbE RJ45 mLOM NIC
  - Two 1 Gigabit Ethernet embedded LOM ports
- Support for up to 2 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, hot-swappable and hot-pluggable components
- One slot for a micro-SD card on PCIe Riser 1 (Option 1 and 1B); the micro-SD card serves as a dedicated local resource for utilities such as the Cisco Host Upgrade Utility (HUU), and 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 that supports either:
  - An SD card module with two SD card slots; mixing of different capacity SD cards is not supported
  - An M.2 module with two SATA M.2 SSD slots; mixing of different capacity M.2 modules is not supported

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

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

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

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

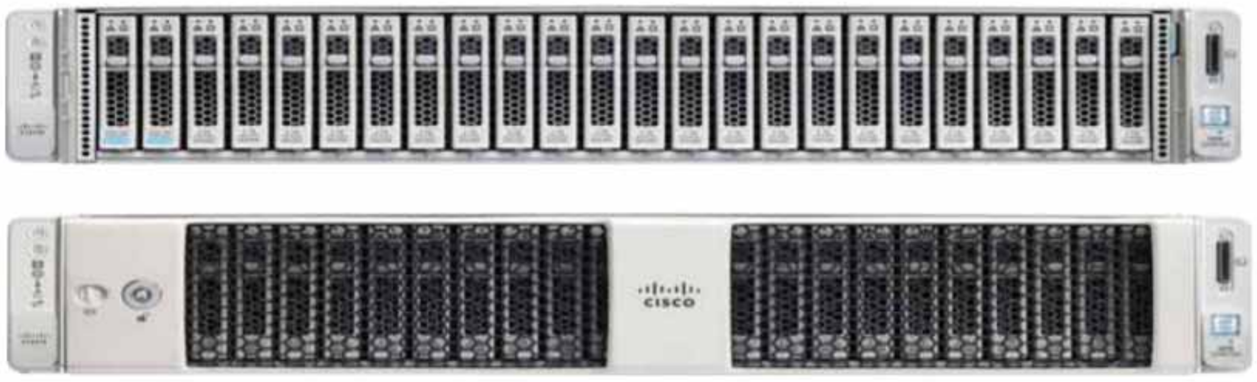


**Figure 5.** Cisco UCS C240 M4 Rack Server rear view

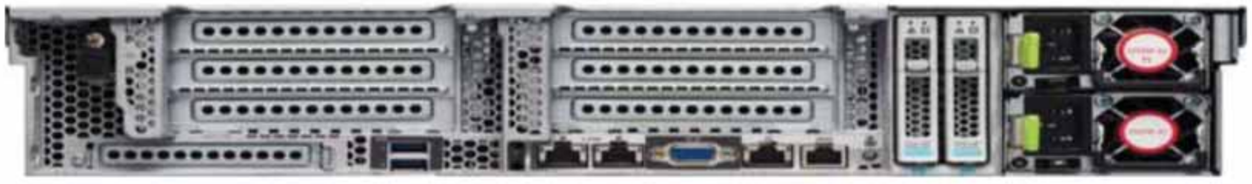


**Table 1.** Cisco UCS C240 M4 PCIe slots

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

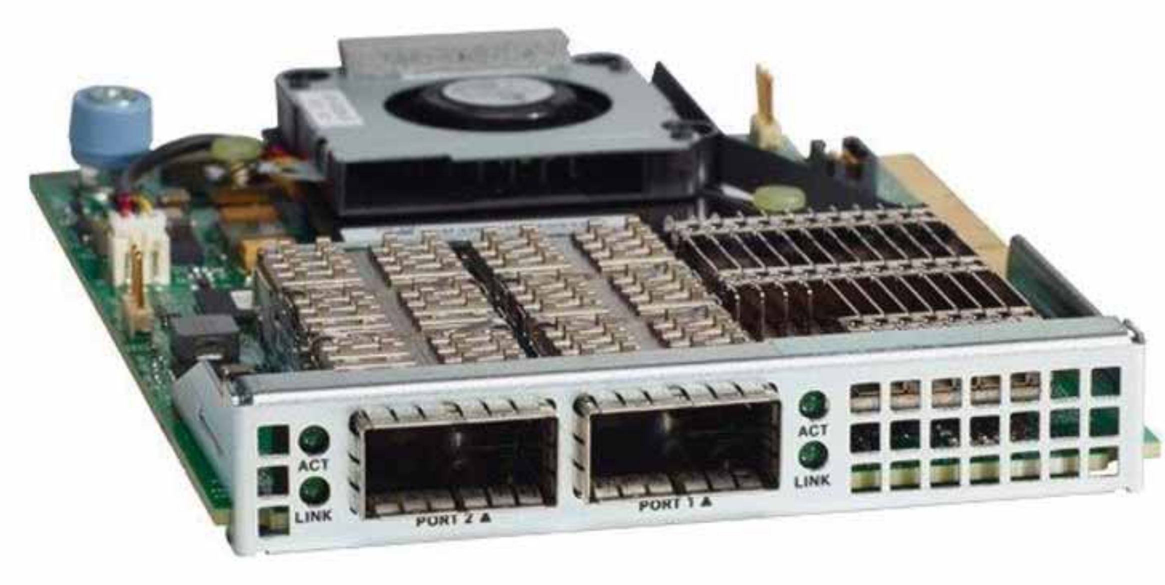
### Cisco UCS VIC 1387

The Cisco UCS VIC 1387 (Figure 6) is a dual-port Enhanced Small Form-Factor Pluggable (SFP+) 40-Gbps Ethernet and FCoE-capable PCIe mLOM adapter installed in the Cisco UCS C-Series Rack Servers. The mLOM slot can be used to install a Cisco VIC without consuming a PCIe slot, which provides greater I/O expandability. It incorporates next-generation converged network adapter (CNA) technology from Cisco, providing investment protection for future feature releases. The card enables a policy-based, stateless, agile server infrastructure that can present more than 256 PCIe standards-compliant interfaces to the host that can be dynamically configured as either NICs or host bus adapters (HBAs). The personality of the card is determined dynamically at boot time using the service profile associated with the server. The number, type (NIC or HBA), identity (MAC address and World Wide

Name [WWN]), failover policy, bandwidth, and quality-of-service (QoS) policies of the PCIe interfaces are all determined using the service profile.

For more information about the VIC, see https://www.cisco.com/c/en/us/products/interfaces-modules/ucs-virtual-interface-card-1387/index.html.

**Figure 6.** Cisco UCS VIC 1387 CNA



### Cisco UCS B200 M5 Blade Server

Delivering performance, versatility and density without compromise, the Cisco UCS B200 M5 Blade Server (Figure 7) addresses the broadest set of workloads, from IT and web infrastructure to distributed database workloads. The enterprise-class Cisco UCS 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 processor scalable family 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.

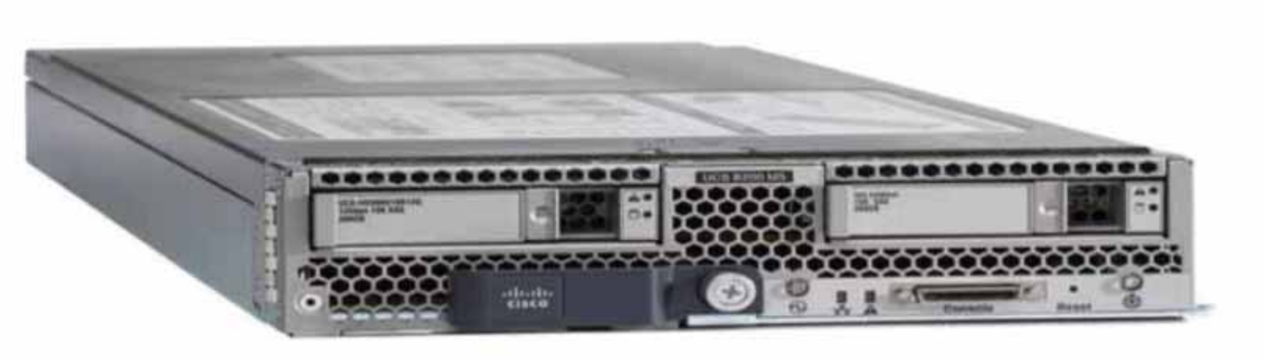
The B200 M5 server mounts in a Cisco UCS 5100 Series Blade Server Chassis or Cisco UCS Mini blade server chassis. It has 24 slots for error-correcting code (ECC) registered DIMMs (RDIMMs) or load-reduced DIMMs (LR DIMMs). It supports one connector for the Cisco UCS VIC 1340 adapter, which provides Ethernet and FCoE.

The B200 M5 has one rear mezzanine adapter slot, which can be configured with a Cisco UCS port expander card for the VIC. This hardware option enables 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 has one front mezzanine slot. The B200 M5 can be ordered with or without the front mezzanine card. The front mezzanine card can accommodate a storage controller or an NVIDIA P6 GPU.

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

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



### Cisco UCS VIC 1340

The Cisco UCS VIC 1340 (Figure 8) is a 2-port 40-Gbps Ethernet or dual 4 x 10-Gbps Ethernet and FCoE-capable mLOM designed exclusively for the M4 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

## NetApp A-Series All Flash FAS

With the new NetApp A-Series All Flash FAS (AFF) controller lineup, NetApp provides industry-leading performance while continuing to provide a full suite of enterprise-class data management and data protection features. The A-Series AFF lineup offers twice the number of I/O operations per second (IOPS), while decreasing latency. The A-Series AFF lineup includes the A200, A300, A700, and A700s controllers. These controllers and their specifications are listed in Table 2. For more information about the A-Series AFF controllers, see:

- http://www.netapp.com/us/products/storage-systems/all-flash-array/aff-a-series.aspx
- https://hwu.netapp.com/Controller/Index

**Table 2.** NetApp A-Series AFF controller specifications

| | AFF A200 | AFF A300 | AFF A700 | AFF A700s |
|---|---|---|---|---|
| NAS scale-out | 2 to 8 nodes | 2 to 24 nodes | 2 to 24 nodes | 2 to 24 nodes |
| SAN scale-out | 2 to 8 nodes | 2 to 12 nodes | 2 to 12 nodes | 2 to 12 nodes |
| **Specifications per high-availability pair (active-active dual controller)** | | | | |
| Maximum number of SSDs | 144 | 384 | 480 | 216 |
| Maximum raw capacity | 2.2 petabytes (PB) | 5.9 PB | 7.3 PB | 3.3 PB |
| Effective capacity | 8.8 PB | 23.8 PB | 29.7 PB | 13 PB |
| Chassis form factor | 2RU chassis with two high-availability controllers and 24 SSD slots | 3RU chassis with two high-availability controllers | 8RU chassis with two high-availability controllers | 4RU chassis with two high-availability controllers and 24 SSD slots |

This solution uses the NetApp AFF A300, shown in Figure 9 and Figure 10. This controller provides the high-performance benefits of 40 Gigabit Ethernet and all-flash SSDs, offering better performance than previous models and occupying only 3RU of rack space compared to 6RU with the AFF8040. When combined with the 2RU disk shelf of 3.8-TB disks, this solution can provide ample horsepower and over 90 TB of raw capacity, while occupying only 5RU of valuable rack space. These features makes it an excellent controller for a shared-workload converged infrastructure. The A700s is an excellent fit for situations in which more performance is needed.

The FlexPod reference architecture supports a variety of NetApp FAS controllers, such as FAS9000, FAS8000, FAS2600, and FAS2500; A-Series AFF platforms such as AFF8000; and traditional NetApp storage.

For more information about the A-Series AFF product family, see http://www.netapp.com/us/products/storage-systems/all-flash-array/aff-a-series.aspx.

**Note:** The 40 Gigabit Ethernet cards are installed in expansion slot 2 and the ports are e2a and e2e.

**Figure 9.** NetApp AFF A300 front view

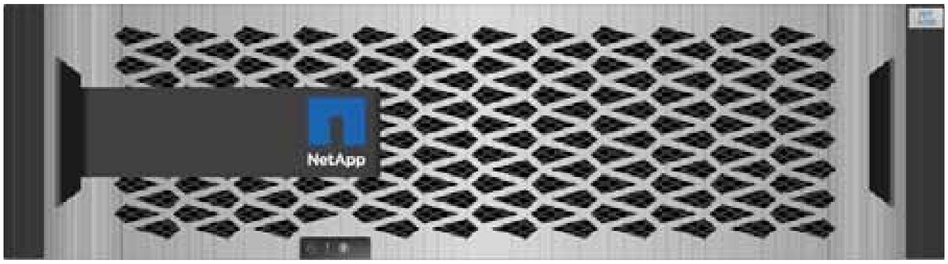
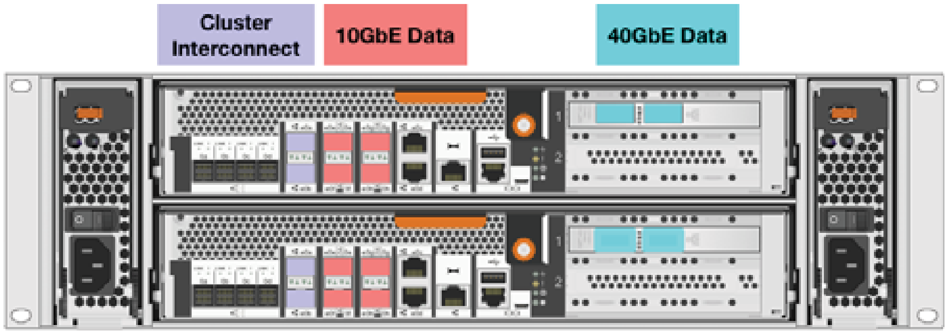


**Figure 10.** NetApp AFF A300 rear view



## NetApp ONTAP

NetApp provides a scalable, unified storage and data management solution. This NetApp solution provides the following benefits:

- **Storage efficiency.** Significant cost savings due to multiple levels of storage efficiency on all VMs.
- **Performance.** An enhanced user experience with the Virtual Storage Tier (VST) and write I/O optimization that complements NetApp storage efficiency.
- **Operational agility.** Enhanced Citrix XenDesktop solution management with tight partner integration.
- **Data protection.** Enhanced protection of virtual desktop OS data and user data with very low overhead in terms of cost and operational data components.

### Multiprotocol Support

NetApp ONTAP 9.3

You can group highly available node pairs together to form a scalable cluster. Creating a cluster allows the nodes to pool their resources and distribute work across the cluster while presenting administrators with a single management entity. Clustering also enables continuous service to end users if individual nodes go offline.

A cluster can contain up to 24 nodes for NAS-based clusters or up to 10 nodes if it contains a storage virtual machine (SVM) with an Infinite Volume. A cluster can also contain up to eight nodes for SAN based clusters. Each node in the cluster can view and manage the same volumes as any other node in the cluster. The total file system namespace, which includes all the volumes and their resultant paths, spans the cluster. For more information, see the NetApp New Features blog.

Adaptive quality of service

There are many features in ONTAP 9.3 that are beneficial to Citrix VDI deployments and they are documented in the base Cisco Validated Design for this whitepaper. The one ONTAP feature that we want to focus on for 3D graphics is Adaptative Quality of Service (AQoS). You can use storage quality of service (QoS) to guarantee that performance of critical workloads is not degraded by competing workloads. You can set a throughput ceiling on a competing workload to limit its impact on system resources, or set a throughput floor for a critical workload, ensuring that it meets minimum throughput targets, regardless of demand by competing workloads. You can even set a ceiling and floor for the same workload.

This is important in 3D graphics because normally the 3D graphic users are a smaller subset of the overall Citrix VDI deployment. The 3D graphic files are large in nature and their applications can be resource intensive. Therefore, you would want to guarantee the quality of service for your graphic user community. You do not want to have a rogue run-away VDI session impact the 3D graphic users. As an example: you could have a situation where Database programmers on the other side of the globe kick off Database dumps to refresh their testing environments at their nightly time zone. Even though it is at night for the database programmers, the database dump process may occur during the production hours of the graphic users on the other side of the globe and could greatly impact there storage performance.  AQoS will ensure that the graphic users have the storage resources that is required regardless of impact from other VDI sessions, even if it is database programmer dumping the database to refresh their desktop testing environments.

Adaptive QoS automatically scales the policy group value to workload size, maintaining the ratio of IOPS to TBs|GBs as the size of the workload changes. That is a significant advantage when you are managing hundreds or thousands of workloads in a large deployment.

You typically use adaptive QoS to adjust throughput ceilings, but you can also use it to manage throughput floors (when workload size increases). For throughput ceilings, workload size is expressed as either the allocated space for the storage object or the space used by the storage object. For throughput floors, only the allocated space is relevant:

- An allocated space policy maintains the IOPS/TB|GB ratio according to the nominal size of the storage object. For example, if the ratio is 100 IOPS/GB, a 150 GB volume will have a throughput ceiling of 15,000 IOPS for as long as the volume remains that size. If the volume is resized to 300 GB, adaptive QoS adjusts the throughput ceiling to 30,000 IOPS.
- A used space policy (the default) maintains the IOPS/TB|GB ratio according to the amount of actual data stored before storage efficiencies. If the ratio is 100 IOPS/GB, a 150 GB volume that has 100 GB of data stored would have a throughput ceiling of 10,000 IOPS. As the amount of used space changes, adaptive QoS adjusts the throughput ceiling according to the ratio

You can expect the following behavior for both throughput ceilings and floors:

- When a workload is assigned to an adaptive QoS policy group, the ceiling or floor is updated immediately.
- When a workload in an adaptive QoS policy group is resized, the ceiling or floor is updated in approximately five minutes.

Throughput must increase by at least 10 IOPS before updates take place.

### Other ONTAP features

There are many other NetApp ONTAP features that we used during this reference architecture and if you want to learn more, please see the base Cisco Validate Design for this whitepaper.

## NVIDIA GRID cards

For desktop virtualization applications, the NVIDIA Tesla P6, P4, and P40 cards are optimal choices for high-performance graphics. Table 3 lists the technical specifications.

Table 3. Technical specifications for NVIDIA GRID cards

| | P6 | P4 | P40 |
|---|---|---|---|
| **Number of GPUs** | Single NVIDIA Pascal | Single NVIDIA Pascal | Single NVIDIA Pascal |
| **NVIDIA Compute Unified Device Architecture (CUDA) cores** | 2048 | 2560 | 3840 |
| **Memory size** | 16-GB GDDR5 | 8-GB GDDR5 | 24-GB GDDR5 |
| **Maximum number of vGPU instances** | 16 (1-GB profile) | 8 (1-GB profile) | 24 (1-GB profile) |
| **Power** | 90 watts (W) | 50 to 75W | 250W |
| **Form factor** | Mobile PCIe Module (MXM), for blade servers, with x16 lanes | PCIe 3.0 single slot (low profile), for rack servers, with x16 lanes | PCIe 3.0 dual slot, for rack servers, with x16 lanes |
| **Cooling solution** | Bare board | Passive | Passive |
| **H.264 1080p30 streams** | 24 | 24 | 24 |
| **Maximum number of users per board** | 16 (1-GB profile) | 8 (1-GB profile) | 24 (1-GB profile) |

## NVIDIA GRID

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

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

### NVIDIA GRID 6.2 GPU

The NVIDIA GRID solution runs on Tesla GPUs based on NVIDIA Volta, NVIDIA Pascal, and NVIDIA Maxwell architectures. These GPUs come in two server form factors: the NVIDIA Tesla P6 for blade servers and converged infrastructure, and the NVIDIA Tesla P4 and P40 for rack servers.

### NVIDIA GRID 6.2 license requirements

NVIDIA GRID 6.2 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 6.2 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 6.2 license requirements, see the NVIDIA documentation.

## Citrix XenServer 7.5

Citrix XenServer is a complete server virtualization platform from Citrix. The XenServer package contains all you need to create and manage a deployment of virtual x86 computers running on Citrix Xen, the open-source paravirtualizing hypervisor with near-native performance. XenServer is optimized for both Windows and Linux virtual servers.

XenServer 7 extends its leadership as the most mature hypervisor for virtual graphics capabilities, the hypervisor most integrated with Citrix XenApp and XenDesktop, and revolutionary security capabilities available only with XenServer.

The XenServer 7.5 platform offers mature graphics capabilities. XenServer is the only hypervisor to support live migration of GPU-enabled virtual machines with vGPU XenMotion at the time this document was written. vGPU XenMotion improves the user experience and administrative flexibility by enabling administrators to rebalance GPU-enabled virtual machines across pool hosts to improve virtual machine performance. vGPU XenMotion also enhances user productivity, enabling users to remain productive during unexpected maintenance operations.

Another feature of XenServer 7 is increased administrative flexibility. XenServer 7 offers new features that simplify snapshot lifecycle management, accelerate VMware vSphere migration, and greatly improve the administrative experience.

- Simplify administration with snapshot lifecycle management. Now you can easily schedule hourly, daily, weekly, or monthly snapshots for a group of virtual machines. The platform also aids in the management of your snapshot lifecycles. XenServer allows you to set the maximum number of snapshots to keep, deleting the oldest snapshot in a rolling fashion.
- Live-patch hosts without downtime, available only with XenServer. Live patching is an industry-first hypervisor capability brought to you with XenServer 7.1. It greatly reduces operational overhead by enabling IT administrators to hot-patch (update) their active XenServers without rebooting. This capability is not found on any other commercial hypervisor and has been used in production by a large cloud computing provider for months.
- Simplify administration and reduce overhead with new automated updates. Simplify patching at scale with new automated updates. XenServer now supports the automated application of fixes to multiple hosts. This new automated update technology batch-patches multiple hosts by automatically downloading the necessary fixes from Citrix, installing these patches in the correct order, and rebooting the hosts in sequence while redistributing virtual machines to prevent outages.
- Get greater migration flexibility with Linux virtual machine conversion. XenServer Conversion Manager now offers the capability to migrate Linux virtual machines from other hypervisors to XenServer. This feature offers greater migration flexibility for customers wanting to migrate from VMware solutions. Customers wanting to deliver secure Linux virtual desktops and applications from XenDesktop can avoid the vTax by using XenServer.

- Increase scalability and simplify administration with increased pool sizes. XenServer 7.5 quadruples the maximum pool size to 64 hosts. The benefits of deploying larger pools include fewer pools to manage, more flexible in-pool migration with shared storage, more flexibility with high availability, and the need for fewer machine catalogs for large XenApp and XenDesktop deployments. This feature saves you time and administrative effort during your image update processes and ongoing maintenance.
- Extend XenApp and XenDesktop integration. XenServer 7 extends alignment with XenApp and XenDesktop release cycles, and the expanded XenServer Enterprise entitlement helps ensure that you can use unique application and desktop performance integrations available only with XenServer.
- Support the Microsoft Windows Continuum experience. With patent-pending technology from XenServer, XenDesktop is the only virtual desktop infrastructure (VDI) platform that enables the Windows Continuum experience with Windows 10 virtual desktops. XenDesktop allows Windows 10 virtual desktops to automatically toggle between tablet and desktop mode, in real-time, as the state of the hardware changes, to provide the most native Windows 10 experience.
- Accelerate Citrix Provisioning Services performance with XenServer. The new Provisioning Services (PVS) Accelerator technology results in up to 25 percent faster desktop boot times, up to 98 percent less network bandwidth use, and up to 93 percent less PVS CPU use. These capabilities are available only with XenServer.
- Enhance support with the first-ever long-term service release (LTSR), for XenServer 7.1. The XenServer LTSR enables up to 10 years of support (5 years of mainstream support and 5 years of extended support), making it the only hypervisor to have product lifecycle dates fully aligned with Citrix XenApp and XenDesktop, which can radically simplify infrastructure maintenance for customers choosing to deploy a full Citrix solution.
- Extend the benefits of XenServer Enterprise to all XenApp and XenDesktop deployments. XenApp and XenDesktop customers are now entitled to all the features of XenServer Enterprise edition, regardless of their XenApp and XenDesktop license type. Additionally, all editions of XenApp and XenDesktop are now entitled to features like such as Machine Creation Services (MCS) Accelerator technology, previously a platinum-only feature.

### 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.

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

### GPU acceleration for Microsoft Windows desktops

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

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

Using GPU virtualization, multiple virtual machines can directly access the graphics processing power of a single physical GPU. The true hardware GPU sharing provides desktops suitable for users with complex and demanding design requirements. GPU virtualization for NVIDIA Tesla cards use the same NVIDIA graphics drivers that are deployed on bare-metal operating systems with NVIDIA Quadro desktop and workstation cards.

HDX 3D Pro offers the following features:

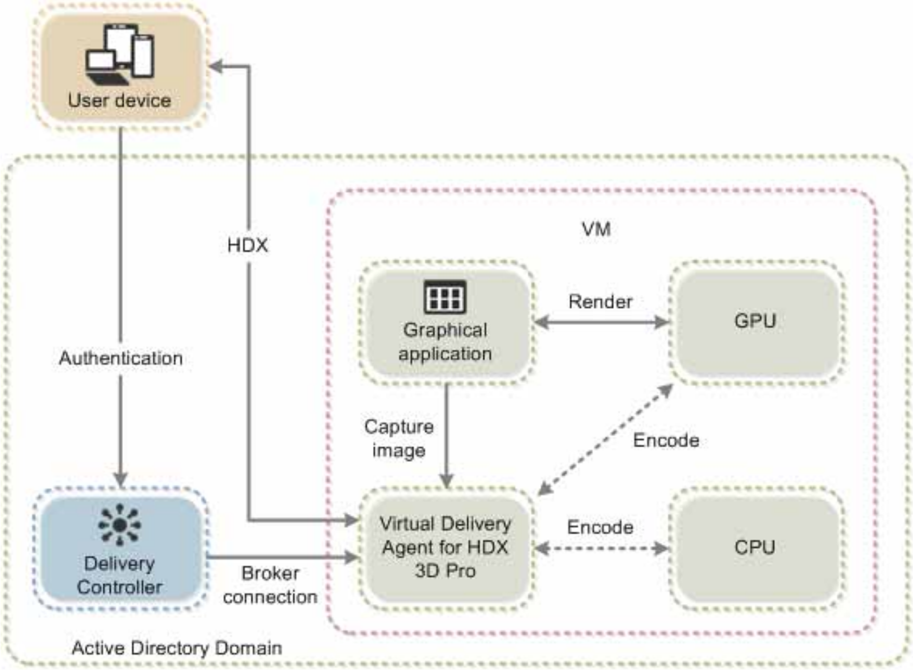
- Adaptive H.264-based deep compression for optimal WAN and wireless performance: Citrix HDX 3D Pro uses CPU-based full-screen H.264 compression as the default compression technique for encoding. Hardware encoding is used with NVIDIA cards that support NVIDIA NVENC.
- Lossless compression option for specialized use cases: Citrix 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, Citrix HDX 3D Pro uses the lossless codec for compression regardless of the codec selected through policy.
    - For lossy compression, Citrix 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, Citrix 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 VMware vSphere and ESX using virtual direct graphics acceleration (vDGA): You can use Citrix HDX 3D Pro with vDGA for both remote desktop service (RDS) and VDI workloads. When you use Citrix HDX 3D Pro with virtual shared graphics acceleration (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.

Note the following details, 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 Citrix HDX 3D Pro to broker a connection to the computer hosting the graphical application.
- The VDA for Citrix 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. With OpenGL, DirectX, Direct3D, and Windows Presentation Foundation (WPF) rendering moved 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 Citrix XenServer, VMware vSphere vDGA, and Intel GVT-d. 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 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, Citrix 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:

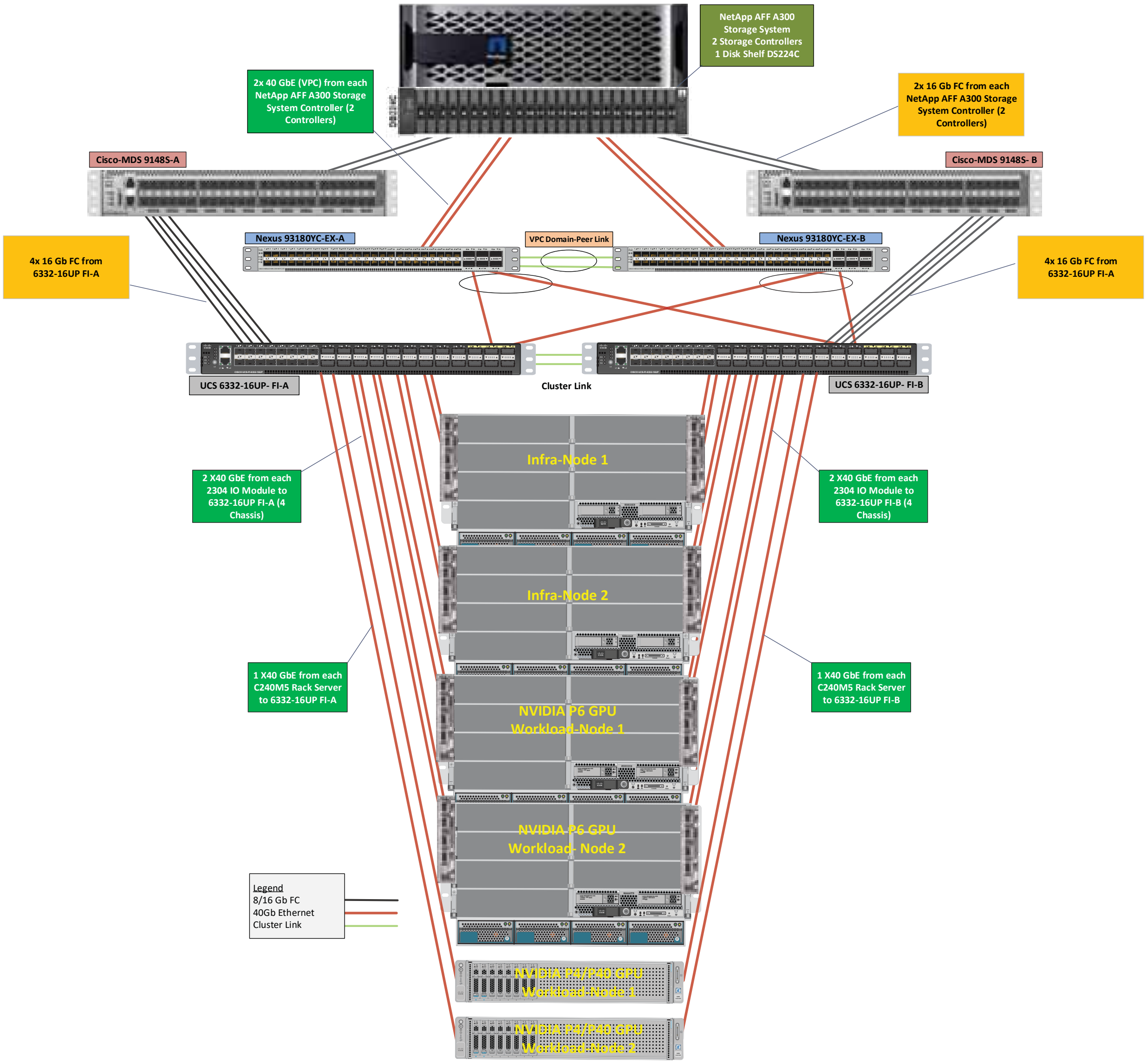
- Citrix HDX 3D Pro supports all monitor resolutions that are supported by the GPU on the host computer. However, for optimal performance with the minimum recommended user device and GPU specifications, Citrix recommends a maximum monitor resolution for user devices of 1920 x 1200 pixels for LAN connections,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

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

## Solution configuration

Figure 12 provides an overview of the physical connectivity configuration of the FlexPod Datacenter solution. The solution is described in a great detail in the Cisco Validated Design FlexPod Datacenter with Citrix XenDesktop and XenApp 7.15 and VMware vSphere 6.5 Update 1 for 6000 Seats. This architecture was used to validate Tesla NVDIA graphic cards using SPECviewperf 13, Citrix XenDesktop HDX 3D Pro, and the Citrix XenServer hypervisor, and to create this document.

**Figure 12.** Cabling diagram for a FlexPod Datacenter with Cisco UCS


NetApp AFF A300
Storage System
2 Storage Controllers
1 Disk Shelf DS224C
2x 40 GbE (VPC) from each NetApp AFF A300 Storage System Controller (2 Controllers)
2x 16 Gb FC from each NetApp AFF A300 Storage System Controller (2 Controllers)
Cisco-MDS 9148S-A
Cisco-MDS 9148S- B
Nexus 93180YC-EX-A
VPC Domain-Peer Link
Nexus 93180YC-EX-B
4x 16 Gb FC from 6332-16UP FI-A
4x 16 Gb FC from 6332-16UP FI-A
UCS 6332-16UP- FI-A
Cluster Link
UCS 6332-16UP- FI-B
2 X40 GbE from each 2304 IO Module to 6332-16UP FI-A (4 Chassis)
2 X40 GbE from each 2304 IO Module to 6332-16UP FI-B (4 Chassis)
Infra-Node 1
Infra-Node 2
1 X40 GbE from each C240M5 Rack Server to 6332-16UP FI-A
1 X40 GbE from each C240M5 Rack Server to 6332-16UP FI-B
NVIDIA P6 GPU
Workload-Node 1
NVIDIA P6 GPU
Workload- Node 2
Legend
8/16 Gb FC
40Gb Ethernet
Cluster Link
NVIDIA P4/P40 GPU
Workload-Node 1
NVIDIA P4/P40 GPU
Workload-Node 2

The hardware components in the solution are as follows:

- Cisco UCS B200 M5 Blade Servers with Intel Xeon Silver 4114 2.20-GHz 10-core processors and 768-GB 2666-MHz RAM for infrastructure
- Cisco UCS B200 M5 Blade Servers with Intel Xeon Gold 6140 2.30-GHz 18-core processors, 768-GB 2666-MHz RAM, and two NVIDIA Tesla P6 GPUs for graphics accelerated virtual client computing workloads
- Cisco UCS C240 M5 Rack Servers with Intel Xeon Gold 6140 2.30-GHz 18-core processors, 768-GB 2666-MHz RAM, and six NVIDIA Tesla P4 or P40 GPUs for graphics accelerated virtual client computing workloads
- Cisco UCS VIC 1387 mLOM (Cisco UCS C240 M5)
- Cisco UCS VIC 1340 mLOM (Cisco UCS B200 M5)

- NetApp AFF A300 system array used for all data
- Cisco Nexus® 93180YC-FX Switches in Cisco NX-OS mode for Layer 2 communications
- Cisco MDS 9148S 16G Multilayer Fabric Switch for Fibre Channel connectivity

The software components of the solution are:

- Cisco UCS Firmware Release 3.2(3d)
- Citrix XenServer 7.5 for VDI hosts
- Citrix XenDesktop 7.15
- Microsoft Windows 10 64-bit
- Microsoft Server 2016
- SPECviewperf 13 graphics benchmark software and commercial license
- NVIDIA GRID 6.2 software and licenses:
  - NVIDIA-vGPU-xenserver-7.5-390.72.x86_64.rpm
  - 391.81_grid_win10_server2016_64bit_international

## Configure Cisco UCS

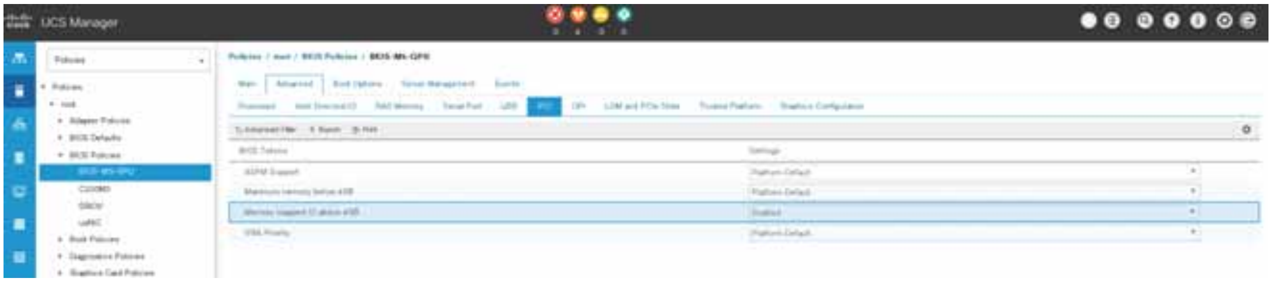
This section describes the Cisco UCS configuration.

### Create BIOS policy

Create a new BIOS policy.

1. Right-click BIOS Policy. On the Advanced tab for the new BIOS policy
2. Click PCI. For "Memory mapped IO above 4GB," choose Enabled (Figure 13).

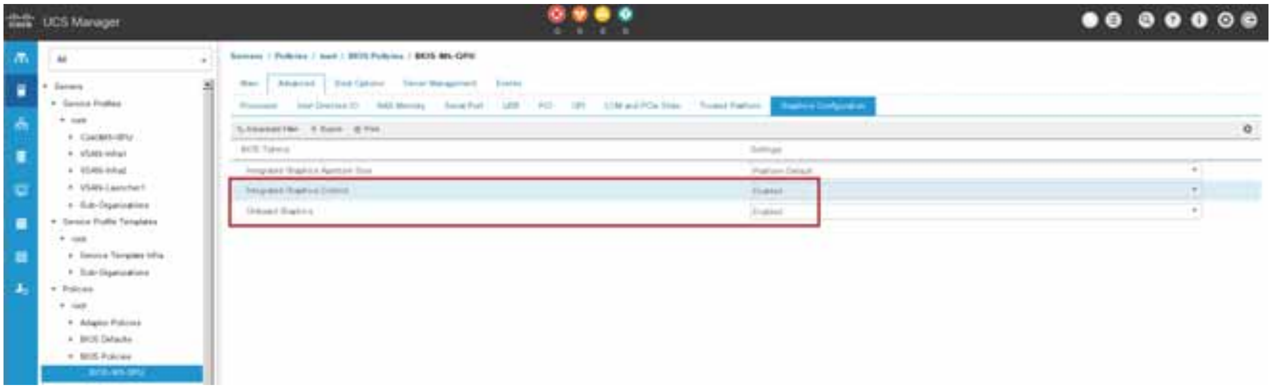
**Figure 13.** PCI setting for BIOS policy: Enable MMIO above 4 GB



| BIOS Setting | Settings |
|---|---|
| ASPM Support | Platform Default |
| Maximum memory below 4GB | Platform Default |
| Memory mapped IO above 4GB | Enabled |
| VGA Priority | Platform Default |

3. Click Graphics Configuration (Figure 14):
   - For Integrated Graphics Control, choose Enabled.
   - For Onboard Graphics, choose Enabled.

**Figure 14.** PCI BIOS policy configuration



Servers / Policies / root / BIOS Policies / BIOS-M5-GPU

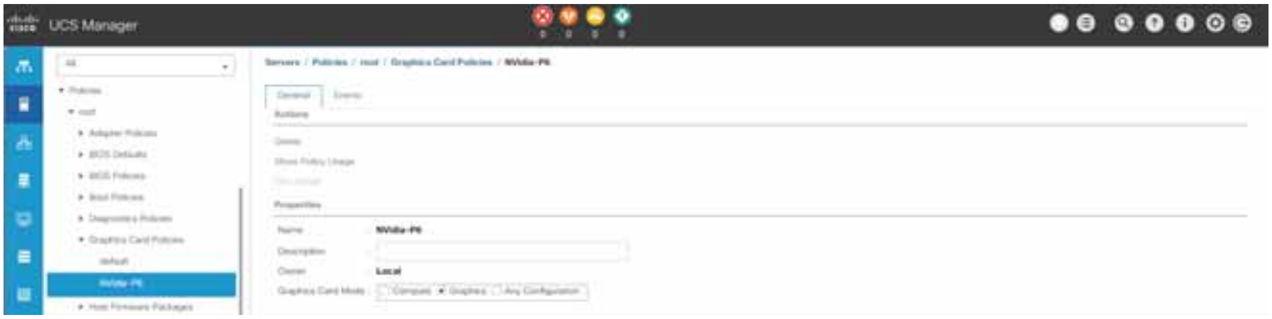
| BIOS Tokens | Settings |
|---|---|
| Integrated Graphics Aperture Size | Platform Default |
| Integrated Graphics Control | Enabled |
| Onboard Graphics | Enabled |

## Create graphics card policy

Create a new graphics card policy with the preferred mode of graphics card.

For the VDI deployment described here, select Graphics mode (Figure 15).

**Figure 15.** Graphics card policy



Servers / Policies / root / Graphics Card Policies / NVidia-P6

Name: NVidia-P6
Description:
Owner: Local
Graphics Card Mode: ○ Compute ● Graphics ○ Any Configuration

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

Install the NVIDIA Tesla GPU card on the Cisco UCS B200 M5 server using one of the methods described here.

### Physically installing a P6 card in the Cisco UCS B200 M5 server

The NVIDIA P6 GPU card provides graphics and computing capabilities to the server. There are two supported versions of the NVIDIA P6 GPU card:

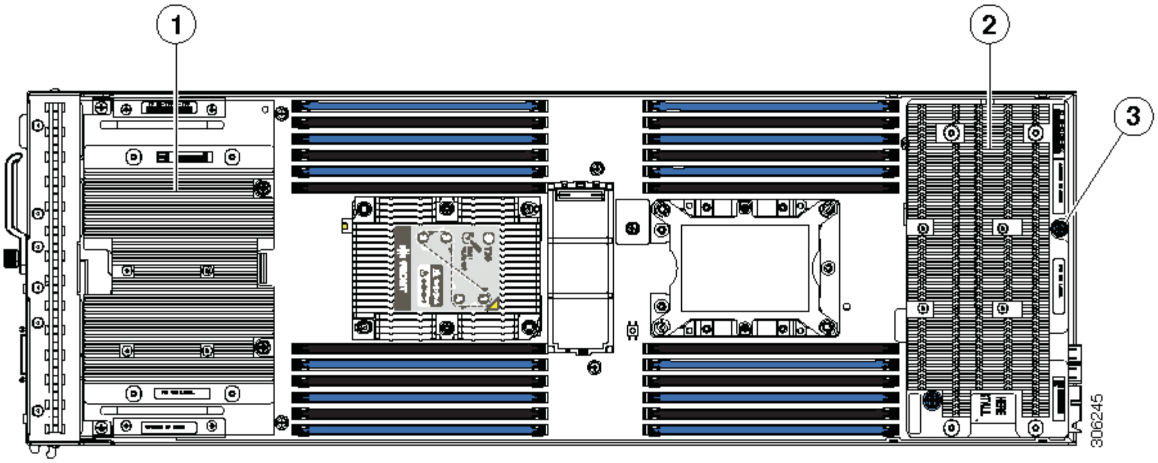
- The UCSB-GPU-P6-F card can be installed only in the front mezzanine slot of the server.

**Note:** No front mezzanine cards can be installed when the server has CPUs using greater than 165W.

- The UCSB-GPU-P6-R can be installed only in the rear mezzanine slot (slot 2) of the server.

Figure 16 shows the installed NVIDIA P6 GPU in the front and rear mezzanine slots.

**Figure 16.** NVIDIA GPU installed in the front and rear mezzanine slots

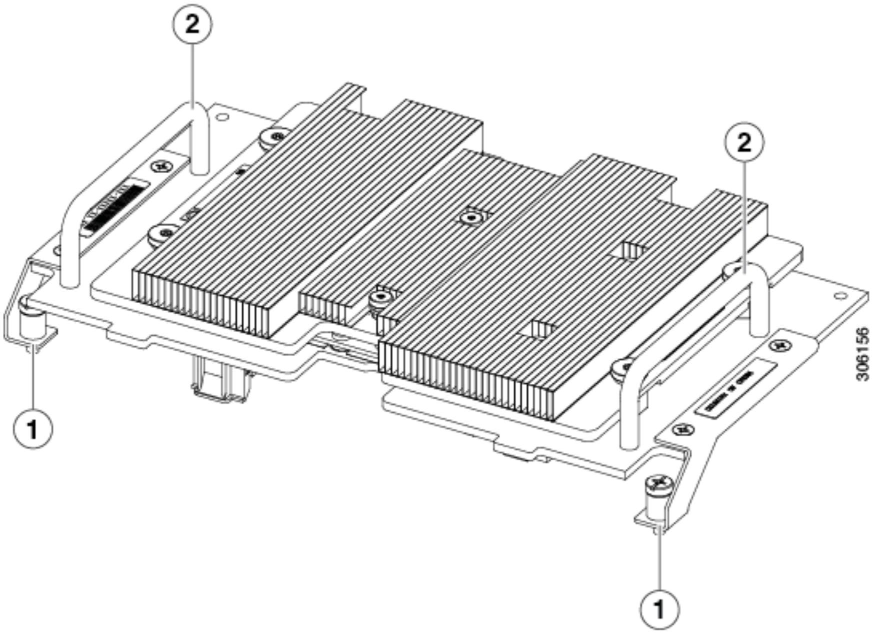


| | | | |
|---|---|---|---|
| 1 | Front GPU | 2 | Rear GPU |
| 3 | Custom standoff screw | - | |

## Installing an NVIDIA GPU card in the front of the server

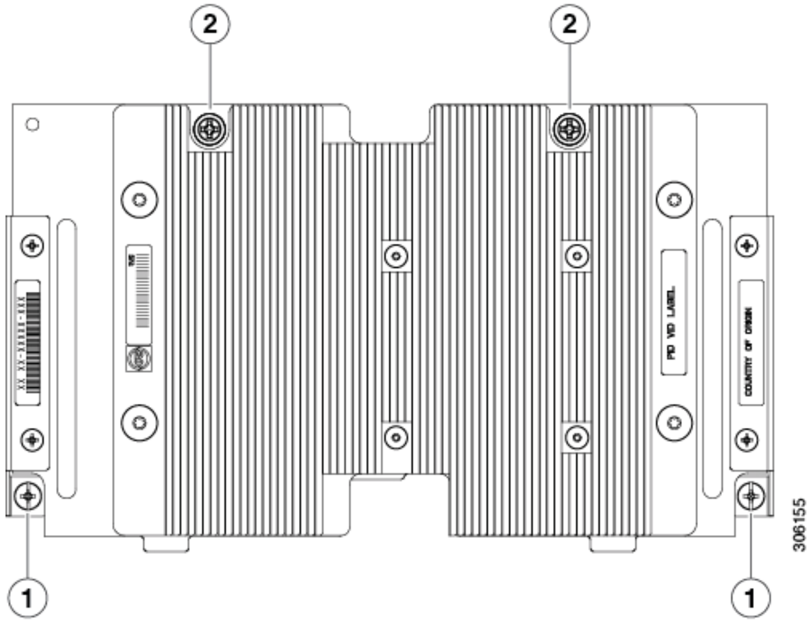
Figure 17 shows the front NVIDIA P6 GPU (UCSB-GPU-P6-F), and Figure 18 shows the top view.

**Figure 17.** NVIDIA P6 GPU that installs in the front of the server



| | | | |
|---|---|---|---|
| 1 | Leg with thumb screw that attaches to the server motherboard at the front | 2 | Handle to press down on when installing the GPU |

**Figure 18.** Top view of the NVIDIA P6 GPU in the front of the server



| | | | |
|---|---|---|---|
| **1** | Leg with thumb screw that attaches to the server motherboard | **2** | Thumb screw that attaches to a standoff below |

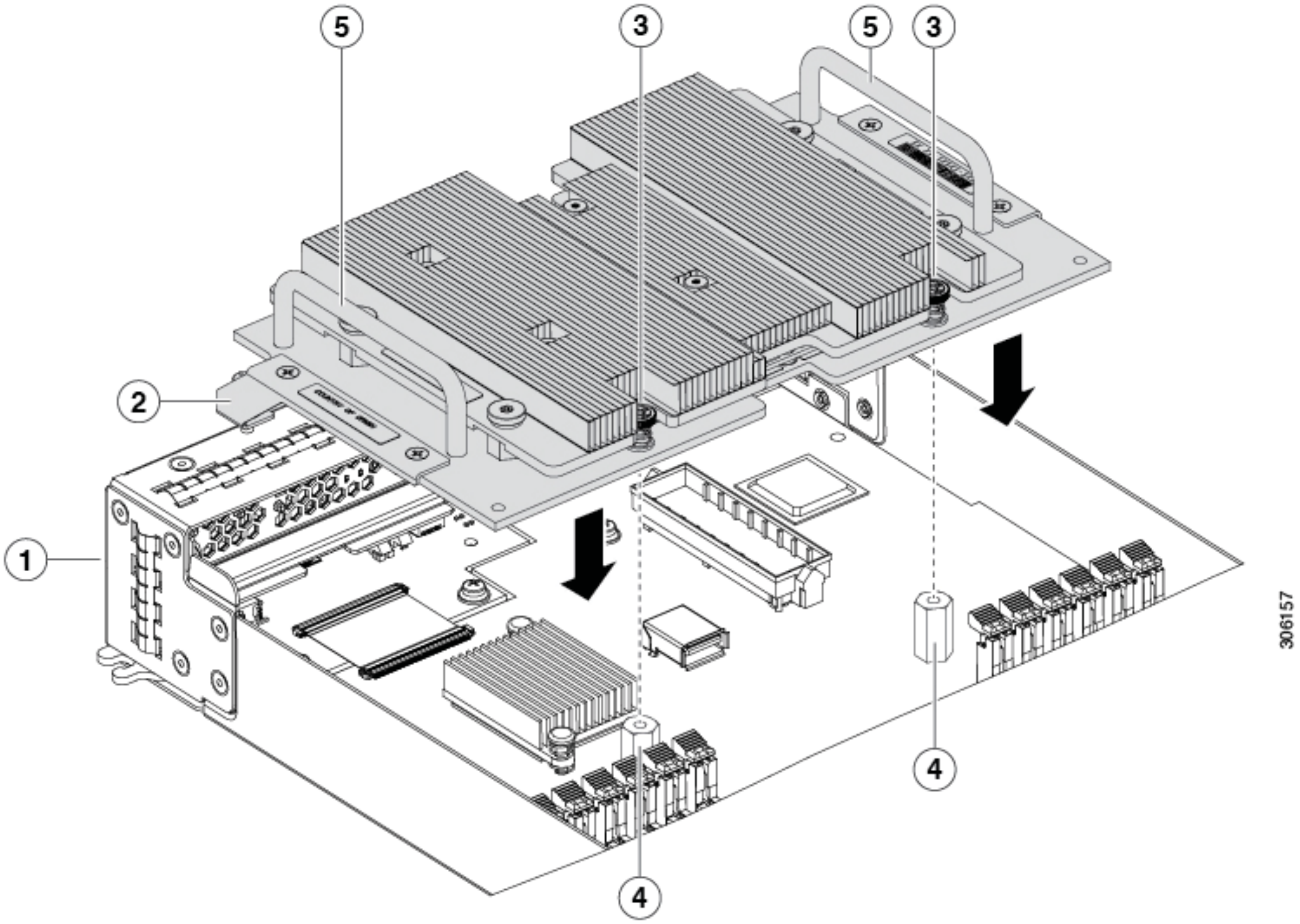
To install the NVIDIA P6 GPU, follow the steps presented here.

**Note:** Before installing the NVIDIA P6 GPU (UCSB-GPU-P6-F) in the front mezzanine slot, do the following:

- Upgrade the Cisco UCS domain that the GPU will be installed into to a version of Cisco UCS Manager that supports this card. Refer to the latest version of the release notes for Cisco UCS software for information about supported hardware: http://www.cisco.com/c/en/us/support/servers-unified-computing/ucs-manager/products-release-notes-list.html.
- Remove the front mezzanine storage module if it is present. You cannot use the storage module in the front mezzanine slot when the NVIDIA P6 GPU is installed in the front of the server.

1. Position the GPU in the correct orientation relative to the front of the server (number 1) as shown in Figure 19.
2. Install the GPU in the server. Press down on the handles (number 5) to firmly secure the GPU.
3. Tighten the thumb screws (number 3) at the back of the GPU with the standoffs (number 4) on the motherboard.
4. Tighten the thumb screws on the legs (number 2) of the motherboard.
5. Install the drive blanking panels

**Figure 19.** Installing the NVIDIA GPU in the front of the server

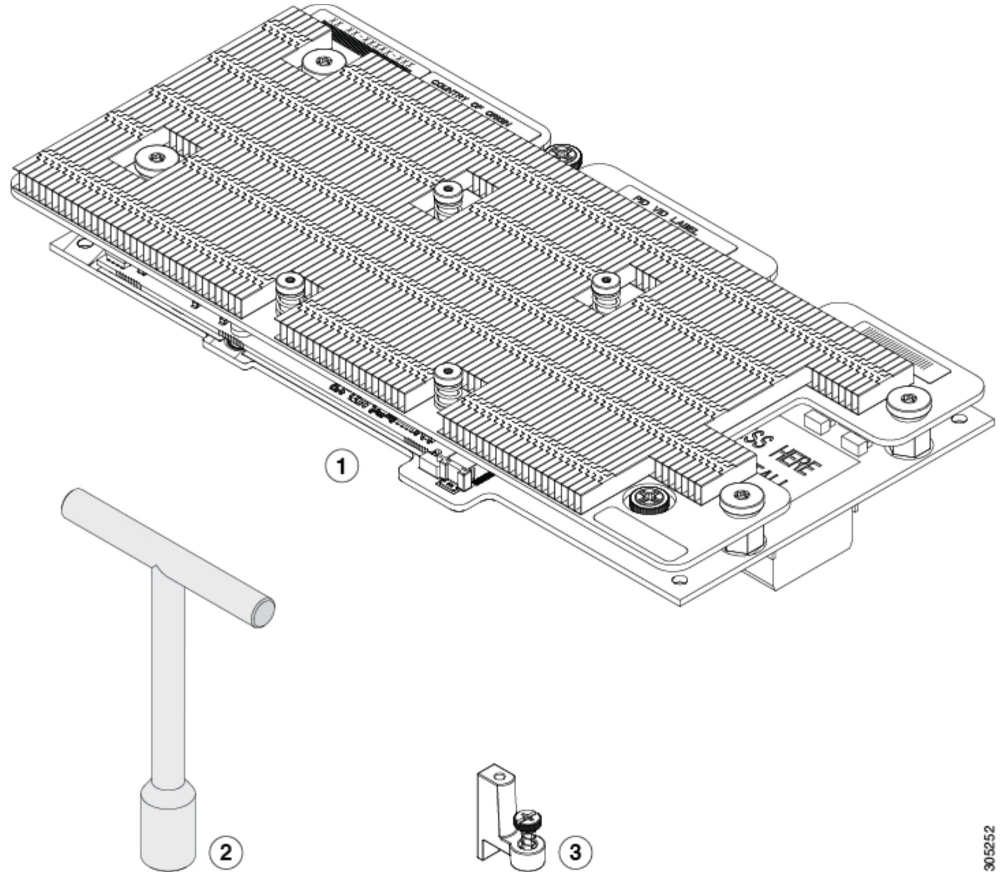


| 1 | Front of the server | 2 | Leg with thumb screw that attaches to the motherboard |
|---|---|---|---|
| 3 | Thumbscrew to attach to standoff below | 4 | Standoff on the motherboard |
| 5 | Handle to press down on to firmly install the GPU | - | |

### Installing an NVIDIA GPU card in the rear of the server

If you are installing the UCSB-GPU-P6-R on a server in the field, the option kit comes with the GPU itself (CPU and heat sink), a T-shaped installation wrench, and a custom standoff to support and attach the GPU to the motherboard. Figure 20 shows the three components of the option kit.

**Figure 20.** NVIDIA P6 GPU (UCSB-GPU-P6-R) option kit



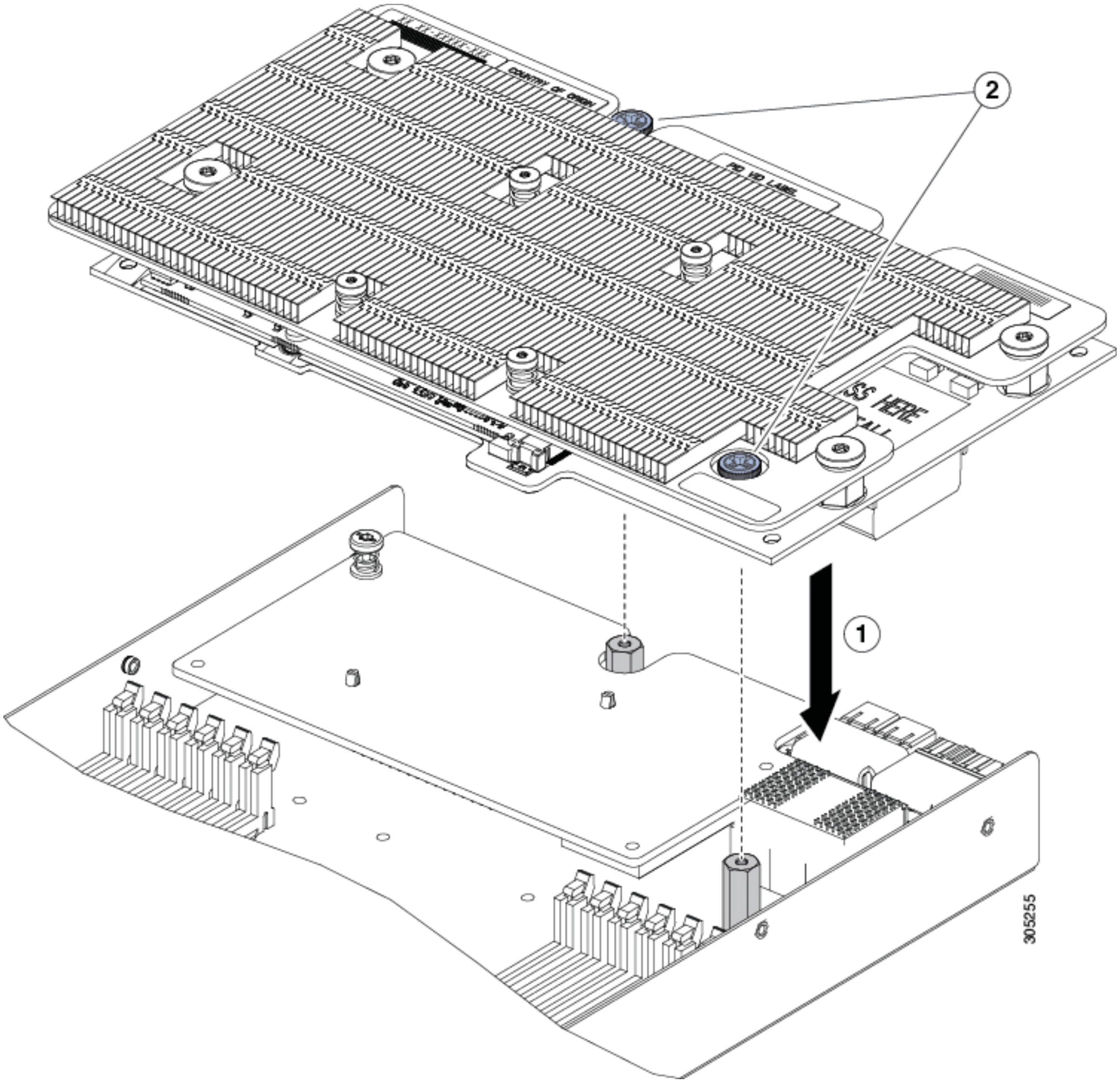
| | | | |
|---|---|---|---|
| **1** | NVIDIA P6 GPU (CPU and heatsink) | **2** | T-shaped wrench |
| **3** | Custom standoff | **-** | |

**Note:** Before installing the NVIDIA P6 GPU (UCSB-GPU-P6-R) in the rear mezzanine slot, do the following:

- Upgrade the Cisco UCS domain that the GPU will be installed into to a version of Cisco UCS Manager that supports this card. Refer to the latest version of the release notes for Cisco UCS software for information about supported hardware: http://www.cisco.com/c/en/us/support/servers-unified-computing/ucs-manager/products-release-notes-list.html.
- Remove any other card, such as a VIC 1480, VIC 1380, or VIC port expander card, from the rear mezzanine slot. You cannot use any other card in the rear mezzanine slot when the NVIDIA P6 GPU is installed.

1. Use the T-shaped wrench that comes with the GPU to remove the existing standoff at the back end of the motherboard.
2. Install the custom standoff in the same location at the back end of the motherboard (Figure 21).
3. Position the GPU over the connector on the motherboard and align all the captive screws with the standoff posts (number 1).
4. Tighten the captive screws (number 2).

**Figure 21.** Installing the NVIDIA P6 GPU in the rear mezzanine slot



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

Install the NVIDIA Tesla GPU card on the Cisco UCS C240 M5 server as described here.

### Physically installing an NVIDIA Tesla P4 card

Use the following procedure to install NVIDIA Tesla P4:

**Note:** This server can support up to six single-wide NVIDIA Tesla P4 GPU cards. These half-height, half-length (HHHL) GPU cards are supported in all PCIe slots.

1. Shut down and remove power from the server.
2. Slide the server out the front of the rack far enough so that you can remove the top cover. You may have to detach cables from the rear panel to provide clearance.

3. Remove the top cover from the server.
4. Install a new single-wide GPU card (Figure 22):

**Note:** Up to six single-wide GPU cards are supported in the PCIe slots.

a. With the hinged card-tab retainer open, align the new single-wide GPU card with the empty socket on the PCIe riser.

b. Push down evenly on both ends of the card until it is fully seated in the socket.

c. Verify that the card's rear panel tab sits flat against the riser rear-panel opening and then close the hinged card-tab retainer over the card's rear-panel tab.

d. Swing the hinged securing plate closed on the bottom of the riser. Verify that the clip on the plate clicks into the locked position.

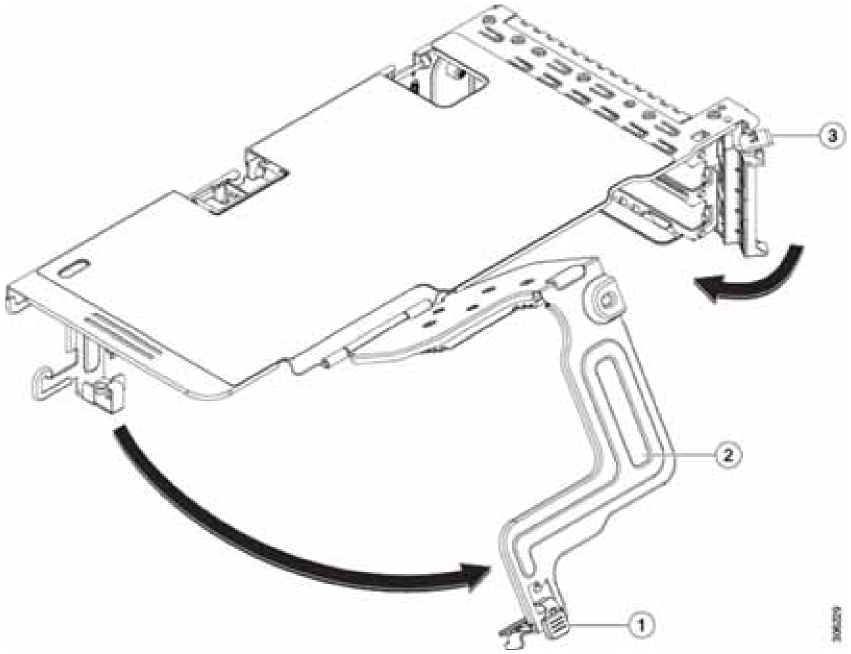
e. Position the PCIe riser over its socket on the motherboard and over the chassis alignment channels.

f. Carefully push down on both ends of the PCIe riser to fully engage its connector with the sockets on the motherboard.

5. Replace the top cover to the server.

6. Replace the server in the rack, replace cables, and then fully power on the server by pressing the Power button.

**Figure 22.** PCIe riser card securing mechanism



| 1 | Release latch on hinged securing plate | 2 | Hinged card-tab retainer |
|---|---|---|---|
| 3 | Hinged securing plate | - | |

### Installing a double-wide GPU card

Use the following procedure to install NVIDIA Tesla P40 card:

1. Shut down and remove power from the server.
2. Slide the server out the front of the rack far enough so that you can remove the top cover. You may have to detach cables from the rear panel to provide clearance.

**Note:** If you cannot safely view and access the component, remove the server from the rack.

3. Remove the top cover from the server.
4. Install a new GPU card:

**Note:** Observe the configuration rules for this server, as described in GPU Card Configuration Rules.

g. Align the GPU card with the socket on the riser, and then gently push the card’s edge connector into the socket. Press evenly on both corners of the card to avoid damaging the connector.

h. Connect the GPU power cable. The straight power cable connectors are color-coded. Connect the cable's black connector into the black connector on the GPU card and the cable's white connector into the white GPU POWER connector on the PCIe riser.

Note: Do not reverse the straight power cable. Connect the black connector on the cable to the black connector on the GPU card. Connect the white connector on the cable to the white connector on the PCIe riser.

i. Close the card-tab retainer over the end of the card.

j. Swing the hinged securing plate closed on the bottom of the riser. Verify that the clip on the plate clicks into the locked position.

k. Position the PCIe riser over its socket on the motherboard and over the chassis alignment channels.

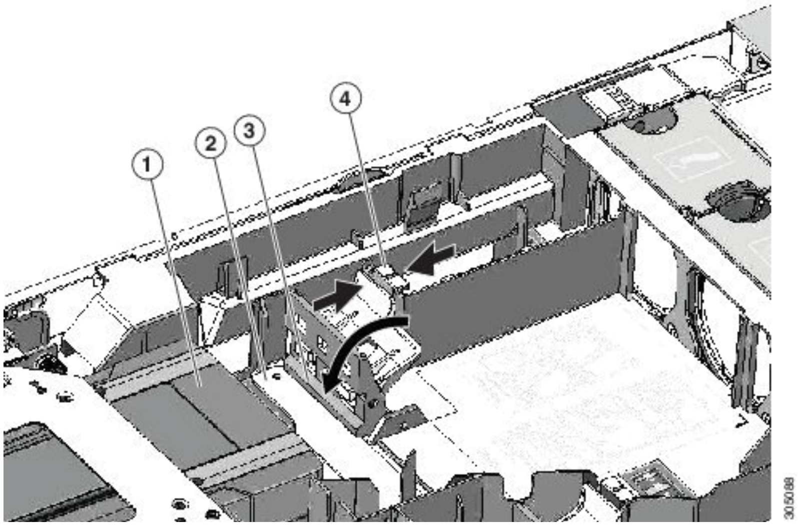
l. Carefully push down on both ends of the PCIe riser to fully engage its connector with the sockets on the motherboard.

m. At the same time, align the GPU front support bracket (on the front end of the GPU card) with the securing latch that is on the server's air baffle.

n. Insert the GPU front support bracket into the latch that is on the air baffle (Figure 23):

- Pinch the latch release tab and hinge the latch toward the front of the server.
- Hinge the latch back down so that its lip closes over the edge of the GPU front support bracket.
- Verify that the latch release tab clicks and locks the latch in place.

**Figure 23.** GPU front support bracket inserted into securing latch on air baffle



| | | | |
|---|---|---|---|
| **1** | Front end of GPU card | **2** | GPU front support bracket |
| **3** | Lip on securing latch | **4** | Securing latch release tab |

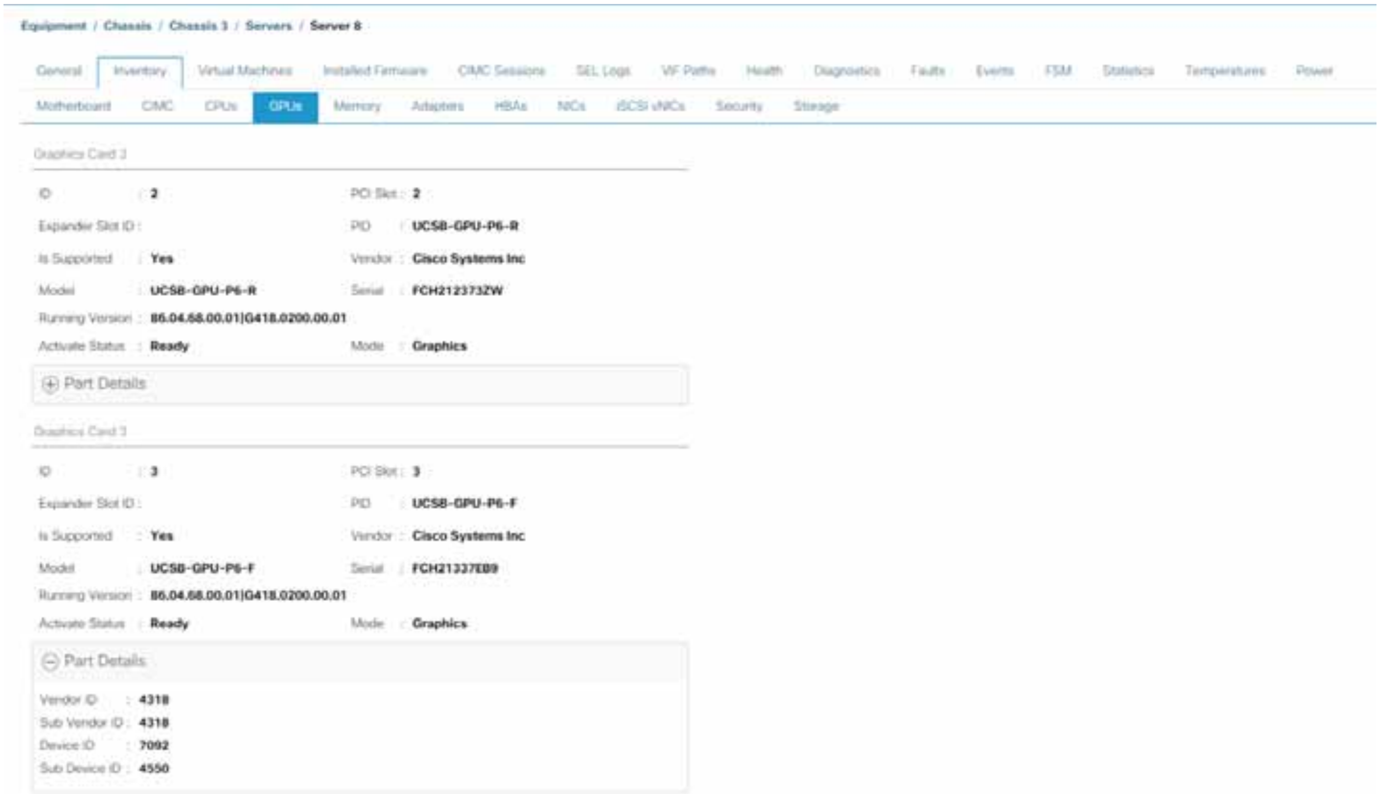
5. Replace the top cover to the server.
6. Replace the server in the rack, replace cables, and then fully power on the server by pressing the Power button.

### 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 24, PCIe slots 2 and 3 are used with two GRID P6 cards.

**Figure 24.** NVIDIA GRID P6 card inventory displayed in Cisco UCS Manager



Graphics Card 2

ID : 2
PCI Slot : 2
Expander Slot ID :
PID : UCSB-GPU-P6-R
Is Supported : Yes
Vendor : Cisco Systems Inc
Model : UCSB-GPU-P6-R
Serial : FCH212373ZW
Running Version : 86.04.68.00.01|G418.0200.00.01
Activate Status : Ready
Mode : Graphics

⊕ Part Details

Graphics Card 3

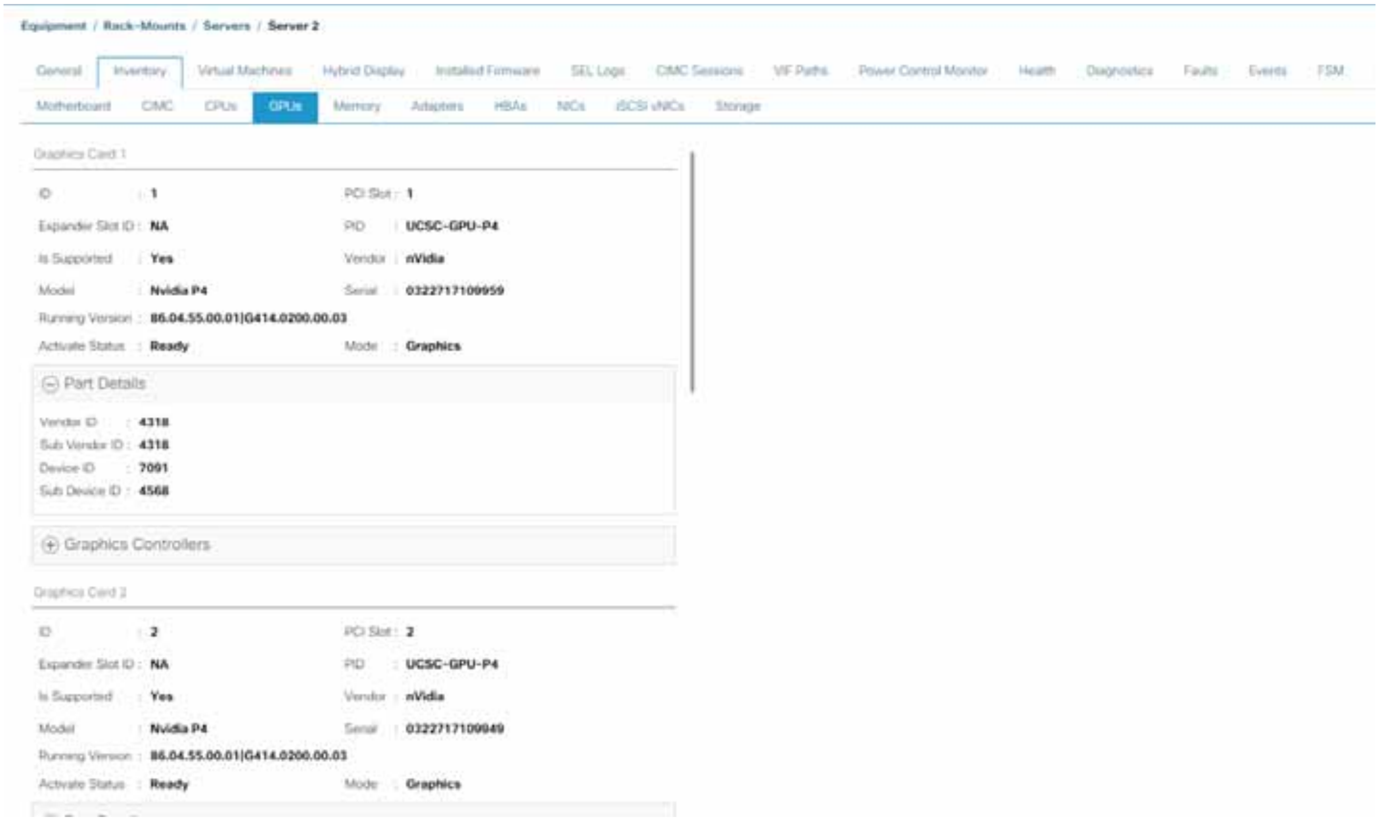
ID : 3
PCI Slot : 3
Expander Slot ID :
PID : UCSB-GPU-P6-F
Is Supported : Yes
Vendor : Cisco Systems Inc
Model : UCSB-GPU-P6-F
Serial : FCH21337EB9
Running Version : 86.04.68.00.01|G418.0200.00.01
Activate Status : Ready
Mode : Graphics

⊖ Part Details

Vendor ID : 4318
Sub Vendor ID : 4318
Device ID : 7092
Sub Device ID : 4550

2. After the NVIDIA P4 GPU cards are physically installed and the Cisco UCS C240 M5 Rack Server is discovered in Cisco UCS Manager, select the server and choose Inventory > GPUs. As shown in Figure 25, PCIe slots 2 and 5 are used with two GRID P4 cards.

**Figure 25.** NVIDIA GRID P4 card inventory displayed in Cisco UCS Manager



Graphics Card 1

ID : 1 PCI Slot : 1
Expander Slot ID : NA PID : UCSC-GPU-P4
Is Supported : Yes Vendor : nVidia
Model : Nvidia P4 Serial : 0322717109959
Running Version : 86.04.55.00.01|G414.0200.00.03
Activate Status : Ready Mode : Graphics

⊖ Part Details

Vendor ID : 4318
Sub Vendor ID : 4318
Device ID : 7091
Sub Device ID : 4568

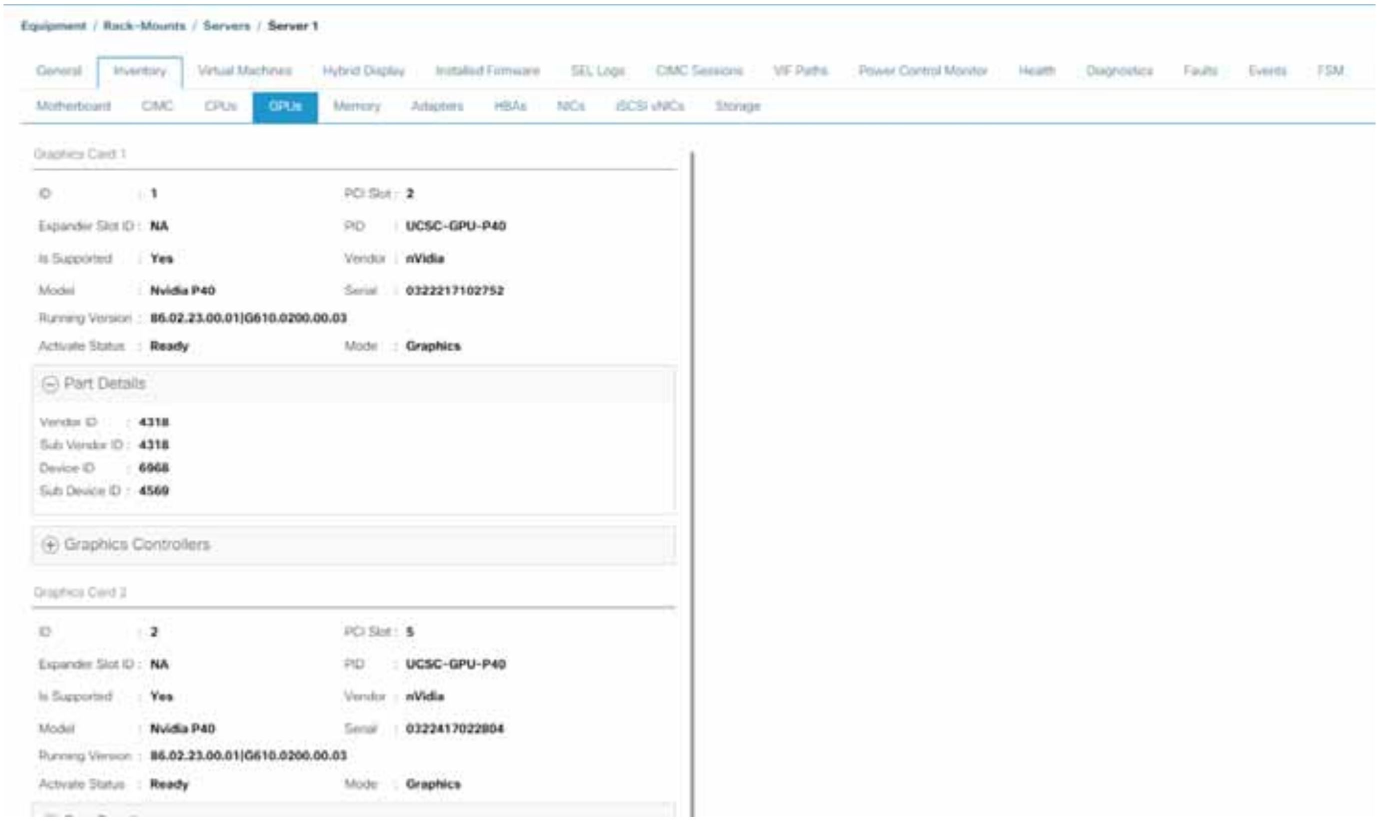
⊕ Graphics Controllers

Graphics Card 2

ID : 2 PCI Slot : 2
Expander Slot ID : NA PID : UCSC-GPU-P4
Is Supported : Yes Vendor : nVidia
Model : Nvidia P4 Serial : 0322717109049
Running Version : 86.04.55.00.01|G414.0200.00.03
Activate Status : Ready Mode : Graphics

3. After the NVIDIA P40 GPU card is physically installed and the Cisco UCS C240 M5 Rack Server is discovered in Cisco UCS Manager, select the server and choose Inventory > GPUs. As shown in Figure 26, PCIe slot 2 and slot 5 are used with the two GRID P40 cards.

**Figure 26.** 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 : Graphics

Vendor ID : 4318
Sub Vendor ID : 4318
Device ID : 6968
Sub Device ID : 4569

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 : Graphics

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

## Install Citrix XenServer and XenCenter 7.5

XenServer installation media is available for download from the XenServer Downloads page.

Installers for both the XenServer host and XenCenter are located on the installation media. The installation media also includes the Readme First file, which provides descriptions of and links to helpful resources, including product documentation for XenServer and XenServer components.

You should install:

- XenServer 7.5 Base Installation ISO
- XenCenter 7.5 Windows Management Console

**Note:** Refer to the XenServer Installation Guide on the Citrix Product Documentation website for comprehensive details about XenServer and XenCenter installation and licensing.

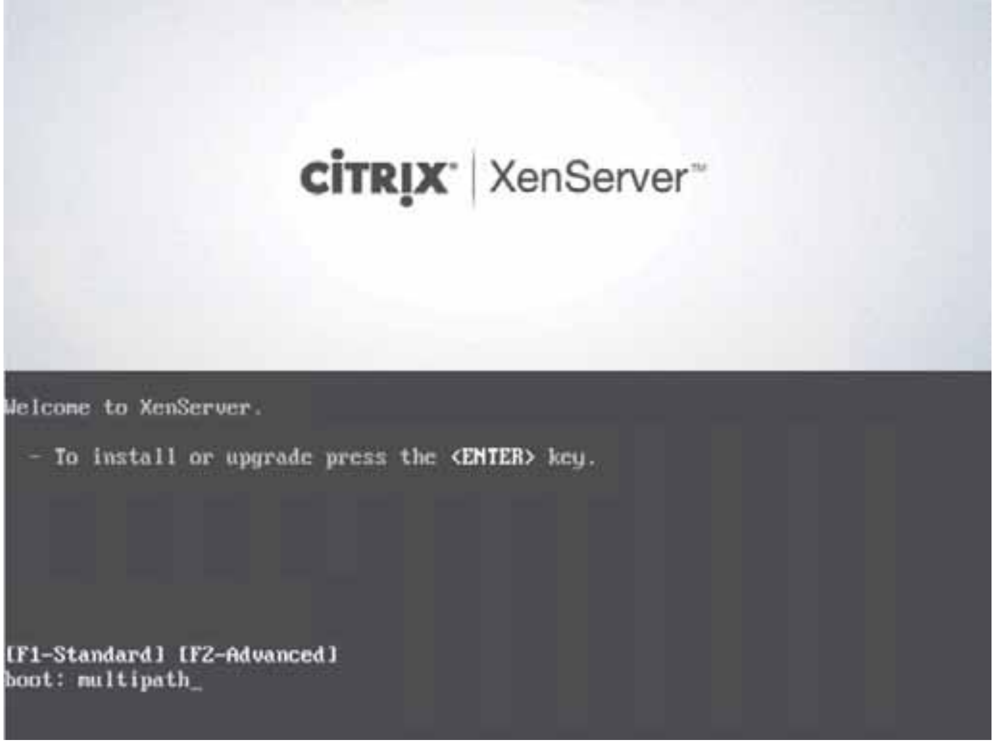
### Install Citrix XenServer

To install XenServer on a remote disk on a NetApp with multipathing enabled, follow these steps:

1. Boot the computer from the installation CD.

2. On the Welcome to XenServer screen, press F2.
3. At the boot prompt, enter multipath (Figure 27). This procedure configures the XenServer host, which boots from a logical unit number (LUN) with multipathing enabled.

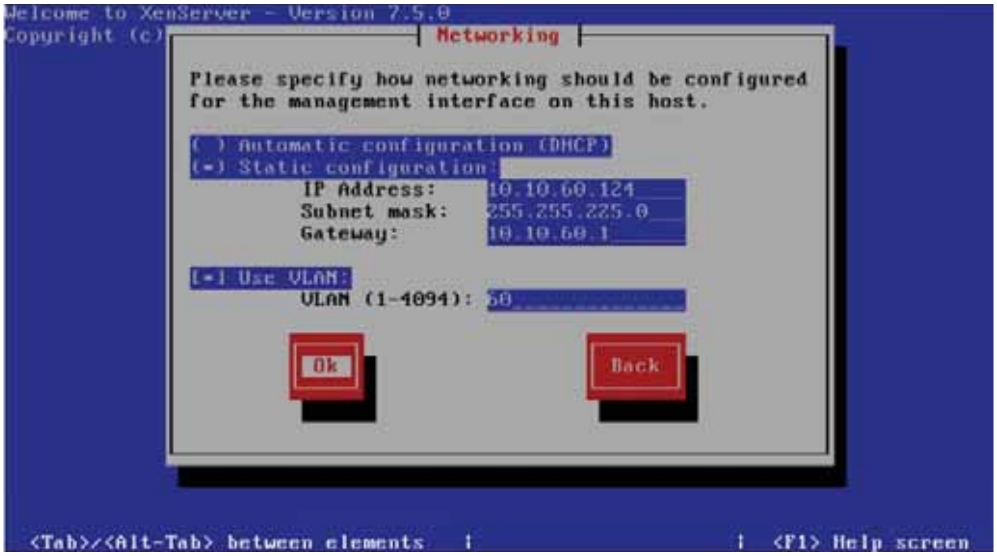
**Figure 27.** Citrix XenServer installation welcome screen



```
Welcome to XenServer.

  - To install or upgrade press the <ENTER> key.



[F1-Standard] [F2-Advanced]
boot: multipath_
```

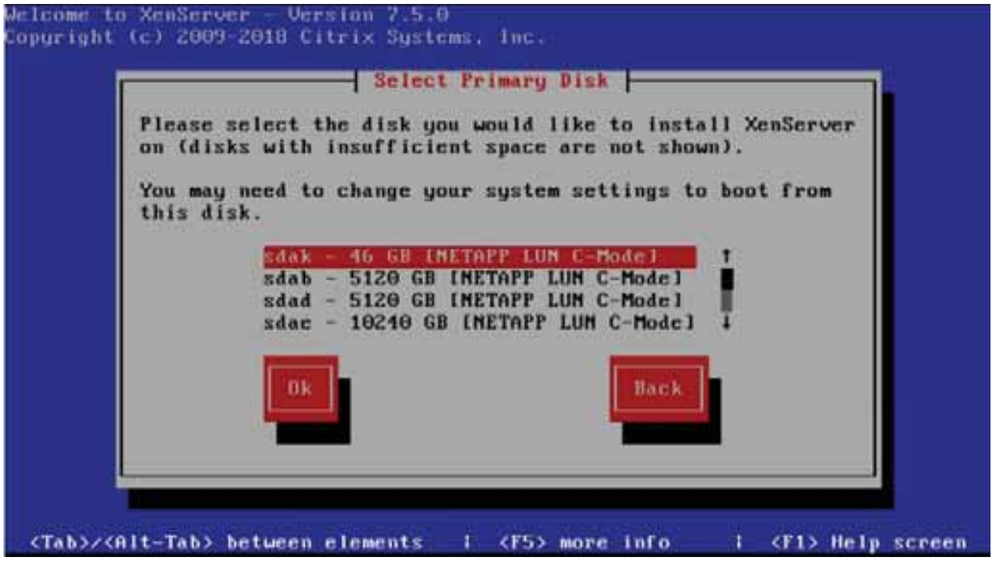
4. Follow the prompts to advance through the installation process, create a management interface on the FlexPod management VLAN. VLAN 60 (Figure 28), and choose a NetApp LUN as the primary disk for the installation (Figure 29). Click Ok.

**Figure 28.** Citrix XenServer installation: Networking



```
Welcome to XenServer - Version 7.5.0
Copyright (c)
                    Networking

  Please specify how networking should be configured
  for the management interface on this host.

  ( ) Automatic configuration (DHCP)
  (*) Static configuration:
           IP Address:      10.10.60.124
           Subnet mask:     255.255.225.0
           Gateway:         10.10.60.1

  [*] Use VLAN:
           VLAN (1-4094):   60

        Ok                          Back

 <Tab>/<Alt-Tab> between elements   |              |  <F1> Help screen
```

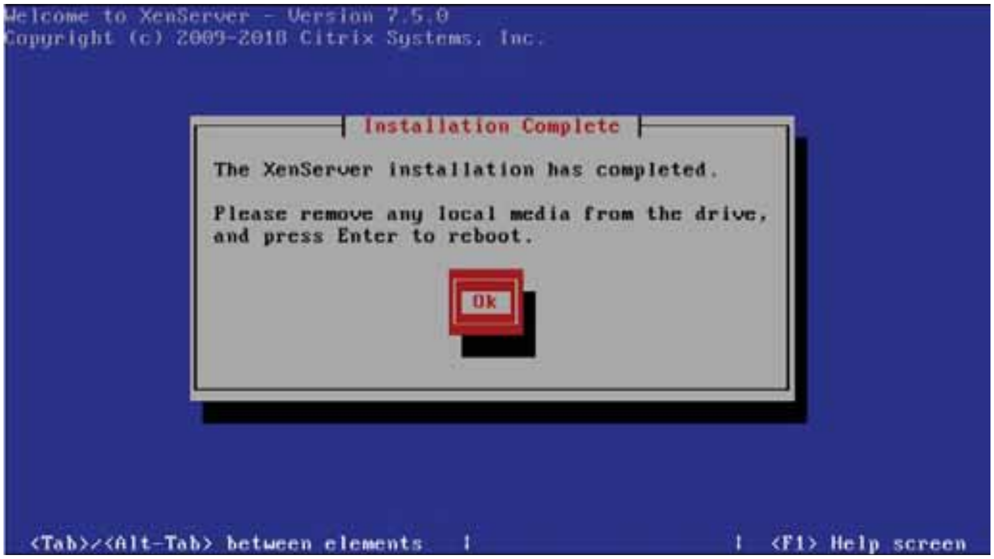
**Figure 29.** Citrix XenServer installation: Select Primary Disk



```
Welcome to XenServer - Version 7.5.0
Copyright (c) 2009-2018 Citrix Systems, Inc.

                    Select Primary Disk

  Please select the disk you would like to install XenServer
  on (disks with insufficient space are not shown).

  You may need to change your system settings to boot from
  this disk.

          sdak - 46 GB [NETAPP LUN C-Mode]
          sdab - 5120 GB [NETAPP LUN C-Mode]
          sdad - 5120 GB [NETAPP LUN C-Mode]
          sdac - 10240 GB [NETAPP LUN C-Mode]

        Ok                          Back

 <Tab>/<Alt-Tab> between elements   |  <F5> more info   |  <F1> Help screen
```

5. On the Installation Complete screen, click Ok to reboot the server (Figure 30).

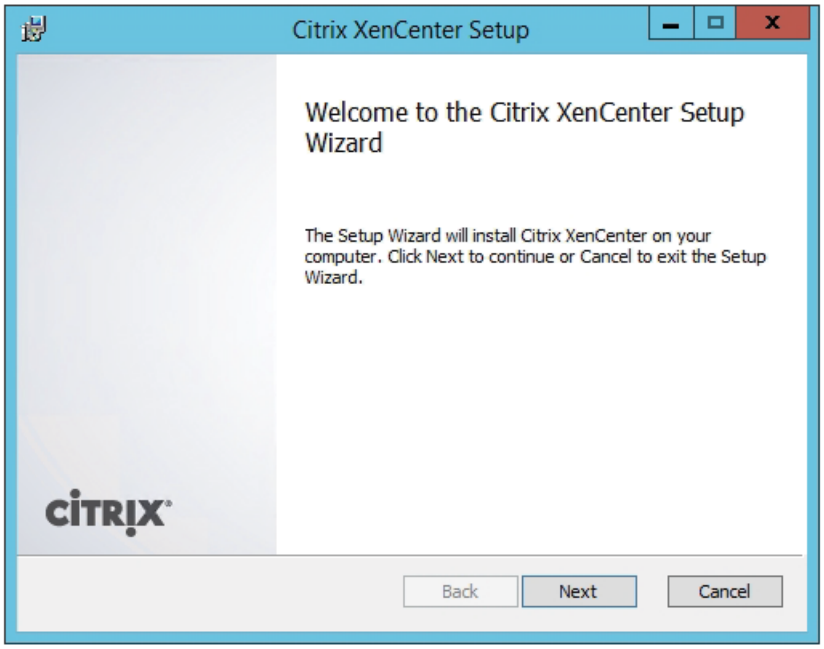
**Figure 30.** Citrix XenServer installation: Installation Complete



```
Welcome to XenServer - Version 7.5.0
Copyright (c) 2009-2018 Citrix Systems, Inc.

                 Installation Complete

    The XenServer installation has completed.

    Please remove any local media from the drive,
    and press Enter to reboot.

                          Ok

 <Tab>/<Alt-Tab> between elements   |        |  <F1> Help screen
```

6. The XenCenter installation media is bundled with the XenServer installation media. Launch the installer and follow the setup wizard (Figure 31).

**Figure 31.** Launch the setup wizard



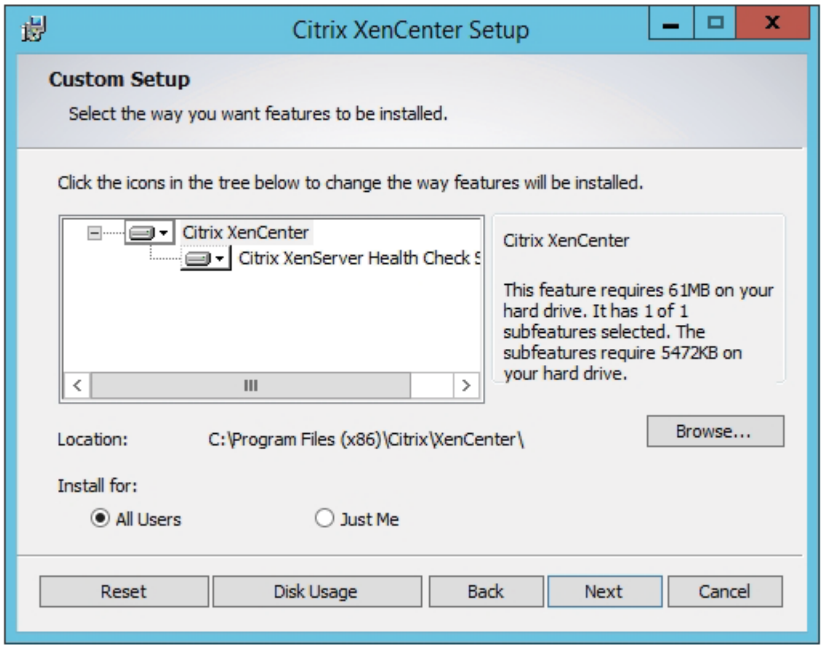
Citrix XenCenter Setup

Welcome to the Citrix XenCenter Setup Wizard

The Setup Wizard will install Citrix XenCenter on your computer. Click Next to continue or Cancel to exit the Setup Wizard.

7. You can modify the default destination folder. Otherwise, choose defaults and proceed with the installation (Figure 32).

**Figure 32.** Choosing the setup



**Custom Setup**

Select the way you want features to be installed.

Click the icons in the tree below to change the way features will be installed.

Citrix XenCenter

This feature requires 61MB on your hard drive. It has 1 of 1 subfeatures selected. The subfeatures require 5472KB on your hard drive.

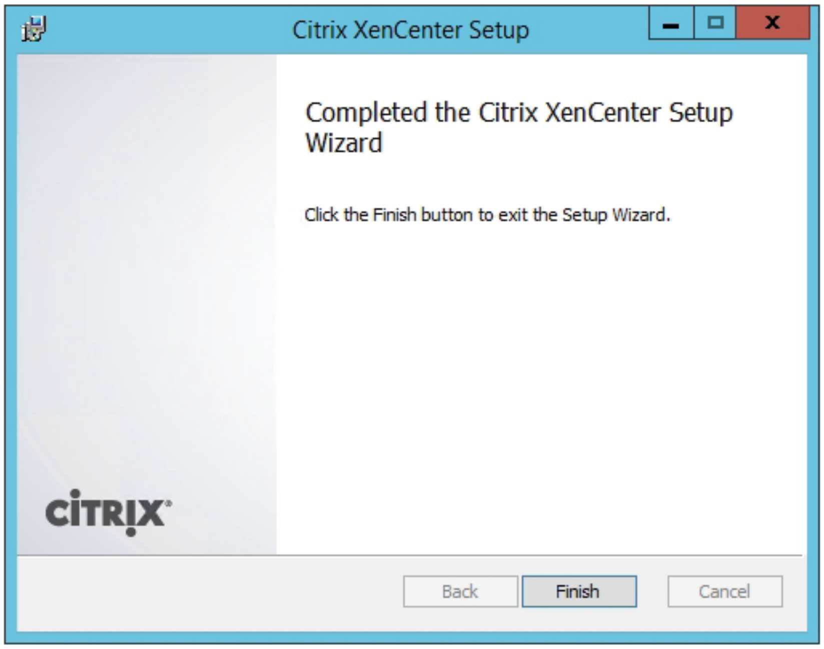
Location: C:\Program Files (x86)\Citrix\XenCenter\

Install for:

● All Users ○ Just Me

8. Click Finish to complete the installation (Figure 33).

**Figure 33.** Completing the installation



Completed the Citrix XenCenter Setup Wizard

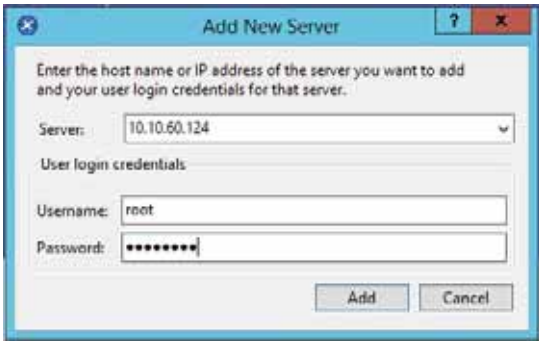
Click the Finish button to exit the Setup Wizard.

### Perform initial Citrix XenServer pool configuration

Launch XenCenter, add XenServer hosts, and configure the resource pool.

1. Click Server > Add. Provide connection details and select Add (Figure 34).

**Figure 34.** Adding a new server



Add New Server

Enter the host name or IP address of the server you want to add and your user login credentials for that server.

Server: 10.10.60.124

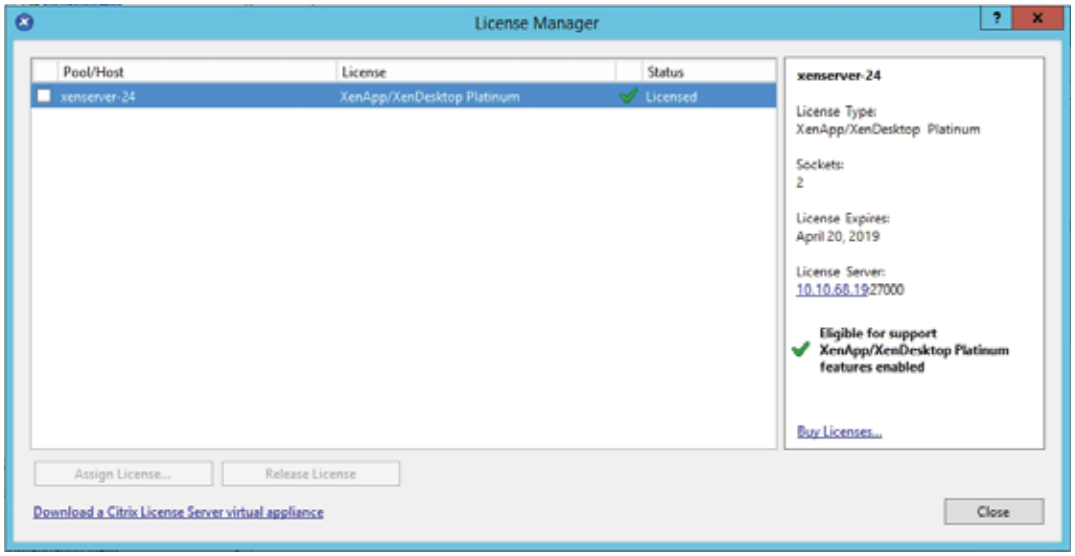
User login credentials

Username: root

Password: ••••••••

2. Click Tools > License Manager and provide licensing details (Figure 35). Click Close to close the License Manager.

**Figure 35.** Providing license details



License Manager

| Pool/Host | License | Status |
|---|---|---|
| xenserver-24 | XenApp/XenDesktop Platinum | ✓ Licensed |

xenserver-24

License Type:
XenApp/XenDesktop Platinum

Sockets:
2

License Expires:
April 20, 2019

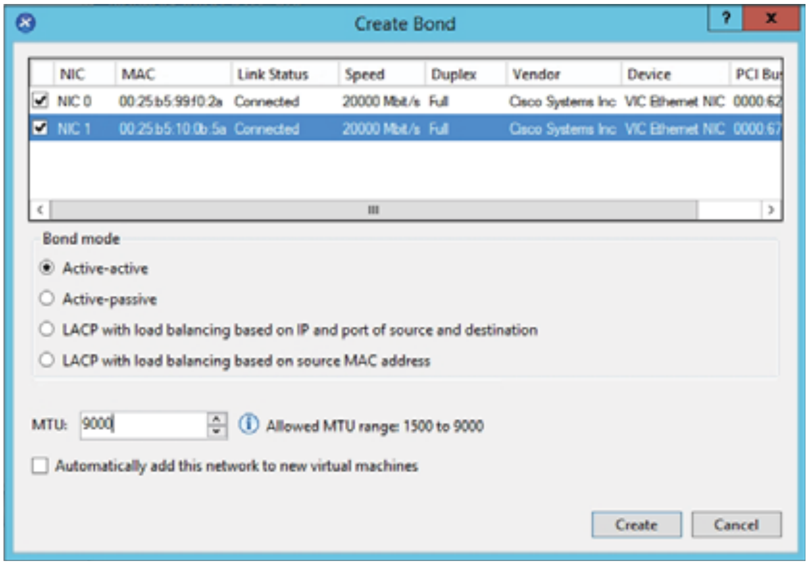
License Server:
10.10.68.1927000

✓ Eligible for support XenApp/XenDesktop Platinum features enabled

3. Repeat the same procedure for the additional XenServer hosts.
4. Create a NIC bond on XenServer (Figure 36):
   a. Select Active-active bond mode.
   b. Set the maximum transmission unit (MTU) to 9000.

**Note:** For simplicity and to prevent misconfiguration, Citrix recommends using XenCenter to create NIC bonds. Whenever possible, create NIC bonds as part of initial resource pool creation prior to joining additional hosts to the pool or creating virtual machines. Doing so allows the bond configuration to be automatically replicated to hosts as they join the pool and reduces the number of steps required.

**Figure 36.** Creating a NIC bond



Create Bond

| | NIC | MAC | Link Status | Speed | Duplex | Vendor | Device | PCI Bus |
|---|---|---|---|---|---|---|---|---|
| ☑ | NIC 0 | 00:25:b5:99:f0:2a | Connected | 20000 Mbit/s | Full | Cisco Systems Inc | VIC Ethernet NIC | 0000:62 |
| ☑ | NIC 1 | 00:25:b5:10:0b:5a | Connected | 20000 Mbit/s | Full | Cisco Systems Inc | VIC Ethernet NIC | 0000:67 |

Bond mode

● Active-active

○ Active-passive

○ LACP with load balancing based on IP and port of source and destination

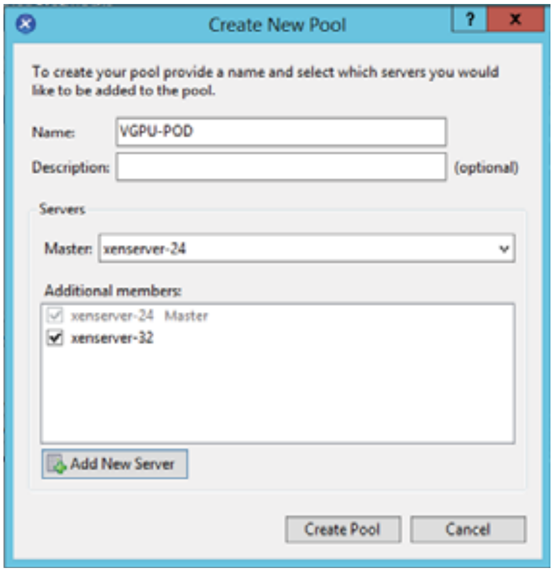
○ LACP with load balancing based on source MAC address

MTU: 9000 Allowed MTU range: 1500 to 9000

□ Automatically add this network to new virtual machines

5. To create a resource pool, choose Pool > New Pool. Provide the name of your pool and assign additional servers. Select Create Pool to finish creation process (Figure 37).

**Figure 37.** Creating a new resource pool



Create New Pool

To create your pool provide a name and select which servers you would like to be added to the pool.

Name: VGPU-POD

Description: (optional)

Servers

Master: xenserver-24

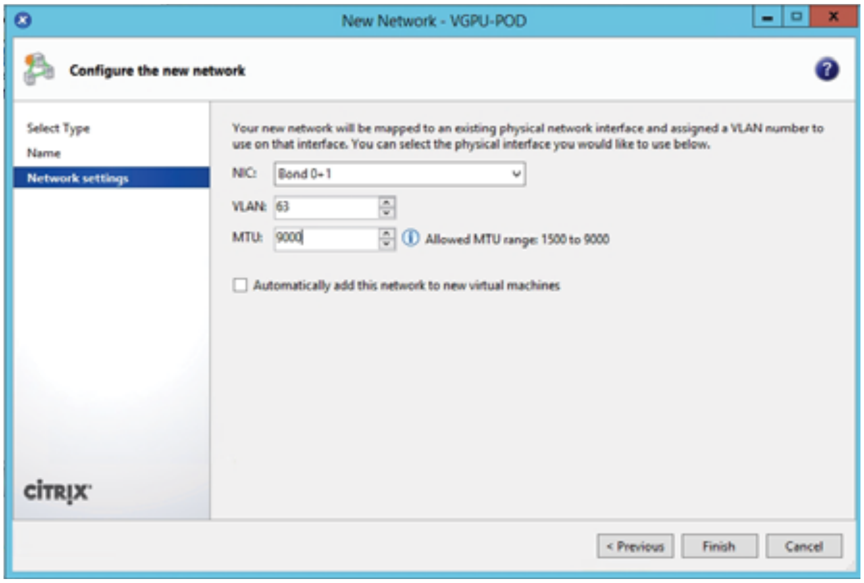
Additional members:

☑ xenserver-24 Master

☑ xenserver-32

6. Configure pool networking by creating two external networks to be used for storage and virtual machine connectivity:
   a. Create the storage network (Figure 38).

Figure 38. Creating a storage network



Your new network will be mapped to an existing physical network interface and assigned a VLAN number to use on that interface. You can select the physical interface you would like to use below.

NIC: Bond 0+1

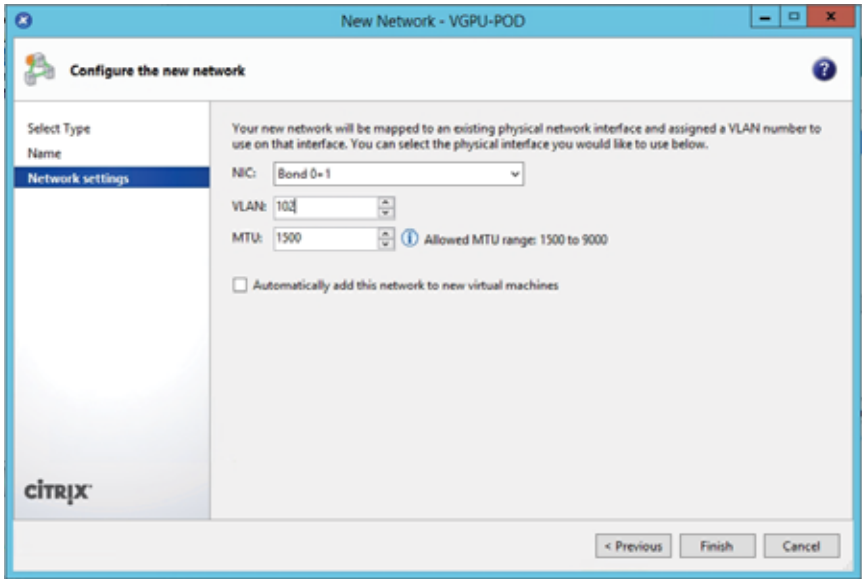
VLAN: 63

MTU: 9000 Allowed MTU range: 1500 to 9000

□ Automatically add this network to new virtual machines

b. Create the virtual machine network (Figure 39).

Figure 39. Creating a virtual machine network



Your new network will be mapped to an existing physical network interface and assigned a VLAN number to use on that interface. You can select the physical interface you would like to use below.

NIC: Bond 0+1

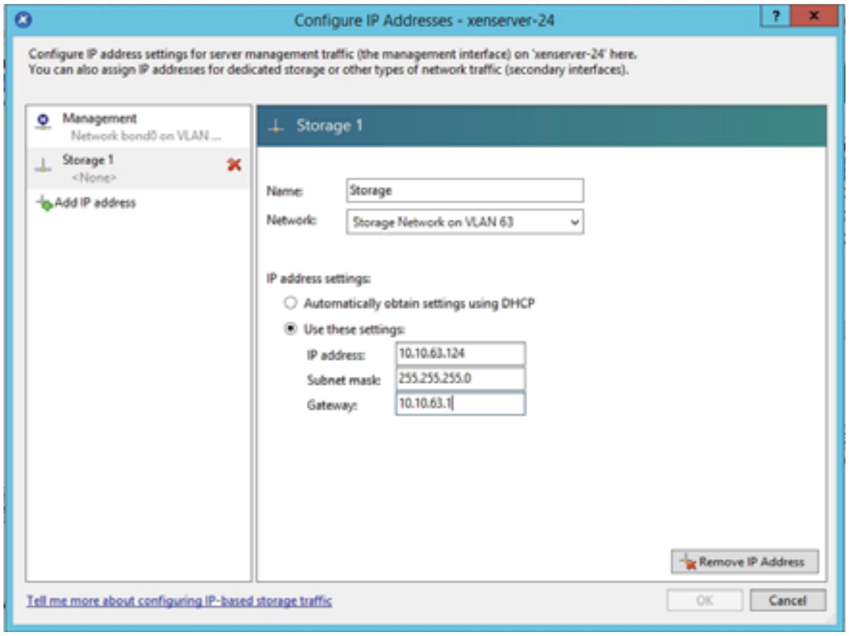
VLAN: 102

MTU: 1500 Allowed MTU range: 1500 to 9000

□ Automatically add this network to new virtual machines

7. Configure dedicated storage NICs on each host in the resource pool (Figure 40).

**Figure 40.** Configuring dedicated NICs



Configure IP Addresses - xenserver-24

Configure IP address settings for server management traffic (the management interface) on 'xenserver-24' here.
You can also assign IP addresses for dedicated storage or other types of network traffic (secondary interfaces).

Management
Network bond0 on VLAN ...

Storage 1
<None>

Add IP address

Storage 1

Name: Storage

Network: Storage Network on VLAN 63

IP address settings:

○ Automatically obtain settings using DHCP

● Use these settings:

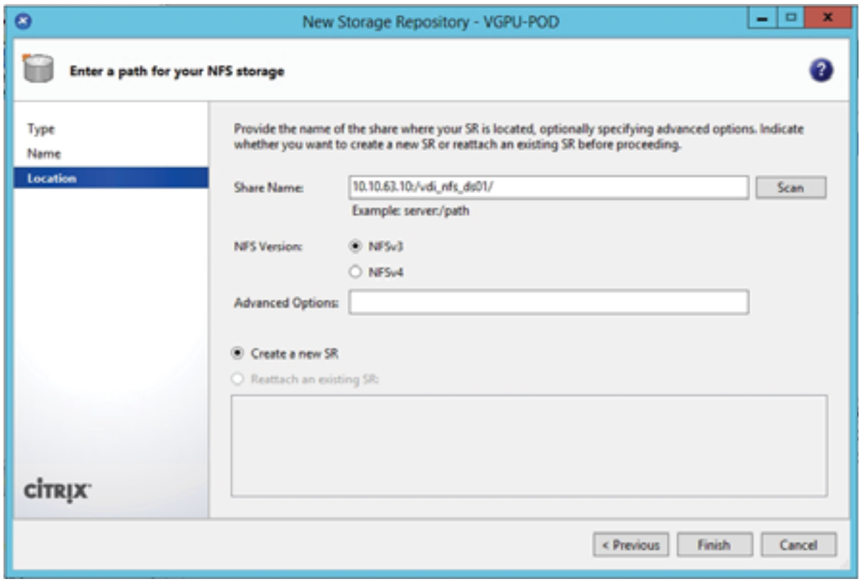
IP address: 10.10.63.124

Subnet mask: 255.255.255.0

Gateway: 10.10.63.1

8. Create an NFS storage repository for virtual machines by connecting to the appropriate NFS share on NetApp (Figure 41).

**Figure 41.** Connecting to an NFS share



New Storage Repository - VGPU-POD

Enter a path for your NFS storage

Provide the name of the share where your SR is located, optionally specifying advanced options. Indicate whether you want to create a new SR or reattach an existing SR before proceeding.

Share Name: 10.10.63.10:/vdi_nfs_ds01/

Example: server:/path

NFS Version: ● NFSv3 ○ NFSv4

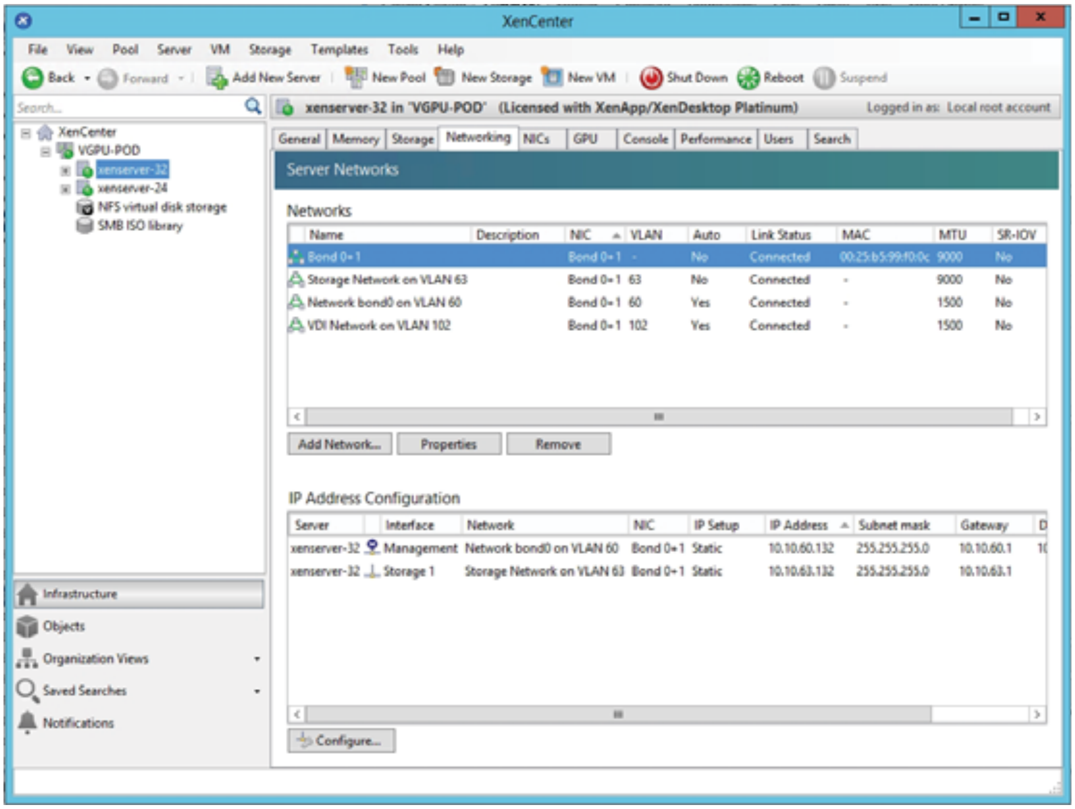
Advanced Options:

● Create a new SR

○ Reattach an existing SR:

The resulting resource pool configuration should look similar to Figure 42.

**Figure 42.** Resource pool configuration example



XenCenter
File View Pool Server VM Storage Templates Tools Help
Back Forward Add New Server New Pool New Storage New VM Shut Down Reboot Suspend
Search...
XenCenter
VGPU-POD
xenserver-32
xenserver-24
NFS virtual disk storage
SMB ISO library
Infrastructure
Objects
Organization Views
Saved Searches
Notifications
General Memory Storage Networking NICs GPU Console Performance Users Search
Add Network... Properties Remove
Configure...

xenserver-32 in 'VGPU-POD' (Licensed with XenApp/XenDesktop Platinum)    Logged in as: Local root account

Server Networks

Networks

| Name | Description | NIC | VLAN | Auto | Link Status | MAC | MTU | SR-IOV |
|---|---|---|---|---|---|---|---|---|
| Bond 0+1 | | Bond 0+1 | - | No | Connected | 00:25:b5:99:f0:0c | 9000 | No |
| Storage Network on VLAN 63 | | Bond 0+1 | 63 | No | Connected | - | 9000 | No |
| Network bond0 on VLAN 60 | | Bond 0+1 | 60 | Yes | Connected | - | 1500 | No |
| VDI Network on VLAN 102 | | Bond 0+1 | 102 | Yes | Connected | - | 1500 | No |

IP Address Configuration

| Server | Interface | Network | NIC | IP Setup | IP Address | Subnet mask | Gateway | D |
|---|---|---|---|---|---|---|---|---|
| xenserver-32 | Management | Network bond0 on VLAN 60 | Bond 0+1 | Static | 10.10.60.132 | 255.255.255.0 | 10.10.60.1 | 1 |
| xenserver-32 | Storage 1 | Storage Network on VLAN 63 | Bond 0+1 | Static | 10.10.63.132 | 255.255.255.0 | 10.10.63.1 | |

## Install the NVIDIA GRID vGPU Manager for Citrix XenServer

NVIDIA GRID allows virtual machines using the same NVIDIA graphics drivers as nonvirtualized operating systems to directly access the physical GPU on the hypervisor host. The NVIDIA GRID vGPU Manager manages multiple vGPU devices, which can be assigned directly to virtual machines.

You should install the latest NVIDIA GRID vGPU software. Instructions for obtaining the software are available from NVIDIA.

The GRID vGPU Manager runs in the XenServer Control Domain (dom0). It is provided as either a supplemental pack or a Red Hat Package Manager (RPM) file.

To install the GRID vGPU Manager in dom0 from an RPM package, follow the steps presented here.

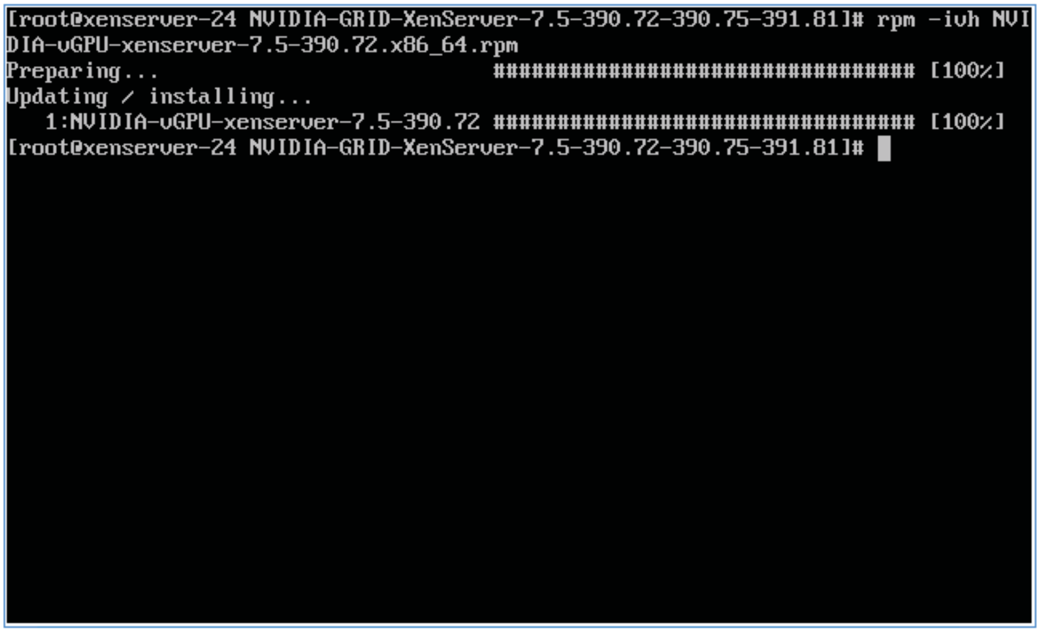
**Note:** When installing the GRID vGPU Manager using an RPM file, you must ensure that the RPM file is accessible from within dom0 prior to the installation.

When using NVIDIA vGPU with XenServer hosts with more than 768 GB of RAM, add the parameter `iommu=dom0-passthrough` to the Xen command line and restart the host.

1. Use the `rpm –ivh` command to install the package (Figure 43):

```
[root@xenserver ~]#  rpm -ivh NVIDIA-vGPU-xenserver-7.5-390.72.x86_64.rpm
```

**Figure 43.** Installing the RPM package

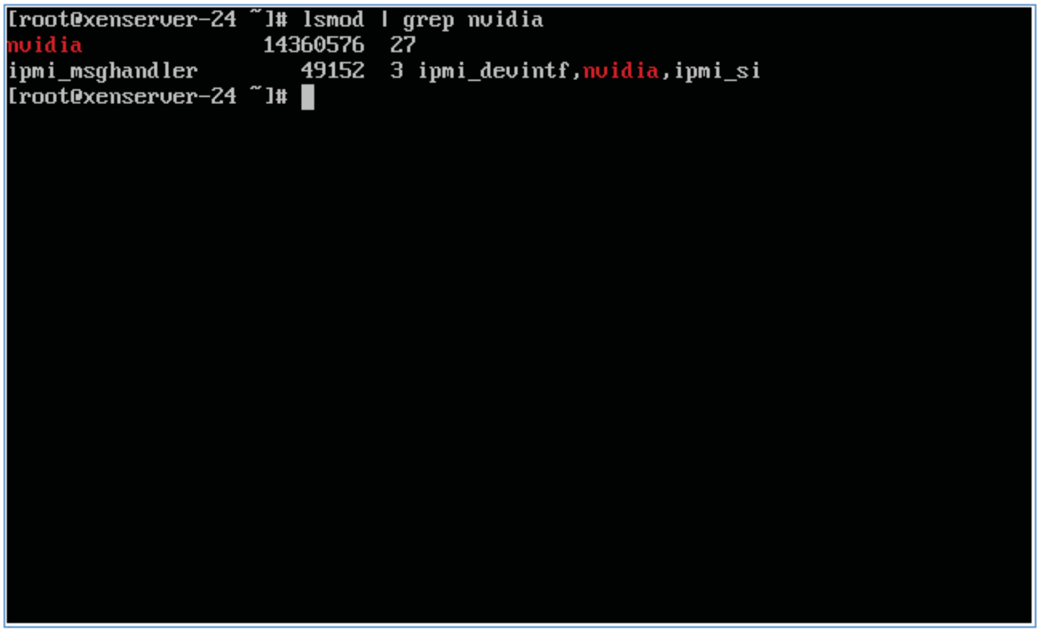


```
[root@xenserver-24 NVIDIA-GRID-XenServer-7.5-390.72-390.75-391.81]# rpm -ivh NVI
DIA-vGPU-xenserver-7.5-390.72.x86_64.rpm
Preparing...                          ################################# [100%]
Updating / installing...
   1:NVIDIA-vGPU-xenserver-7.5-390.72 ################################# [100%]
[root@xenserver-24 NVIDIA-GRID-XenServer-7.5-390.72-390.75-391.81]#
```

2. Reboot the XenServer host.
3. After rebooting the XenServer host, verify that the GRID package is installed and loaded correctly by checking for the NVIDIA kernel driver in the list of kernel loaded modules (Figure 44):

   ```
   [root@xenserver ~]#lsmod |grep nvidia
   ```

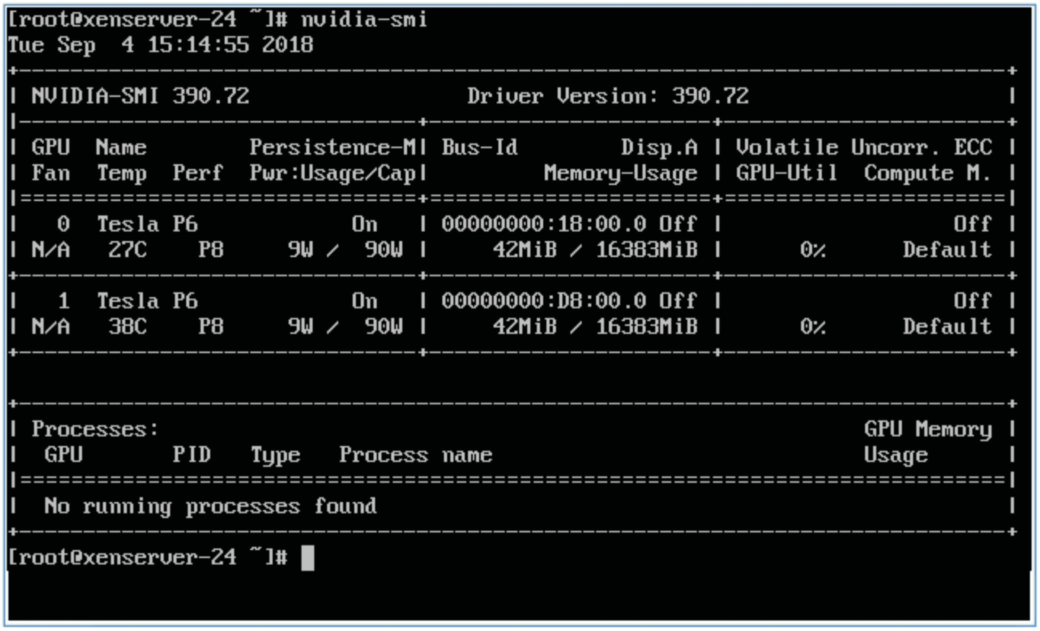
**Figure 44.** Verifying that the package is installed and loaded correctly



```
[root@xenserver-24 ~]# lsmod | grep nvidia
nvidia              14360576  27
ipmi_msghandler        49152  3 ipmi_devintf,nvidia,ipmi_si
[root@xenserver-24 ~]#
```

4. Verify that the NVIDIA kernel driver can successfully communicate with the GRID physical GPUs in your host by running the nvidia-smi command, which produces a list of the GPUs on your platform similar to Figure 45.

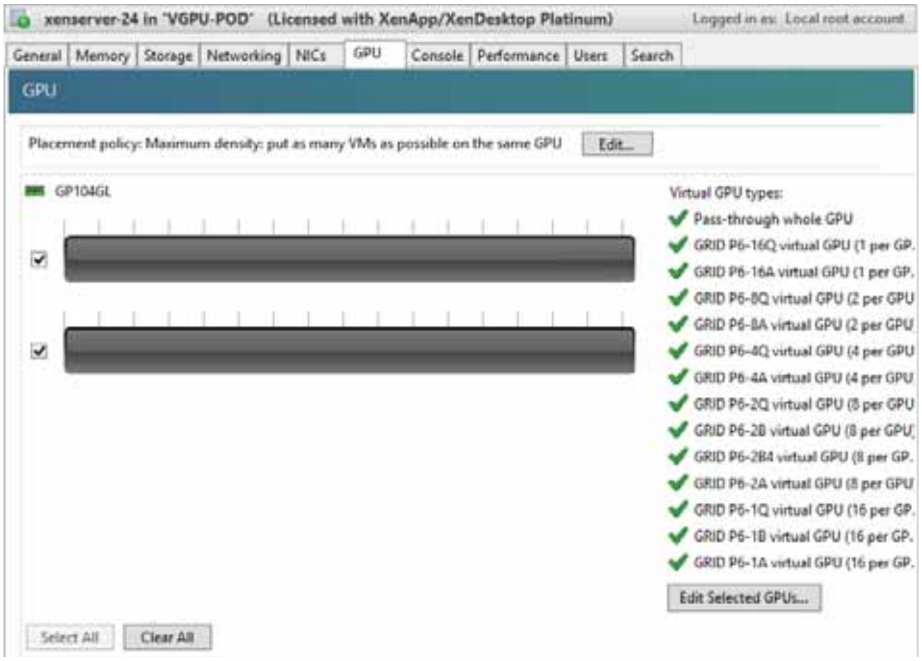
**Figure 45.** List of GPUs



```
[root@xenserver-24 ~]# nvidia-smi
Tue Sep  4 15:14:55 2018
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 390.72                 Driver Version: 390.72                    |
|-------------------------------+----------------------+----------------------+
| 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   27C    P8     9W /  90W |     42MiB / 16383MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   1  Tesla P6            On   | 00000000:D8:00.0 Off |                  Off |
| N/A   38C    P8     9W /  90W |     42MiB / 16383MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
[root@xenserver-24 ~]#
```

5. Repeat the same process for all hosts in the pool.
6. After the GRID vGPU Manager has been installed on all hosts in the resource pool, verify that the GPUs and vGPUs are visible on the XenCenter Host GPU tab as shown in Figure 46, Figure 47, and Figure 48.

**Figure 46.** Example of dual P6 availability on the Cisco UCS B200 M5 host for XenServer



xenserver-24 in 'VGPU-POD' (Licensed with XenApp/XenDesktop Platinum)

Logged in as: Local root account

Placement policy: Maximum density: put as many VMs as possible on the same GPU

GP104GL

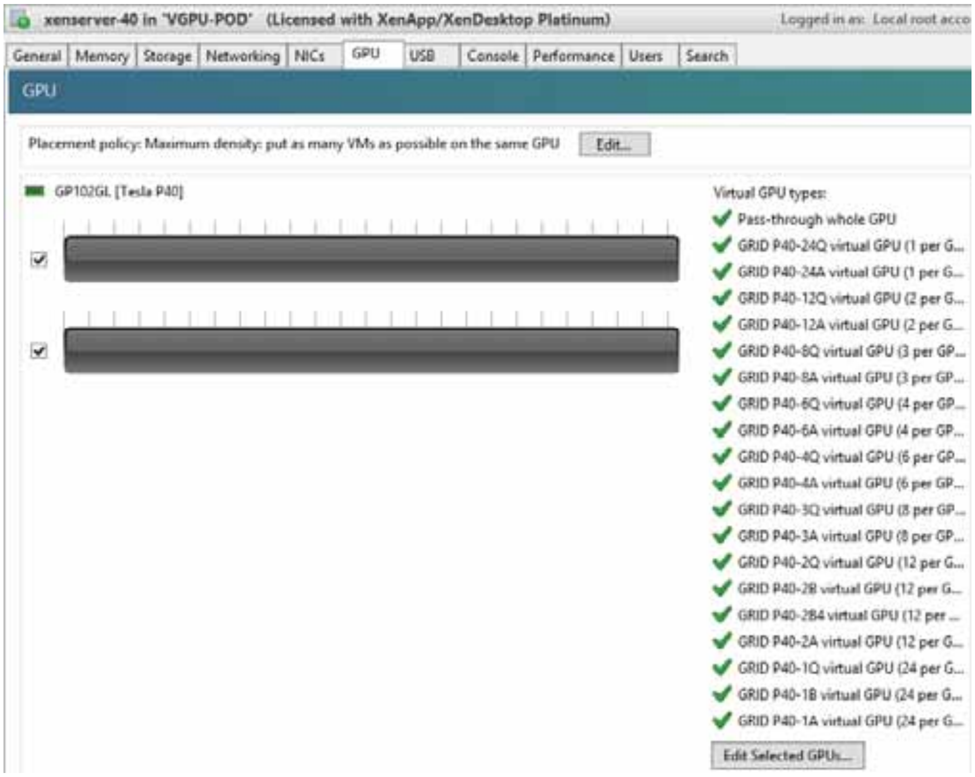
☑

☑

Virtual GPU types:

- ✓ Pass-through whole GPU
- ✓ GRID P6-16Q virtual GPU (1 per GP.
- ✓ GRID P6-16A virtual GPU (1 per GP.
- ✓ GRID P6-8Q virtual GPU (2 per GPU
- ✓ GRID P6-8A virtual GPU (2 per GPU
- ✓ GRID P6-4Q virtual GPU (4 per GPU
- ✓ GRID P6-4A virtual GPU (4 per GPU
- ✓ GRID P6-2Q virtual GPU (8 per GPU
- ✓ GRID P6-2B virtual GPU (8 per GPU
- ✓ GRID P6-2B4 virtual GPU (8 per GP.
- ✓ GRID P6-2A virtual GPU (8 per GPU
- ✓ GRID P6-1Q virtual GPU (16 per GP.
- ✓ GRID P6-1B virtual GPU (16 per GP.
- ✓ GRID P6-1A virtual GPU (16 per GP.

**Figure 47.** Example of dual P40 availability on the Cisco UCS C240 M5 host for XenServer



xenserver-40 in 'VGPU-POD' (Licensed with XenApp/XenDesktop Platinum)

Logged in as: Local root acco

Placement policy: Maximum density: put as many VMs as possible on the same GPU

GP102GL [Tesla P40]

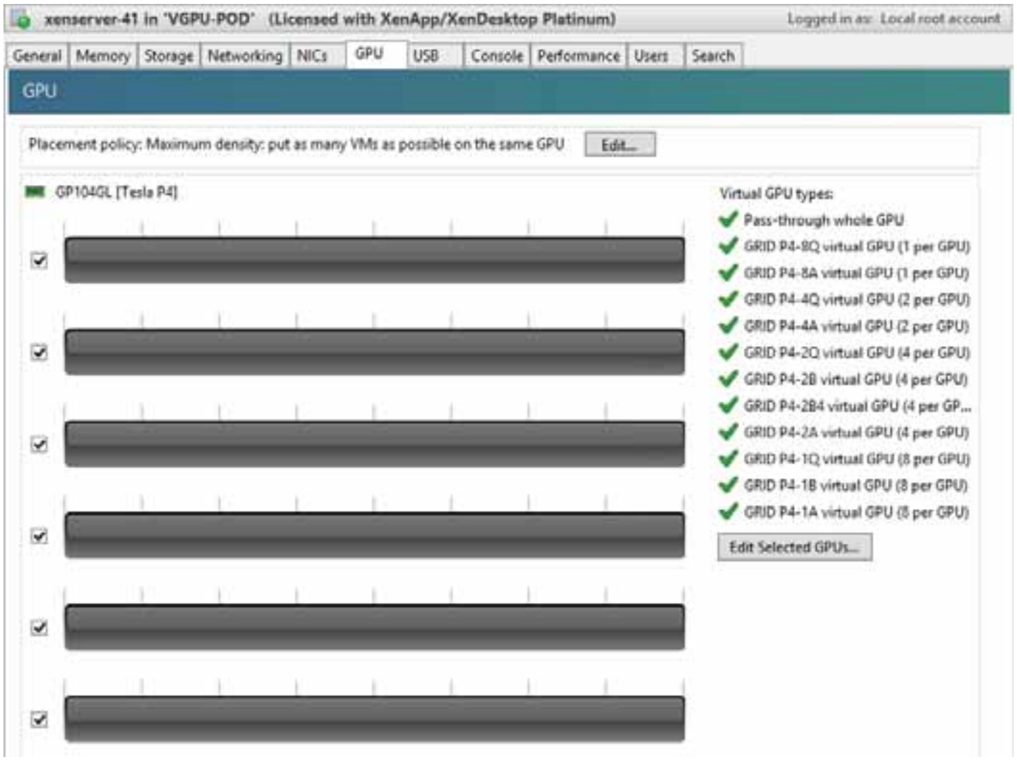
☑

☑

Virtual GPU types:

- ✓ Pass-through whole GPU
- ✓ GRID P40-24Q virtual GPU (1 per G...
- ✓ GRID P40-24A virtual GPU (1 per G...
- ✓ GRID P40-12Q virtual GPU (2 per G...
- ✓ GRID P40-12A virtual GPU (2 per G...
- ✓ GRID P40-8Q virtual GPU (3 per GP...
- ✓ GRID P40-8A virtual GPU (3 per GP...
- ✓ GRID P40-6Q virtual GPU (4 per GP...
- ✓ GRID P40-6A virtual GPU (4 per GP...
- ✓ GRID P40-4Q virtual GPU (6 per GP...
- ✓ GRID P40-4A virtual GPU (6 per GP...
- ✓ GRID P40-3Q virtual GPU (8 per GP...
- ✓ GRID P40-3A virtual GPU (8 per GP...
- ✓ GRID P40-2Q virtual GPU (12 per G...
- ✓ GRID P40-2B virtual GPU (12 per G...
- ✓ GRID P40-2B4 virtual GPU (12 per ...
- ✓ GRID P40-2A virtual GPU (12 per G...
- ✓ GRID P40-1Q virtual GPU (24 per G...
- ✓ GRID P40-1B virtual GPU (24 per G...
- ✓ GRID P40-1A virtual GPU (24 per G...

**Figure 48.** Example of sextuple P4 availability on the Cisco UCS C240 M5 host for XenServer



xenserver-41 in 'VGPU-POD' (Licensed with XenApp/XenDesktop Platinum) Logged in as: Local root account

GPU

Placement policy: Maximum density: put as many VMs as possible on the same GPU

GP104GL [Tesla P4]

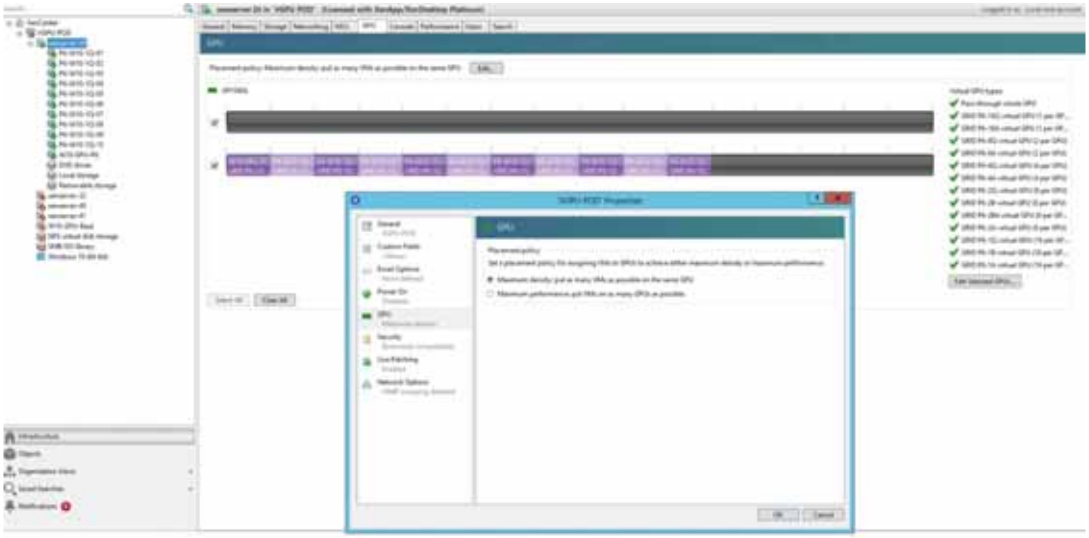
☑
☑
☑
☑
☑
☑

Virtual GPU types:

- ✓ Pass-through whole GPU
- ✓ GRID P4-8Q virtual GPU (1 per GPU)
- ✓ GRID P4-8A virtual GPU (1 per GPU)
- ✓ GRID P4-4Q virtual GPU (2 per GPU)
- ✓ GRID P4-4A virtual GPU (2 per GPU)
- ✓ GRID P4-2Q virtual GPU (4 per GPU)
- ✓ GRID P4-2B virtual GPU (4 per GPU)
- ✓ GRID P4-2B4 virtual GPU (4 per GP...
- ✓ GRID P4-2A virtual GPU (4 per GPU)
- ✓ GRID P4-1Q virtual GPU (8 per GPU)
- ✓ GRID P4-1B virtual GPU (8 per GPU)
- ✓ GRID P4-1A virtual GPU (8 per GPU)

7. Additionally, the GPU tab allows you to set a hostwide policy to assign virtual machines to available GPUs to achieve either maximum density or maximum performance. Select an option based on your requirements. Figure 49 and Figure 50 show examples of GPU utilization with these policies.

**Figure 49.** Virtual machine distribution policy for maximum density



'VGPU-POD' Properties

GPU

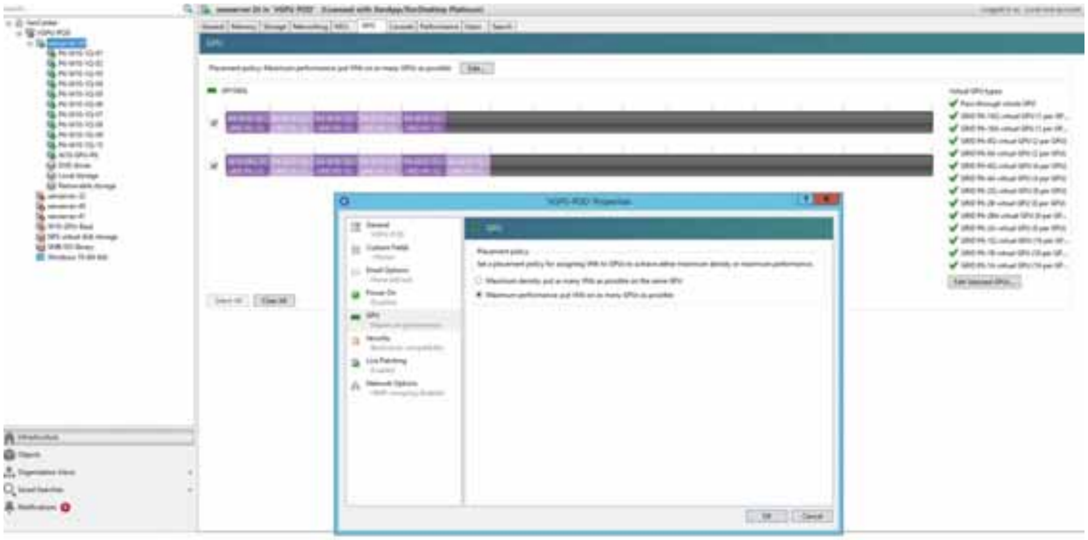
Placement policy

Set a placement policy for assigning VMs to GPUs to achieve either maximum density or maximum performance.

● Maximum density: put as many VMs as possible on the same GPU

○ Maximum performance: put VMs on as many GPUs as possible

**Figure 50.** Virtual machine distribution policy for maximum performance



Placement policy

Set a placement policy for assigning VMs to GPUs to achieve either maximum density or maximum performance.

○ Maximum density: put as many VMs as possible on the same GPU

● Maximum performance: put VMs on as many GPUs as possible

**Note:** GPU virtualization is available for XenServer Enterprise Edition customers and for those who have access to XenServer through their XenApp or XenDesktop entitlement. The GPU tab is displayed when the pool meets the license requirements and also has GPUs that support various vGPU types.

## Install and configure the NVIDIA GRID license server

This section summarizes the installation and configuration process for the GRID 6.2 license server.

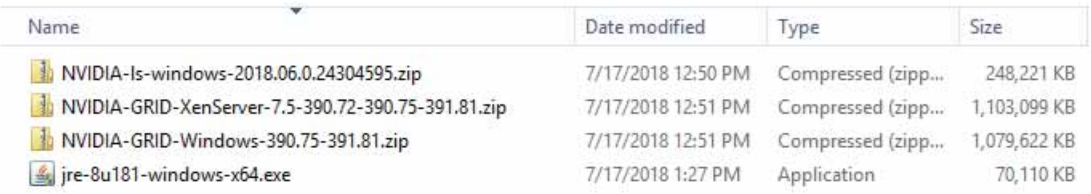
The NVIDIA GRID vGPU is a licensed feature on Tesla P6, P40, and P4. A software license is required to use the full vGPU features on a guest virtual machine. An NVIDIA license server with the appropriate licenses is required.

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

The following packages are required to set up the Citrix environment (Figure 51):

- NVIDIA GRID license server installer
- NVIDIA GRID Manager software, which is installed on XenServer dom0
- NVIDIA drivers and software that are installed in Microsoft Windows

**Figure 51.** Software required for NVIDIA GRID 6.2 setup on the VMware ESXi host



| Name | Date modified | Type | Size |
|---|---|---|---|
| NVIDIA-ls-windows-2018.06.0.24304595.zip | 7/17/2018 12:50 PM | Compressed (zipp... | 248,221 KB |
| NVIDIA-GRID-XenServer-7.5-390.72-390.75-391.81.zip | 7/17/2018 12:51 PM | Compressed (zipp... | 1,103,099 KB |
| NVIDIA-GRID-Windows-390.75-391.81.zip | 7/17/2018 12:51 PM | Compressed (zipp... | 1,079,622 KB |
| jre-8u181-windows-x64.exe | 7/17/2018 1:27 PM | Application | 70,110 KB |

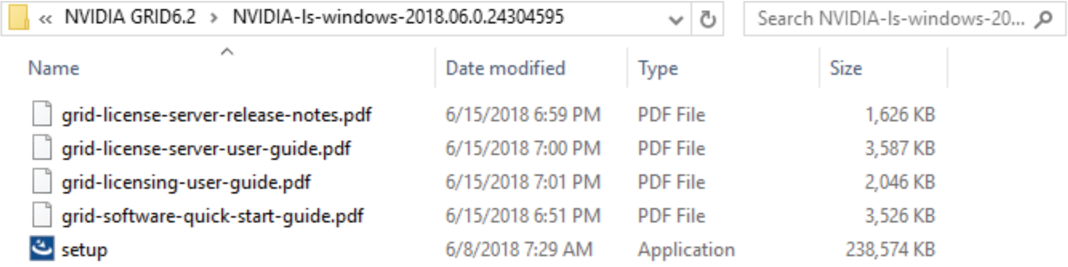
## Install the NVIDIA GRID 6.2 license server

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

The GRID 6.2 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-$version folder. Run setup.exe (Figure 52).

**Figure 52.** Run setup.exe

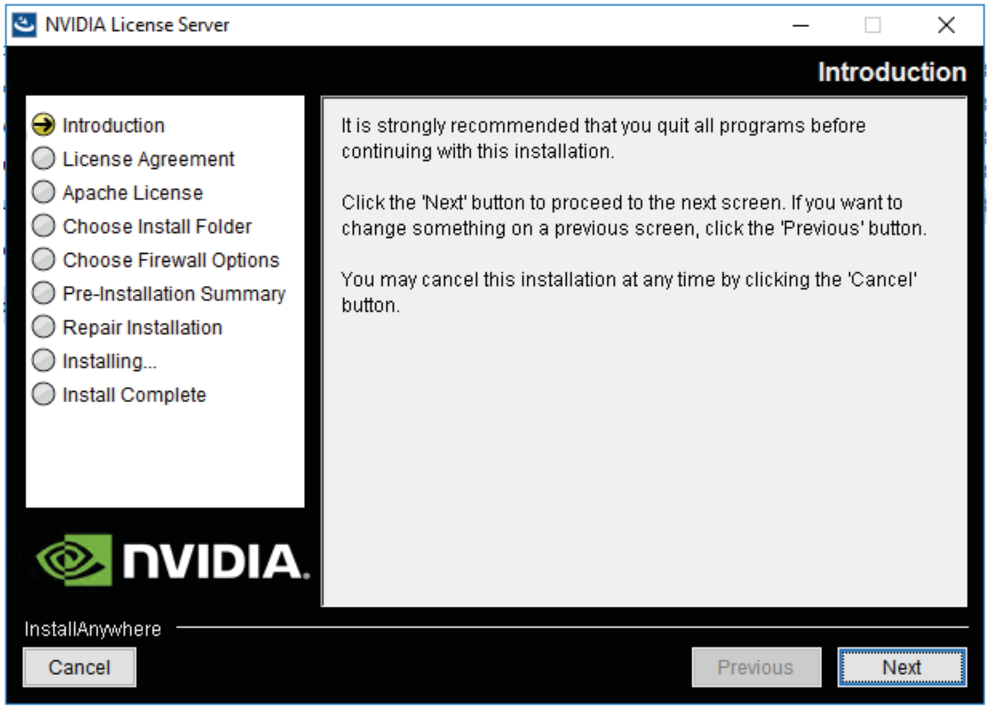


« NVIDIA GRID6.2 › NVIDIA-ls-windows-2018.06.0.24304595
Search NVIDIA-ls-windows-20...

| Name | Date modified | Type | Size |
|---|---|---|---|
| grid-license-server-release-notes.pdf | 6/15/2018 6:59 PM | PDF File | 1,626 KB |
| grid-license-server-user-guide.pdf | 6/15/2018 7:00 PM | PDF File | 3,587 KB |
| grid-licensing-user-guide.pdf | 6/15/2018 7:01 PM | PDF File | 2,046 KB |
| grid-software-quick-start-guide.pdf | 6/15/2018 6:51 PM | PDF File | 3,526 KB |
| setup | 6/8/2018 7:29 AM | Application | 238,574 KB |

2. Click Next (Figure 53).

**Figure 53.** NVIDIA License Server page



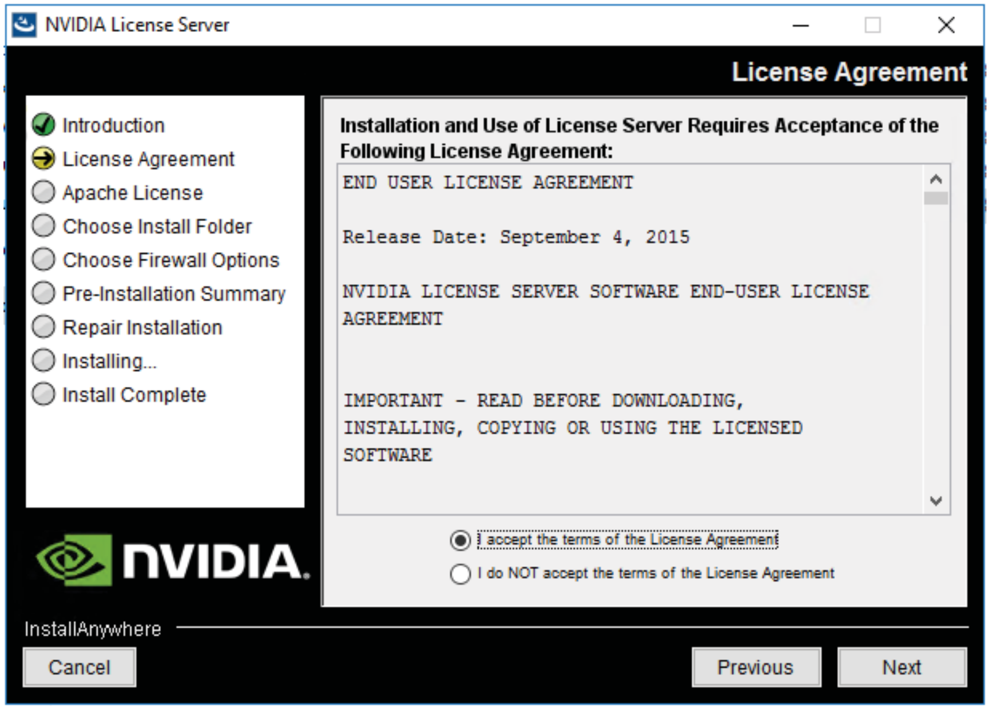
It is strongly recommended that you quit all programs before continuing with this installation.

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 54).

**Figure 54.** NVIDIA License Agreement page



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

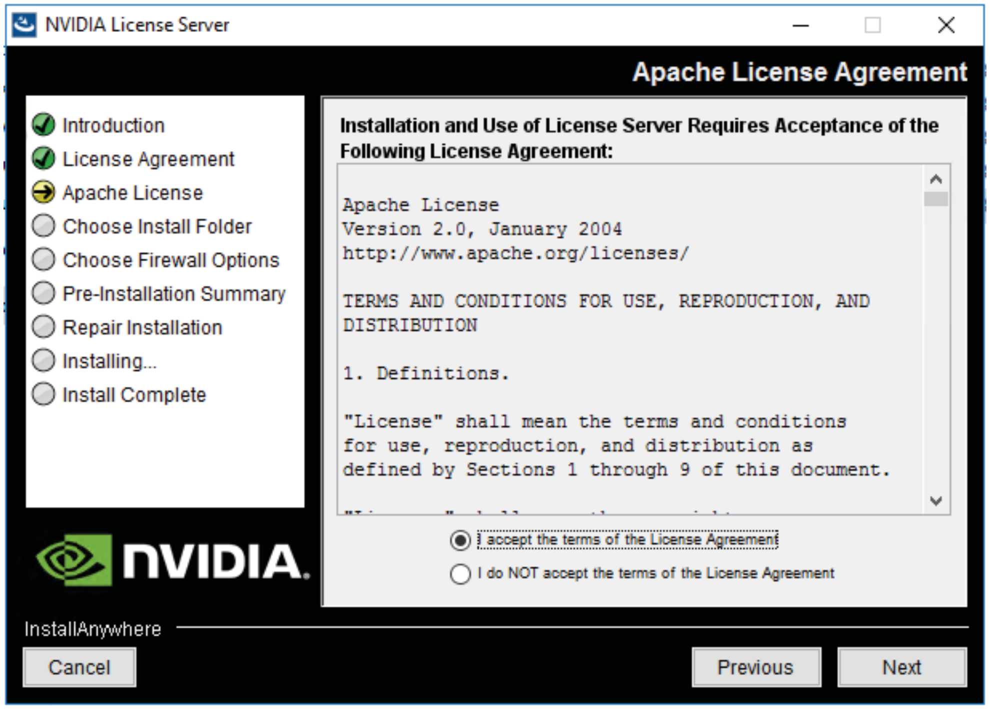
```
END USER LICENSE AGREEMENT

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 55).

**Figure 55.** Apache License Agreement page



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

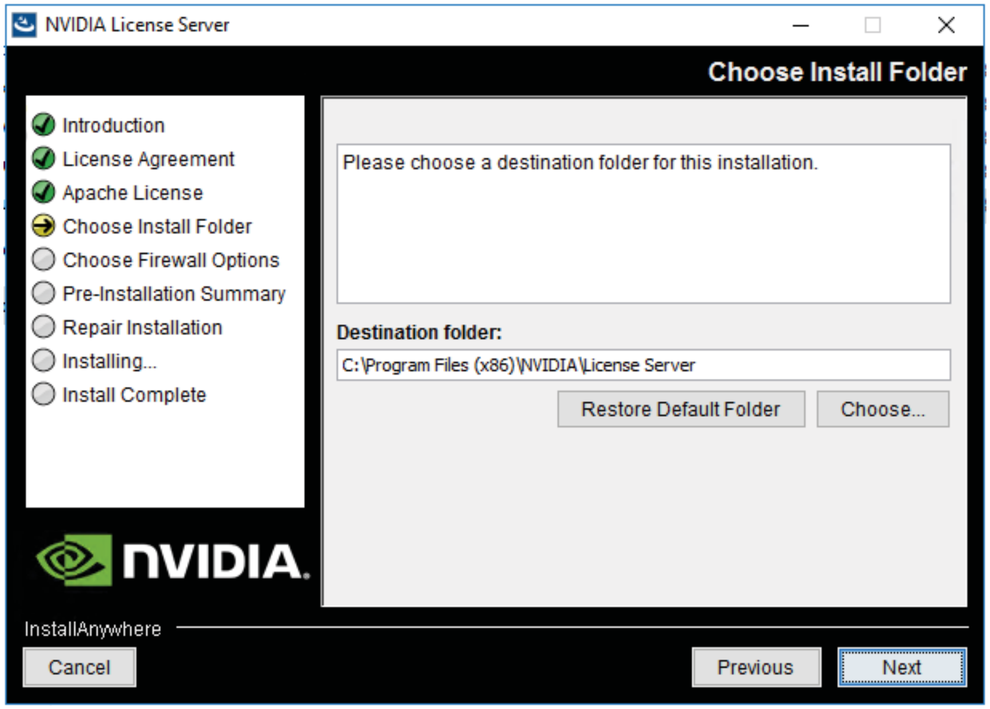
```
Apache License
Version 2.0, January 2004
http://www.apache.org/licenses/

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 56).

**Figure 56.** Choosing a destination folder



Please choose a destination folder for this installation.

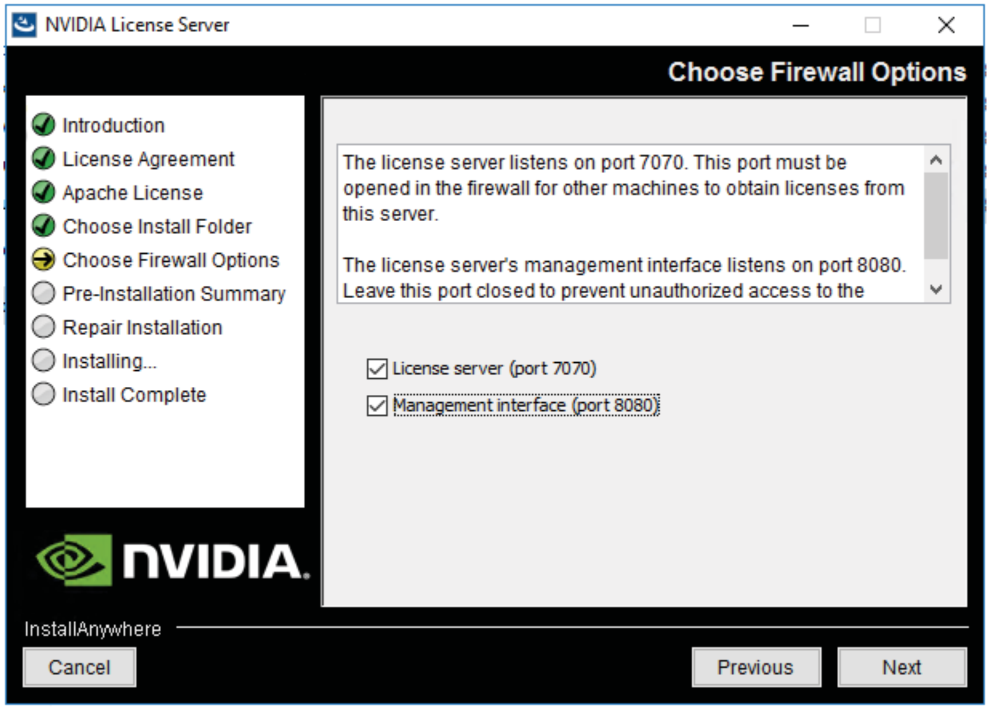
Destination folder: C:\Program Files (x86)\NVIDIA\License Server

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 need to open port 8080. Select the "Management interface (port 8080)" option.

8. Click Next (Figure 57).

**Figure 57.** 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.

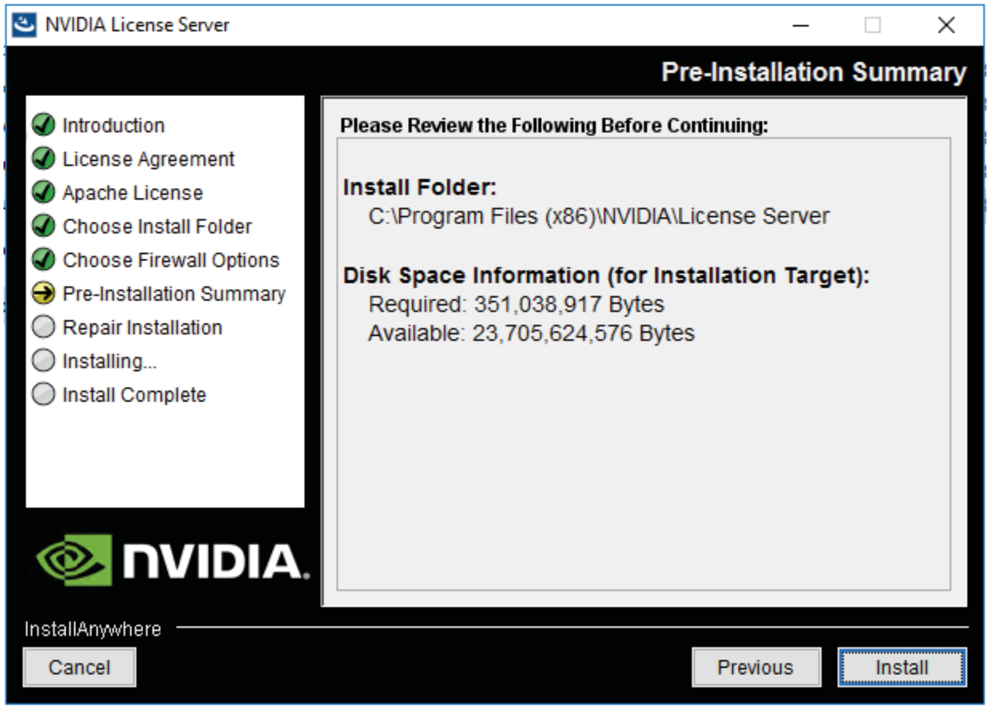
The license server's management interface listens on port 8080. Leave this port closed to prevent unauthorized access to the

☑ License server (port 7070)

☑ Management interface (port 8080)

9. The Pre-installation Summary and Repair Installation options automatically progress without user input (Figure 58).

**Figure 58.** Installing the license server



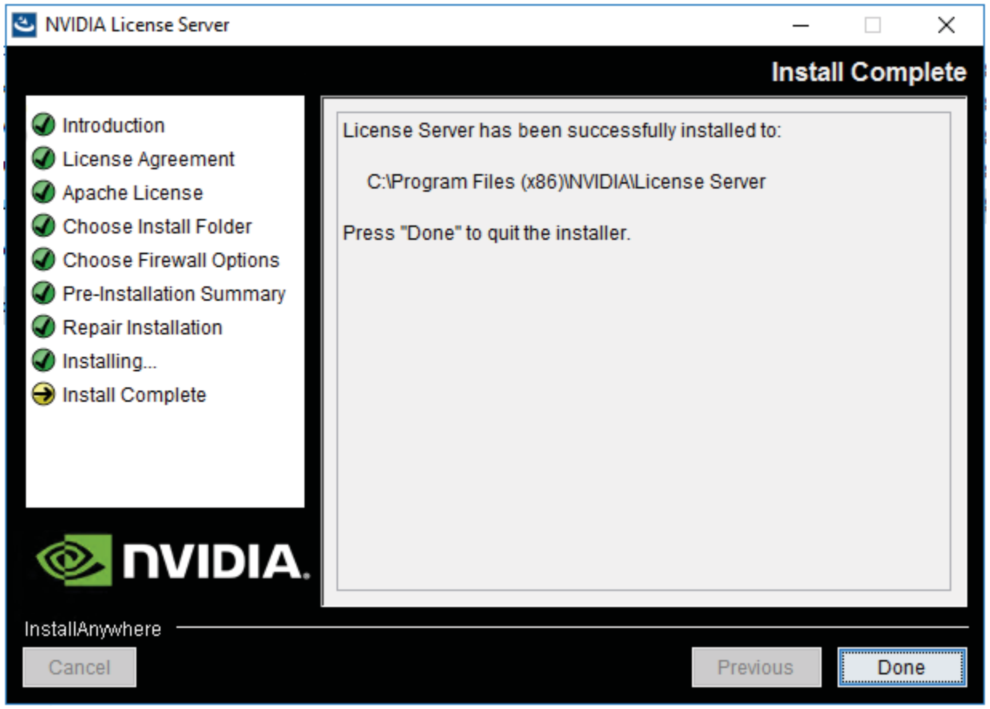
Please Review the Following Before Continuing:

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

Disk Space Information (for Installation Target):
Required: 351,038,917 Bytes
Available: 23,705,624,576 Bytes

10. When the installation process is complete, click Done (Figure 59).

**Figure 59.** Installation complete



License Server has been successfully installed to:

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

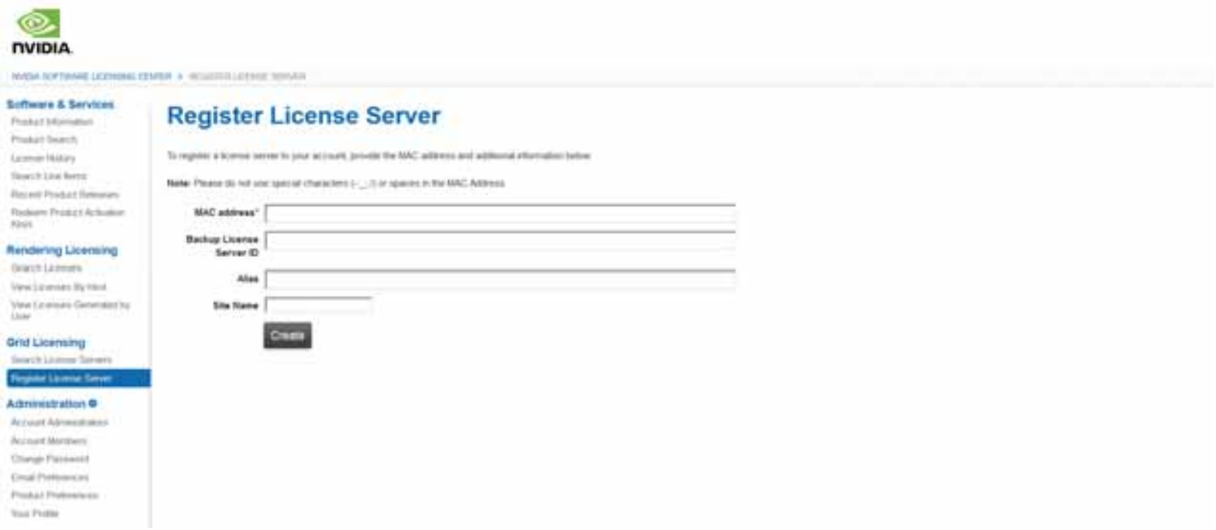
Press "Done" to quit the installer.

### Configure the NVIDIA GRID 6.2 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 Register License Server.
3. Specify the fields as shown in Figure 60. 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 60.** Registering the license server



Register License Server

To register a license server to your account, provide the MAC address and additional information below.

**Note:** Please do not use special characters (-_:) or spaces in the MAC Address.

**MAC address***

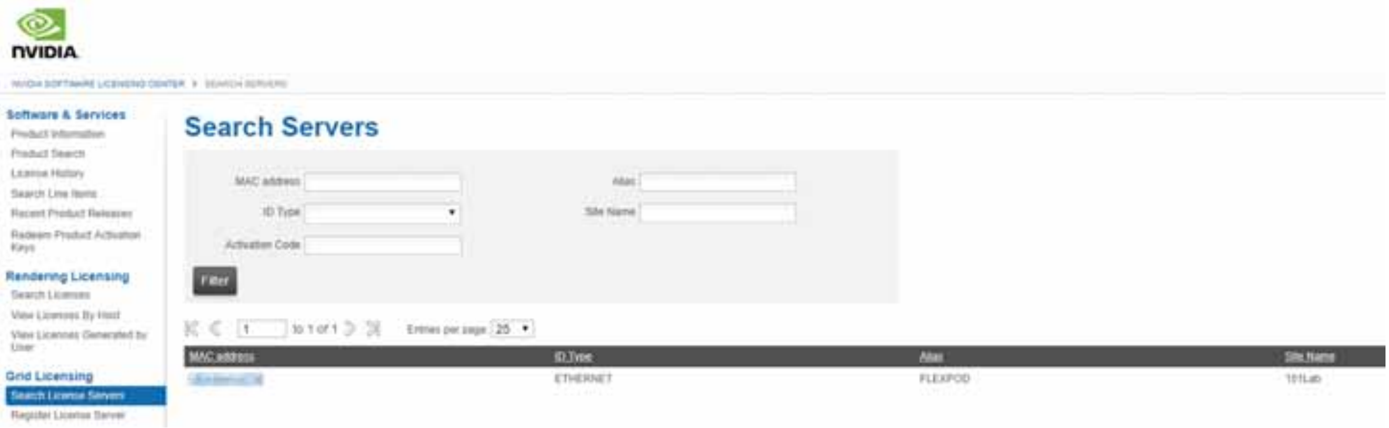
**Backup License Server ID**

**Alias**

**Site Name**

4. Click the Search License Servers node.
5. Click your license server ID (Figure 61).

**Figure 61.** Selecting the license server ID



Search Servers

MAC address

ID Type

Activation Code

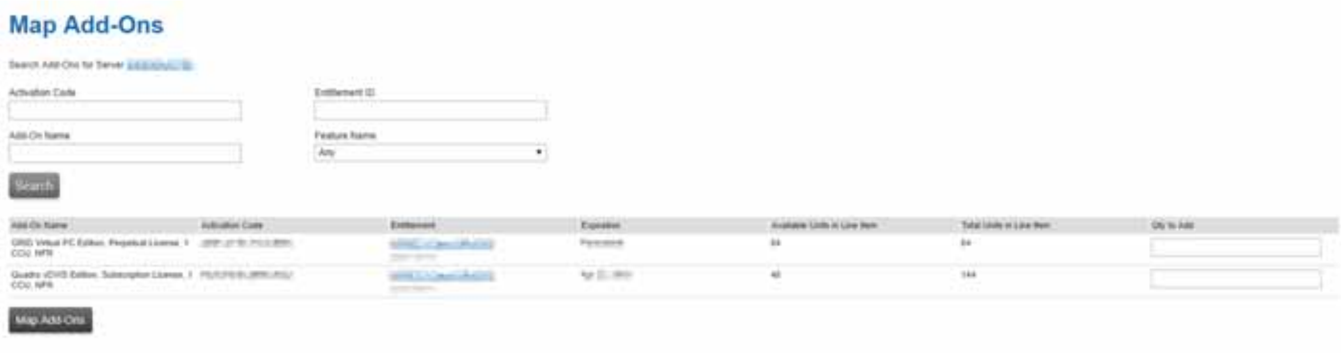
Alias

Site Name

| MAC address | ID Type | Alias | Site Name |
|---|---|---|---|
| [illegible] | ETHERNET | FLEXPOD | 101Lab |

6. Click Map Add-Ons, choose the number of license units from your total pool to allocate to this license server, and click Map Add-Ons (Figure 62 and Figure 63).

**Figure 62.** Choosing the number of license units



## Map Add-Ons

Search Add-Ons for Server [redacted]

Activation Code

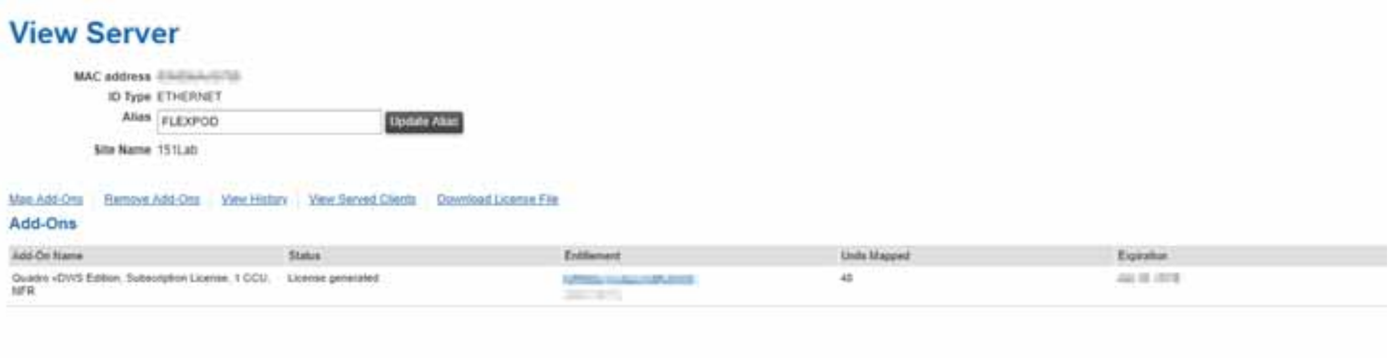
Entitlement ID

Add-On Name

Feature Name: Any

| Add-On Name | Activation Code | Entitlement | Expiration | Available Units in Line Item | Total Units in Line Item | Qty to Add |
|---|---|---|---|---|---|---|
| GRID Virtual PC Edition, Perpetual License, 1 CCU, NFR | [redacted] | [redacted] | Perpetual | 64 | 64 | |
| Quadro vDWS Edition, Subscription License, 1 CCU, NFR | [redacted] | [redacted] | Apr [redacted] | 48 | 144 | |

**Figure 63.** Mapped add-ons after successful mapping



## View Server

MAC address [redacted]
ID Type ETHERNET
Alias FLEXPOD
Site Name 151Lab

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

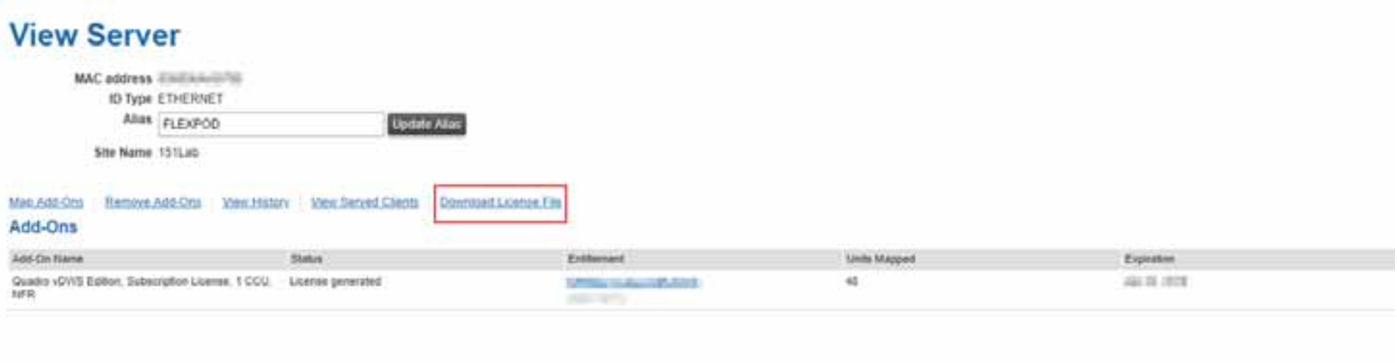
Add-Ons

| Add-On Name | Status | Entitlement | Units Mapped | Expiration |
|---|---|---|---|---|
| Quadro vDWS Edition, Subscription License, 1 CCU, NFR | License generated | [redacted] | 48 | [redacted] |

7. Click Download License File and save the .bin file to your license server (Figure 64).

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

**Figure 64.** Saving the .bin file



## View Server

MAC address [redacted]
ID Type ETHERNET
Alias FLEXPOD
Site Name 151Lab

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

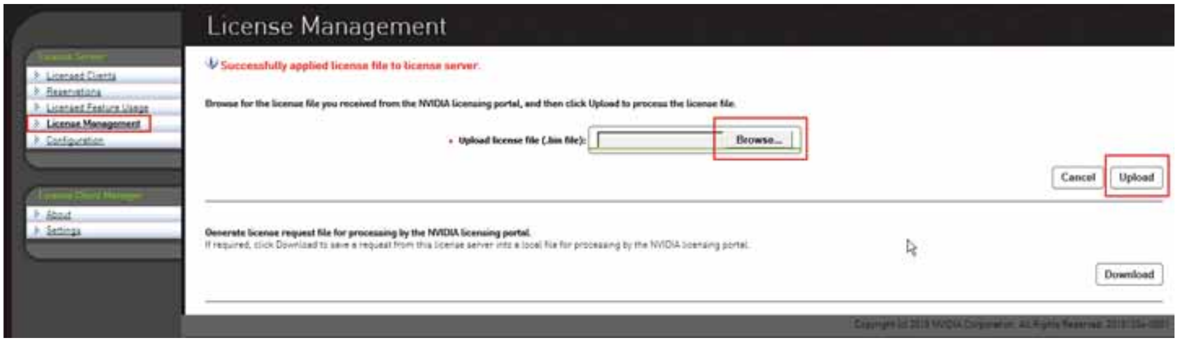
Add-Ons

| Add-On Name | Status | Entitlement | Units Mapped | Expiration |
|---|---|---|---|---|
| Quadro vDWS Edition, Subscription License, 1 CCU, NFR | License generated | [redacted] | 48 | [redacted] |

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 65). The features are now available (Figure 66).

**Figure 65.** License file successfully applied



License Management

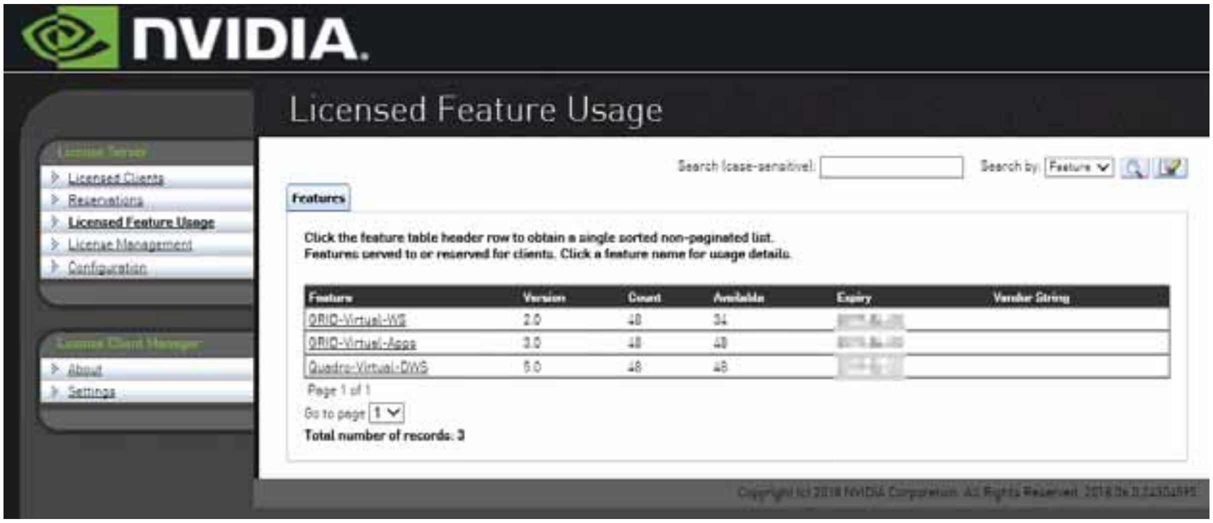
Successfully applied license file to license server.

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 66.** NVIDIA Licensed Feature Usage page



Licensed Feature Usage

Search (case-sensitive):

Search by: Feature

Click the feature table header row to obtain a single sorted non-paginated list.
Features served to or reserved for clients. Click a feature name for usage details.

| Feature | Version | Count | Available | Expiry | Vendor String |
|---|---|---|---|---|---|
| GRID-Virtual-WS | 2.0 | 48 | 34 | [redacted] | |
| GRID-Virtual-Apps | 3.0 | 48 | 48 | [redacted] | |
| Quadro-Virtual-DWS | 5.0 | 48 | 48 | [redacted] | |

Page 1 of 1

Go to page 1

Total number of records: 3

## NVIDIA Tesla P6, P40, and P4 profile specifications

The Tesla P6, P4, and P40 cards have a single physical GPU. 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 4 lists the vGPU types supported by GRID GPUs.

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

**Table 4.** User profile specifications for NVIDIA Tesla cards

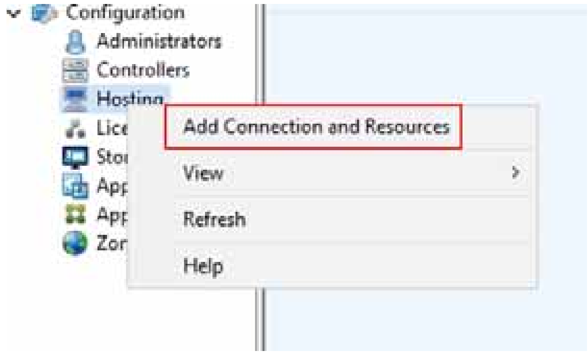
| **End-user GRID options** | | | |
|---|---|---|---|
| **End-user profile** | **GRID virtual application profiles** | **GRID virtual PC profiles** | **NVIDIA Quadro Virtual Data Center Workstation (Quadro vDWS) profiles** |
| 1 GB | P6-1A<br>P4-1A<br>P40-1A | P6-1B<br>P4-1B<br>P40-1B | P6-1Q<br>P4-1Q<br>P40-1Q |
| 2 GB | P6-2A<br>P4-2A<br>P40-2A | P6-2B<br>P4-2B<br>P40-2B | P6-2Q<br>P4-2Q<br>P40-2Q |
| 3 GB | P40-3A | – | P40-3Q |
| 4 GB | P6-4A<br>P4-4A<br>P40-4A | – | P6-4Q<br>P4-4Q<br>P40-4Q |
| 6GB | P40-6A | – | P40-6Q |
| 8 GB | P6-8A<br>P4-8A<br>P40-8A | – | P6-8Q<br>P4-8Q<br>P40-8Q |
| 12 GB | P40-12A | – | P40-12Q |
| 16 GB | P6-16A | – | P6-16Q |
| 24 GB | P40-24A | – | P40-24Q |

## Prepare Citrix delivery controllers for MCS provisioning with vGPU support

Follow these steps to prepare Citrix delivery controllers for MCS provisioning with vGPU support.

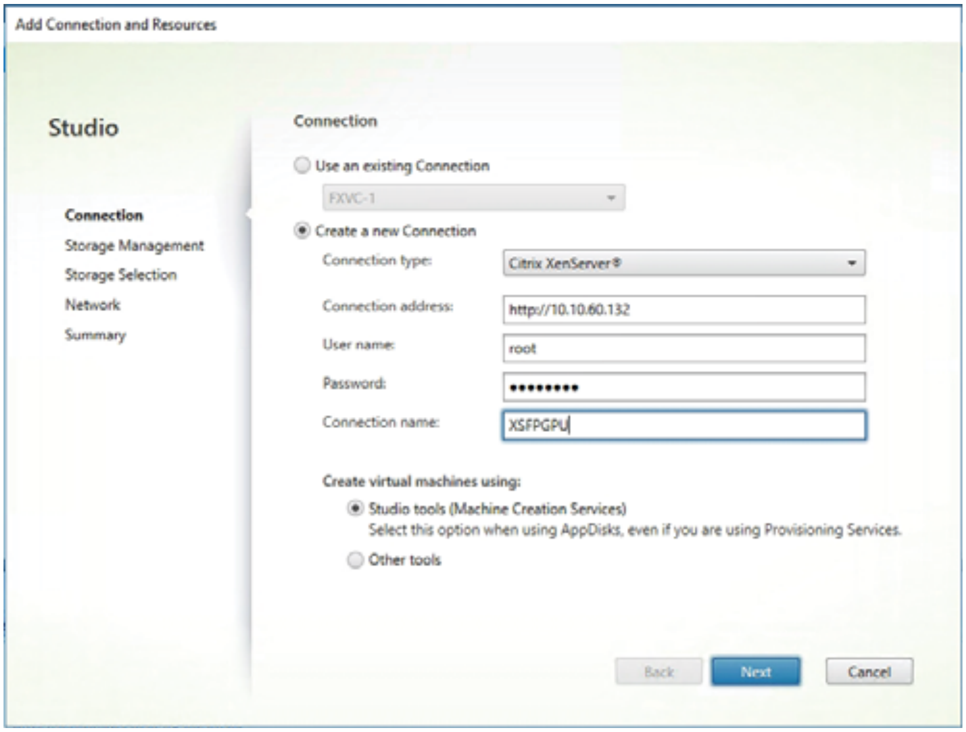
1. In the Studio navigation pane, select Configuration > Hosting > Add Connection and Resources (Figure 67).

**Figure 67.** Adding a connection and resources



2. On the Connection page, select Create a new Connection and provide the necessary credentials (Figure 68).

**Figure 68.** Creating a new connection



Add Connection and Resources

Studio

Connection

○ Use an existing Connection

FXVC-1

● Create a new Connection

Connection type: Citrix XenServer®

Connection address: http://10.10.60.132

User name: root

Password: ••••••••

Connection name: XSFPGPU

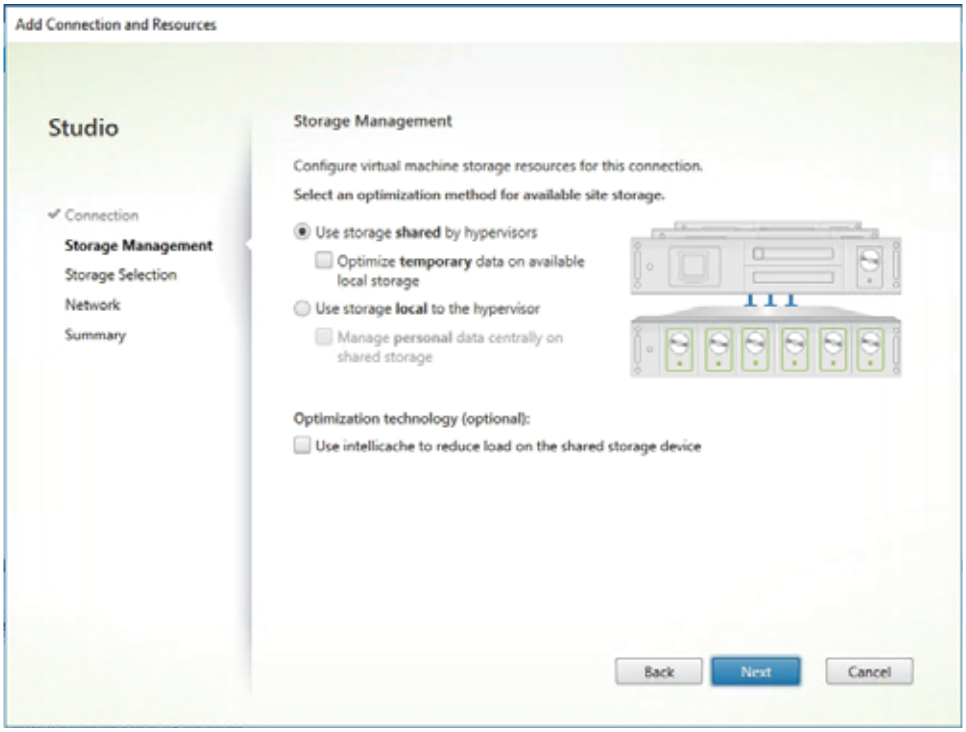
Create virtual machines using:

● Studio tools (Machine Creation Services)
Select this option when using AppDisks, even if you are using Provisioning Services.

○ Other tools

3. On the Storage Management page, select a storage management method. Select “Use storage shared by hypervisors” (Figure 69).

**Figure 69.** Selecting a storage management method



Add Connection and Resources

Studio

- ✓ Connection
- **Storage Management**
- Storage Selection
- Network
- Summary

Storage Management

Configure virtual machine storage resources for this connection.

Select an optimization method for available site storage.

● Use storage **shared** by hypervisors

□ Optimize **temporary** data on available local storage

○ Use storage **local** to the hypervisor

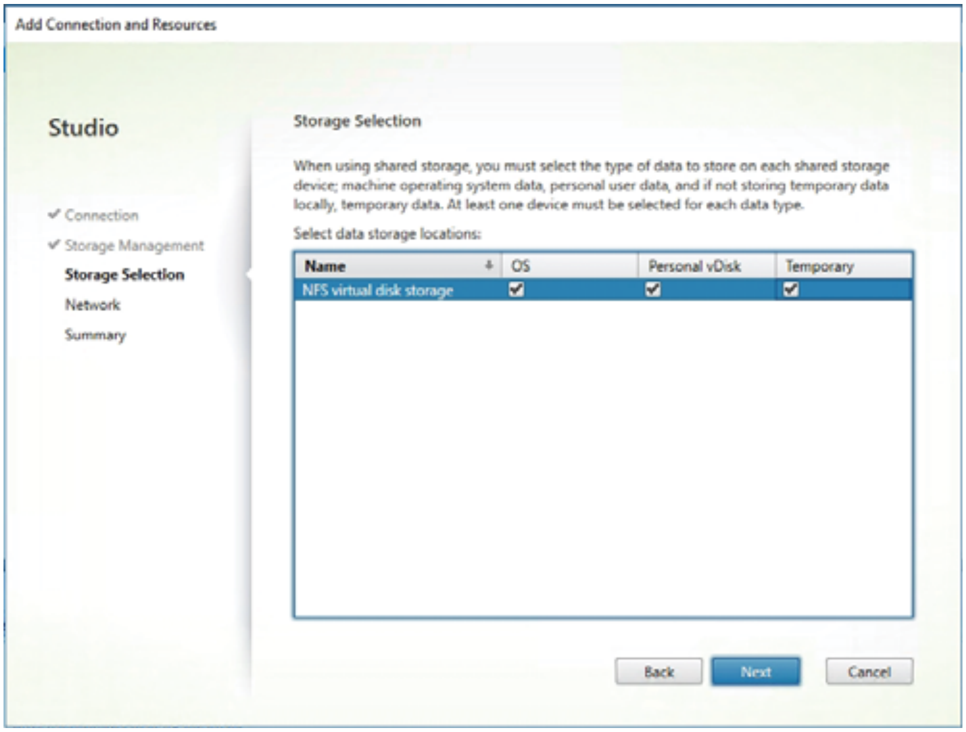
□ Manage **personal** data centrally on shared storage

Optimization technology (optional):

□ Use intellicache to reduce load on the shared storage device

4. Perform storage selection (Figure 70).

**Figure 70.** Storage Selection page



Add Connection and Resources

Studio

✓ Connection
✓ Storage Management
**Storage Selection**
Network
Summary

Storage Selection

When using shared storage, you must select the type of data to store on each shared storage device; machine operating system data, personal user data, and if not storing temporary data locally, temporary data. At least one device must be selected for each data type.

Select data storage locations:

| Name | OS | Personal vDisk | Temporary |
|---|---|---|---|
| NFS virtual disk storage | ☑ | ☑ | ☑ |

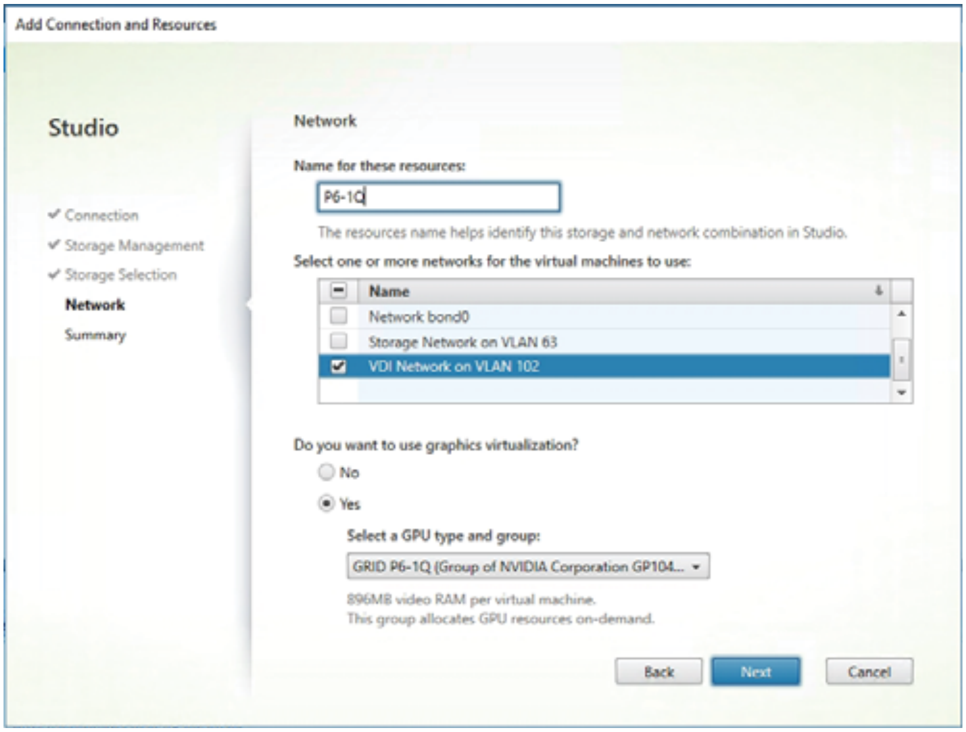
5. On the Network page, specify the following settings (Figure 71):

   a. Enter a name for the resources; this name appears in Studio to identify the storage and network combination associated with the connection.

   b. Select the networks to be used by your virtual machines.

   c. Enable the use of graphics virtualization, and from the drop-down list choose the appropriate profile type.

**Figure 71.** Specifying network settings



Add Connection and Resources

Studio

Network

Name for these resources: P6-1Q

The resources name helps identify this storage and network combination in Studio.

Select one or more networks for the virtual machines to use:

| | Name |
|---|---|
| □ | Network bond0 |
| □ | Storage Network on VLAN 63 |
| ☑ | VDI Network on VLAN 102 |

Do you want to use graphics virtualization?

○ No

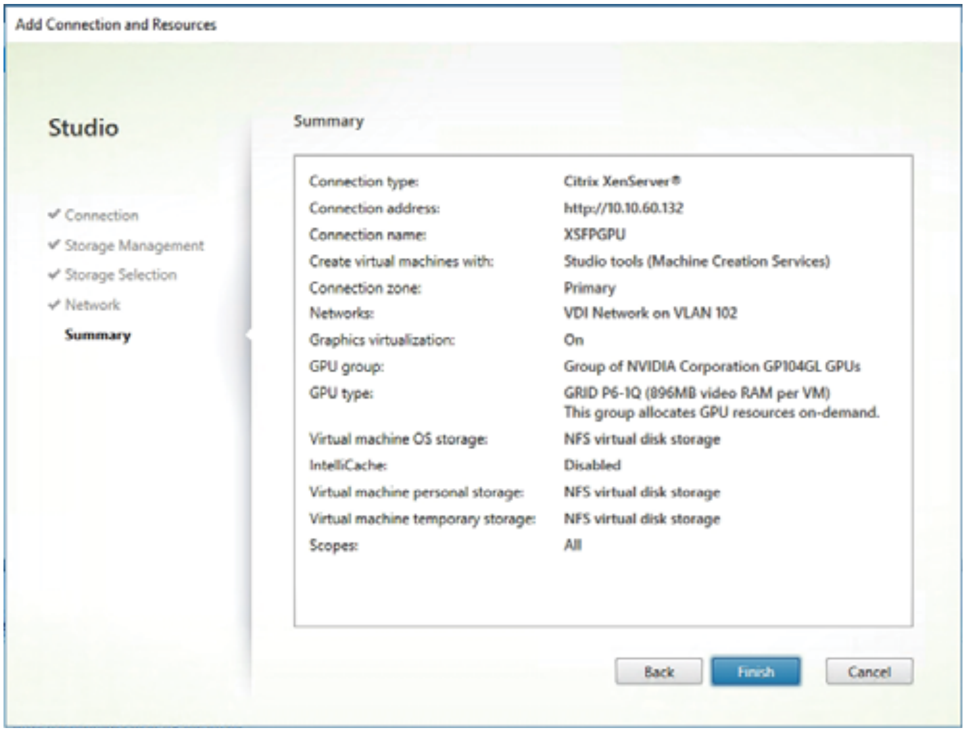
● Yes

Select a GPU type and group: GRID P6-1Q (Group of NVIDIA Corporation GP104...

896MB video RAM per virtual machine.
This group allocates GPU resources on-demand.

6. Review the Summary page and click Finish to create the connection (Figure 72).

**Figure 72.** Connection Summary page

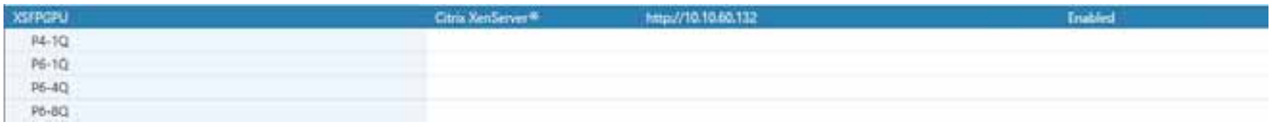


Summary

| | |
|---|---|
| Connection type: | Citrix XenServer® |
| Connection address: | http://10.10.60.132 |
| Connection name: | XSFPGPU |
| Create virtual machines with: | Studio tools (Machine Creation Services) |
| Connection zone: | Primary |
| Networks: | VDI Network on VLAN 102 |
| Graphics virtualization: | On |
| GPU group: | Group of NVIDIA Corporation GP104GL GPUs |
| GPU type: | GRID P6-1Q (896MB video RAM per VM)<br>This group allocates GPU resources on-demand. |
| Virtual machine OS storage: | NFS virtual disk storage |
| IntelliCache: | Disabled |
| Virtual machine personal storage: | NFS virtual disk storage |
| Virtual machine temporary storage: | NFS virtual disk storage |
| Scopes: | All |

After the connection has been created, you can modify it by working through the Add Connection and Resources wizard to configure additional resources to deploy a variety of the vGPUs, as shown in Figure 73.

**Figure 73.** Connection with multiple vGPU resources



| | | | |
|---|---|---|---|
| XSFPGPU | Citrix XenServer® | http://10.10.60.132 | Enabled |
| P4-1Q | | | |
| P6-1Q | | | |
| P6-4Q | | | |
| P6-8Q | | | |

## Create virtual desktops with vGPU support

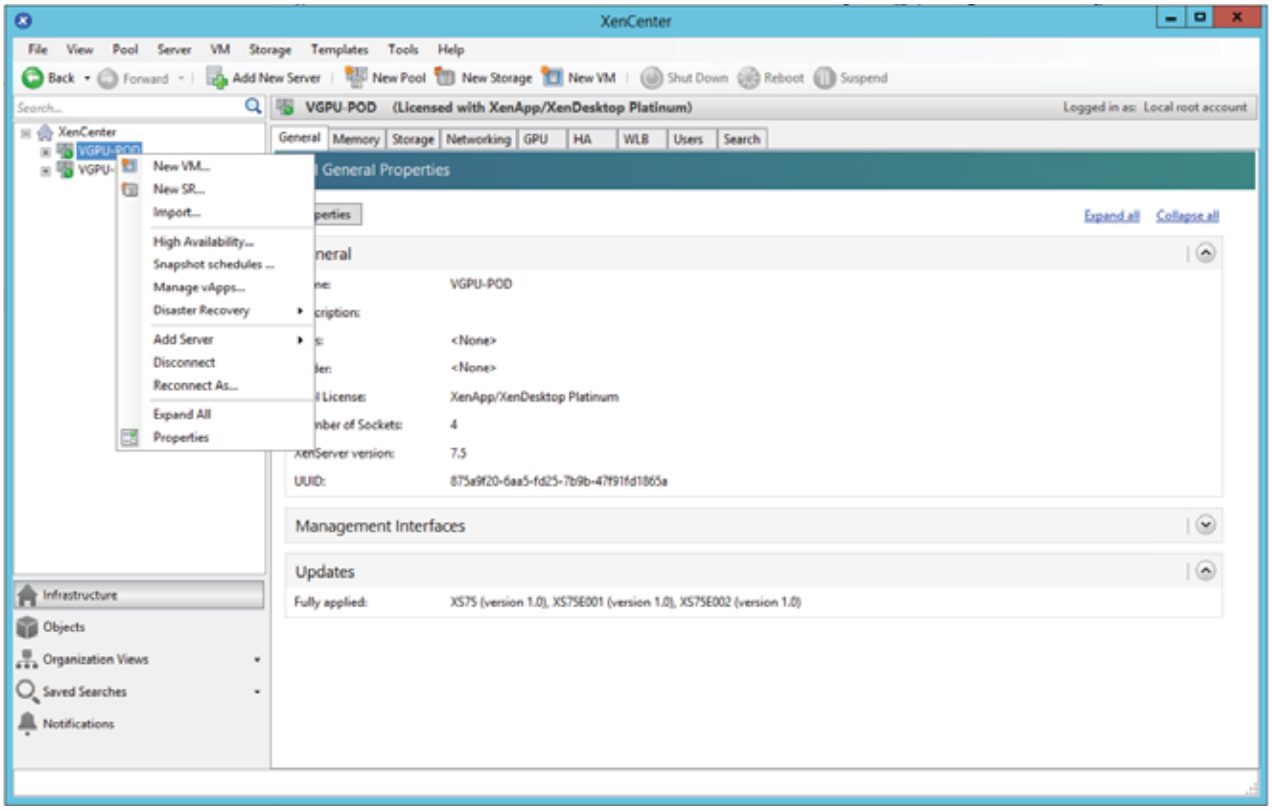
Now create virtual desktops with vGPU support.

### Create the Citrix XenDesktop base image

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

1. Using XenCenter, create a new virtual machine. To do this, right-click a host or resource pool and choose New VM. Work through the New VM wizard. Unless another configuration is specified, select the configuration settings appropriate for your environment (Figure 74).

**Figure 74.** Creating a new virtual machine in Citrix XenCenter



VGPU-POD (Licensed with XenApp/XenDesktop Platinum)

Logged in as: Local root account

ne: VGPU-POD

cription:

s: <None>

der: <None>

l License: XenApp/XenDesktop Platinum

nber of Sockets: 4

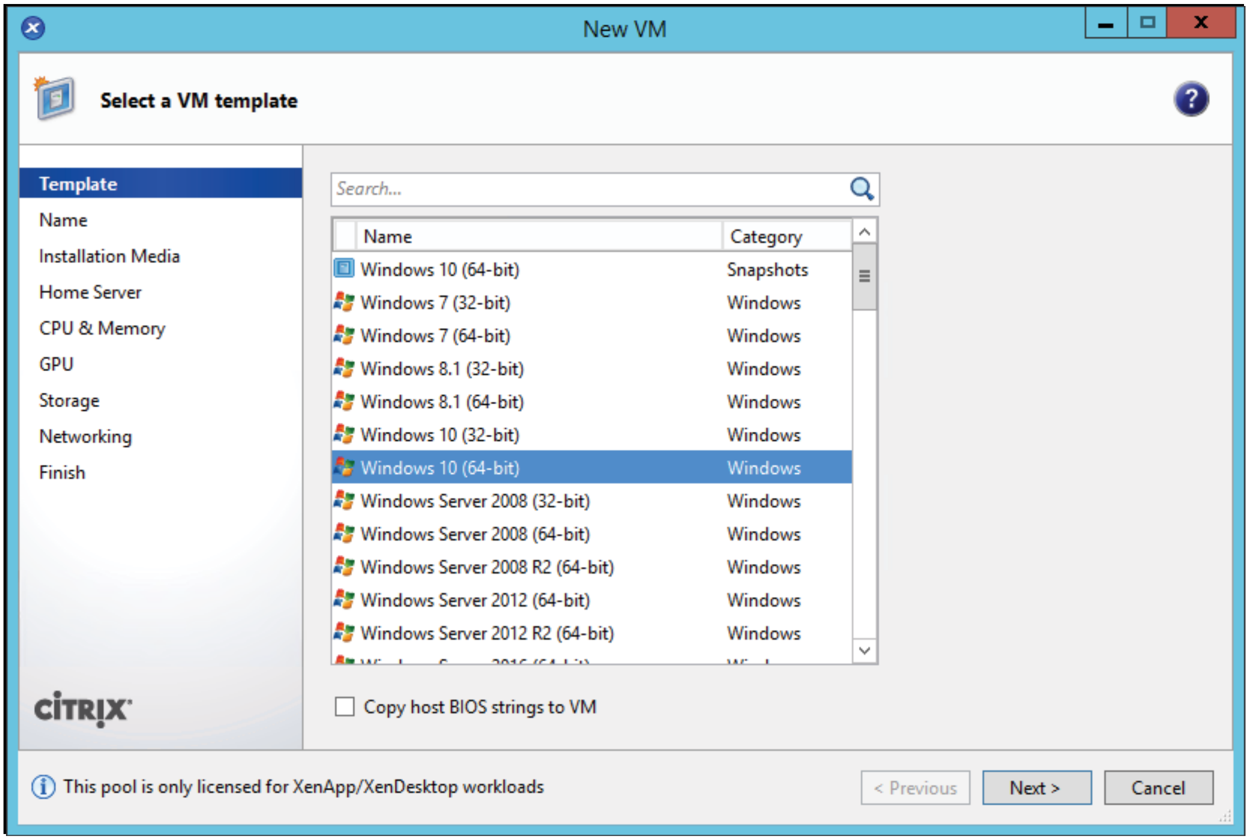
XenServer version: 7.5

UUID: 875a9f20-6aa5-fd25-7b9b-47f91fd1865a

Fully applied: XS75 (version 1.0), XS75E001 (version 1.0), XS75E002 (version 1.0)

2. Choose a virtual machine template appropriate for your environment. Figure 75 shows the Windows 10 (64-bit) template required by SPECviewperf13.

**Figure 75.** Selecting a virtual machine template



Select a VM template

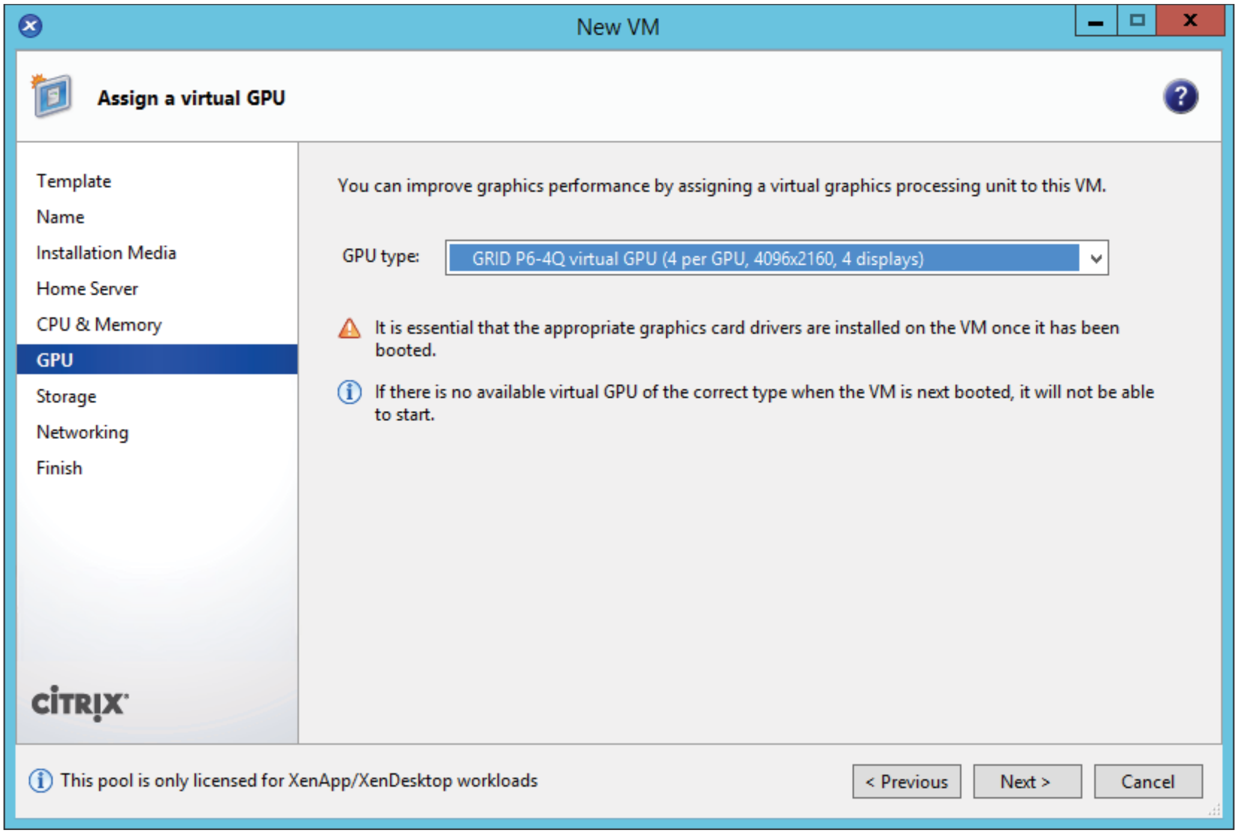
| Name | Category |
|---|---|
| Windows 10 (64-bit) | Snapshots |
| Windows 7 (32-bit) | Windows |
| Windows 7 (64-bit) | Windows |
| Windows 8.1 (32-bit) | Windows |
| Windows 8.1 (64-bit) | Windows |
| Windows 10 (32-bit) | Windows |
| Windows 10 (64-bit) | Windows |
| Windows Server 2008 (32-bit) | Windows |
| Windows Server 2008 (64-bit) | Windows |
| Windows Server 2008 R2 (64-bit) | Windows |
| Windows Server 2012 (64-bit) | Windows |
| Windows Server 2012 R2 (64-bit) | Windows |

☐ Copy host BIOS strings to VM

This pool is only licensed for XenApp/XenDesktop workloads

3. Follow New VM wizard to customize the hardware of the new virtual machine and assign the appropriate GPU type to be used by the virtual machine template (Figure 76).

**Figure 76.** Selecting a new virtual machine GPU



**Assign a virtual GPU**

You can improve graphics performance by assigning a virtual graphics processing unit to this VM.

GPU type: GRID P6-4Q virtual GPU (4 per GPU, 4096x2160, 4 displays)

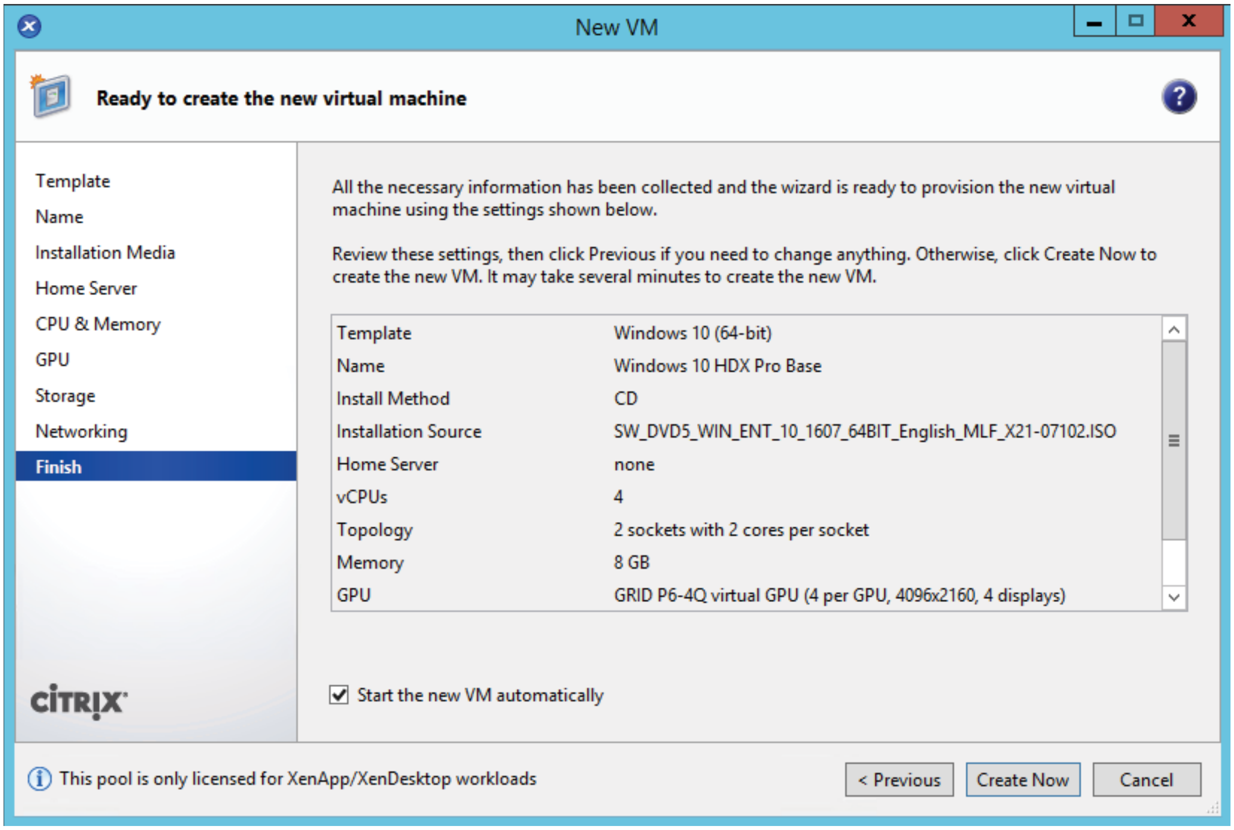
⚠ It is essential that the appropriate graphics card drivers are installed on the VM once it has been booted.

ⓘ If there is no available virtual GPU of the correct type when the VM is next booted, it will not be able to start.

ⓘ This pool is only licensed for XenApp/XenDesktop workloads

4. Continue to follow the New VM wizard. On the Finish screen, select Create Now to create a virtual machine that will be used for the base desktop image (Figure 77).

**Figure 77.** New VM summary page



All the necessary information has been collected and the wizard is ready to provision the new virtual machine using the settings shown below.

Review these settings, then click Previous if you need to change anything. Otherwise, click Create Now to create the new VM. It may take several minutes to create the new VM.

| | |
|---|---|
| Template | Windows 10 (64-bit) |
| Name | Windows 10 HDX Pro Base |
| Install Method | CD |
| Installation Source | SW_DVD5_WIN_ENT_10_1607_64BIT_English_MLF_X21-07102.ISO |
| Home Server | none |
| vCPUs | 4 |
| Topology | 2 sockets with 2 cores per socket |
| Memory | 8 GB |
| GPU | GRID P6-4Q virtual GPU (4 per GPU, 4096x2160, 4 displays) |

☑ Start the new VM automatically

This pool is only licensed for XenApp/XenDesktop workloads

**Note:** A virtual machine with a vGPU assigned will not start if ECC is enabled. As a workaround, disable ECC by entering the following commands:

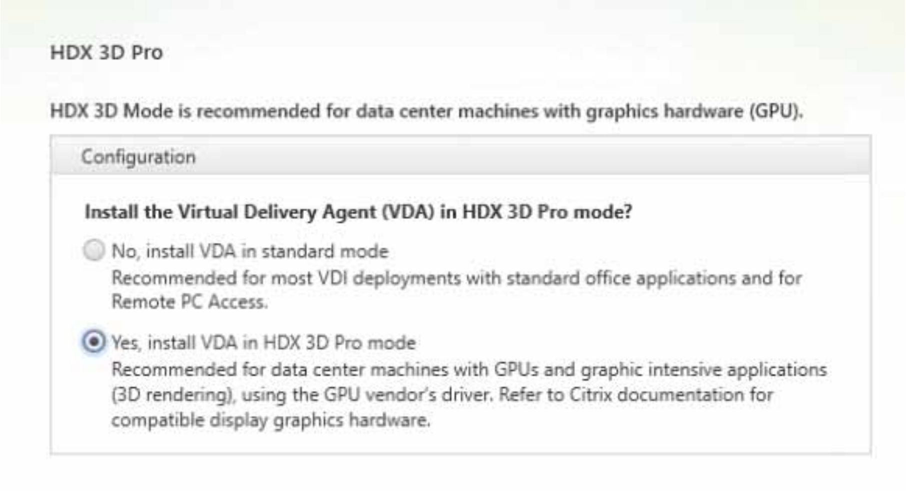
```
#nvidia-smi -i 0 -e 0
#nvidia-smi -i 1 -e 0
```

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

5. Install and configure Microsoft Windows on the virtual machine:
   a. Install XenServer Tools.
   b. Install SPECviewperf13.
   c. Join the virtual machine to the Microsoft Active Directory domain.
   d. Install or upgrade Citrix HDX 3D Pro Virtual Desktop Agent using the CLI (Figure 78).

- When you use the installer's GUI to install a VDA for a Windows desktop, simply select Yes on the HDX 3D Pro page. When you use the CLI, include the **/enable_hdx_3d_pro** option with the XenDesktop **VdaSetup.exe** command.
- To upgrade HDX 3D Pro, uninstall both the separate HDX 3D for Professional Graphics component and the VDA before installing the VDA for HDX 3D Pro. Similarly, to switch from the standard VDA for a Windows desktop to the HDX 3D Pro VDA, uninstall the standard VDA and then install the VDA for HDX 3D Pro.

**Figure 78.** Installing the VDA in HDX 3D Pro



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.

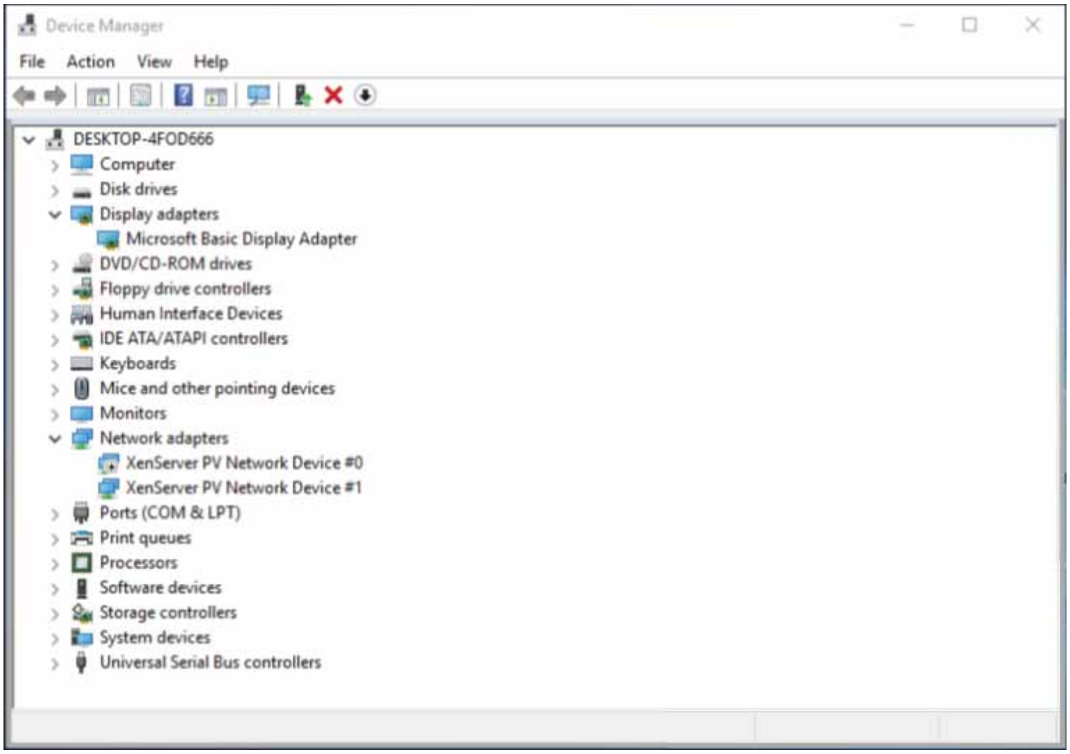
e. Optimize the Windows OS. CitrixOptimizer, the optimization tool, includes customizable templates to enable or disable Windows system services. Most Windows system services are enabled by default, but you can use the optimization tool to easily disable unnecessary services and features to improve performance.

### Install the NVIDIA vGPU software driver

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

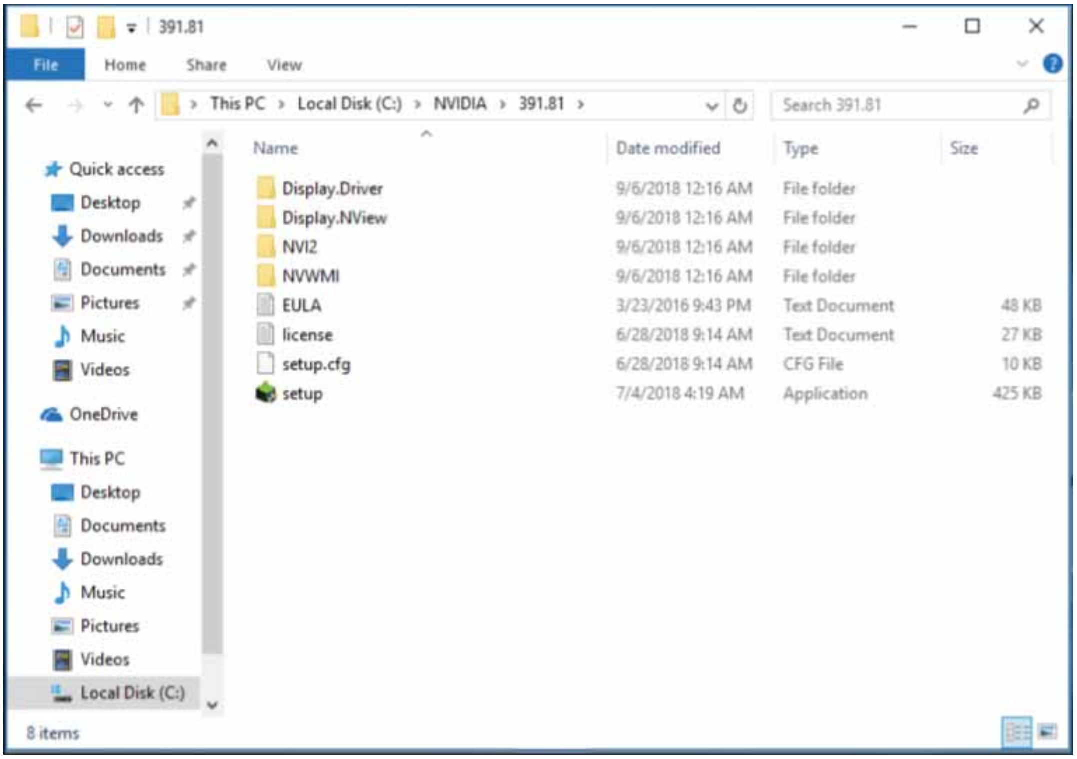
Before the NVIDIA driver is installed on the guest virtual machine, the Device Manager shows the Microsoft Basic Display adapter installed (Figure 79).

**Figure 79.** Device Manager before the NVIDIA driver is installed



1. Copy the Windows drivers from the NVIDIA GRID vGPU driver pack downloaded earlier to the master virtual machine.
2. Copy the 32- or 64-bit NVIDIA Windows driver from the vGPU driver pack to the desktop virtual machine and run setup.exe (Figure 80).

**Figure 80.** NVIDIA driver pack



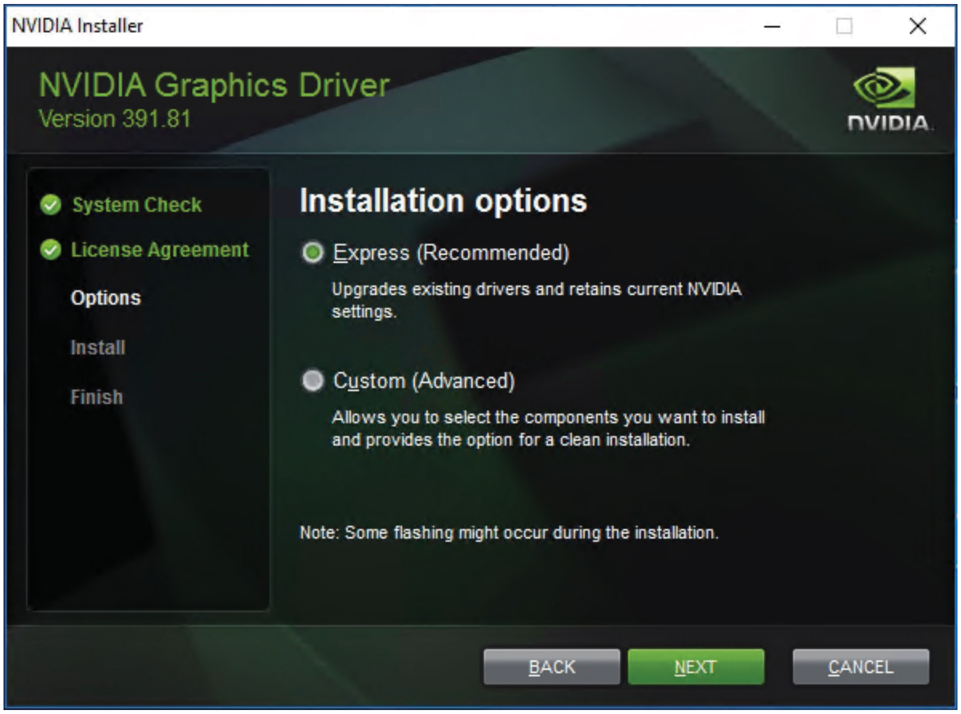
| Name | Date modified | Type | Size |
|---|---|---|---|
| Display.Driver | 9/6/2018 12:16 AM | File folder | |
| Display.NView | 9/6/2018 12:16 AM | File folder | |
| NVI2 | 9/6/2018 12:16 AM | File folder | |
| NVWMI | 9/6/2018 12:16 AM | File folder | |
| EULA | 3/23/2016 9:43 PM | Text Document | 48 KB |
| license | 6/28/2018 9:14 AM | Text Document | 27 KB |
| setup.cfg | 6/28/2018 9:14 AM | CFG File | 10 KB |
| setup | 7/4/2018 4:19 AM | Application | 425 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. Install the graphics drivers using the Express Option (Figure 81). After the installation is completed successfully, click Close (Figure 82) and restart the virtual machine.

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

**Figure 81.** Select the Express or Custom installation option



Installation options

● Express (Recommended)

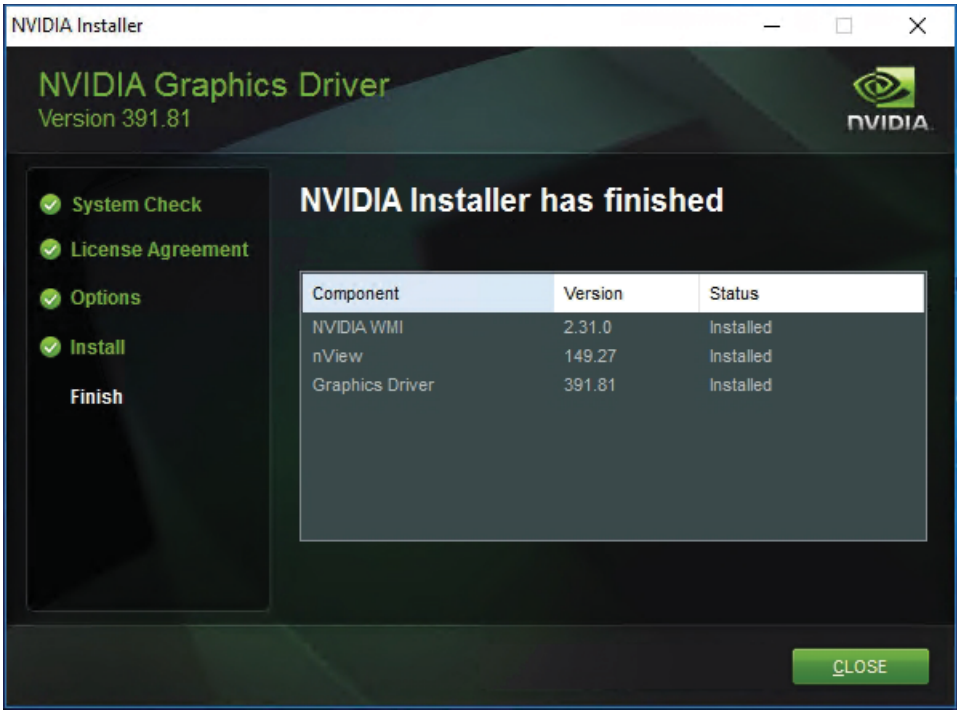
Upgrades existing drivers and retains current NVIDIA settings.

○ 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 82.** Components installed during the NVIDIA graphics driver installation process



NVIDIA Installer has finished

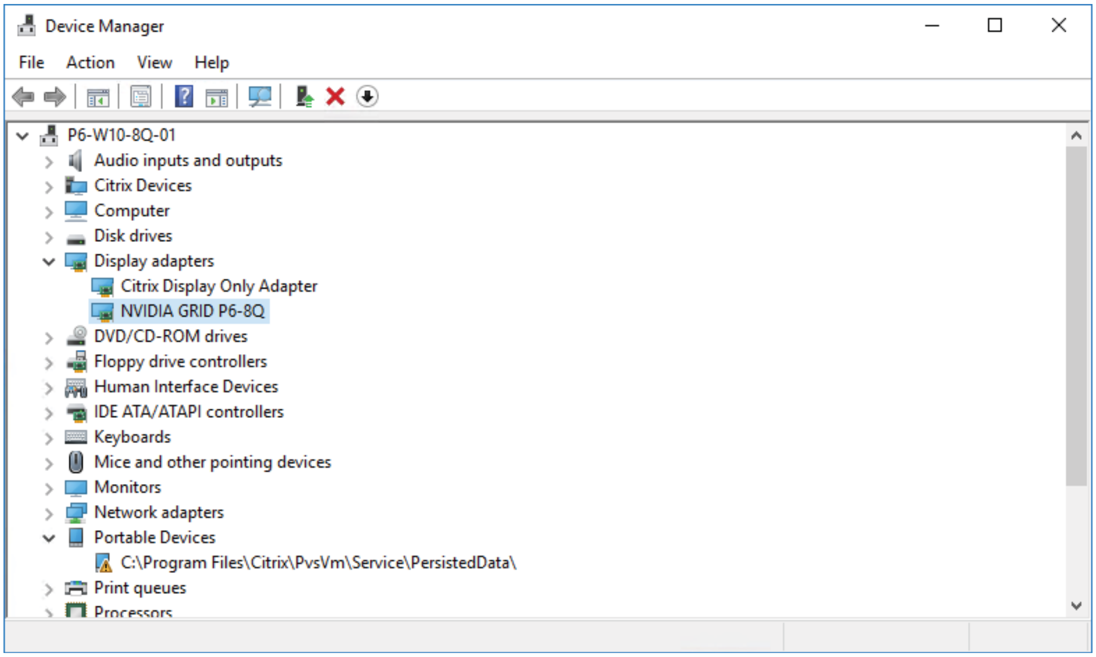
| Component | Version | Status |
|---|---|---|
| NVIDIA WMI | 2.31.0 | Installed |
| nView | 149.27 | Installed |
| Graphics Driver | 391.81 | Installed |

## Verify that the virtual machine is ready to support the vGPU

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

1. Open Windows Device Manager and expand the Display Adapter section. The device will reflect the chosen profile (Figure 83).

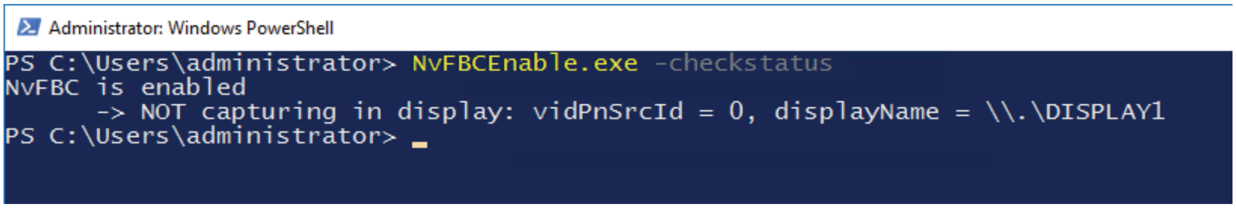
**Figure 83.** Verifying the driver installation: Device Manager



P6-W10-8Q-01
- Audio inputs and outputs
- Citrix Devices
- Computer
- Disk drives
- Display adapters
  - Citrix Display Only Adapter
  - NVIDIA GRID P6-8Q
- DVD/CD-ROM drives
- Floppy drive controllers
- Human Interface Devices
- IDE ATA/ATAPI controllers
- Keyboards
- Mice and other pointing devices
- Monitors
- Network adapters
- Portable Devices
  - C:\Program Files\Citrix\PvsVm\Service\PersistedData\
- Print queues
- Processors

2. Verify that NVIDIA GRID is enabled by using the NVFBCEnable tool provided by NVIDIA (Figure 84).

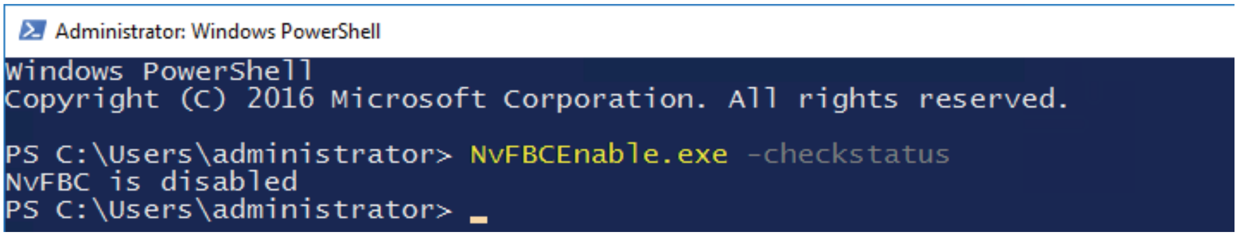
**Figure 84.** Validating the driver installation: NVFBCEnable tool



```
PS C:\Users\administrator> NvFBCEnable.exe -checkstatus
NvFBC is enabled
      -> NOT capturing in display: vidPnSrcId = 0, displayName = \\.\DISPLAY1
PS C:\Users\administrator> _
```

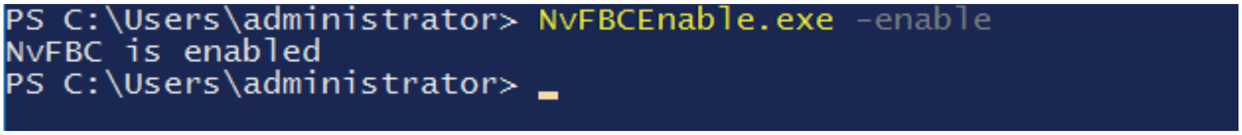
3. If NVBC is disabled as shown in Figure 85, enable it with the NVFBCEnable tool as shown in Figure 86 and then reboot the virtual machine.

**Figure 85.** Verifying the driver installation: NVFBCEnable tool (NVFBC disabled)



```
Windows PowerShell
Copyright (C) 2016 Microsoft Corporation. All rights reserved.

PS C:\Users\administrator> NvFBCEnable.exe -checkstatus
NvFBC is disabled
PS C:\Users\administrator> _
```

**Figure 86.** Verifying the driver installation: NVFBCEnable tool (enable NVFBC)



```
PS C:\Users\administrator> NvFBCEnable.exe -enable
NvFBC is enabled
PS C:\Users\administrator> _
```

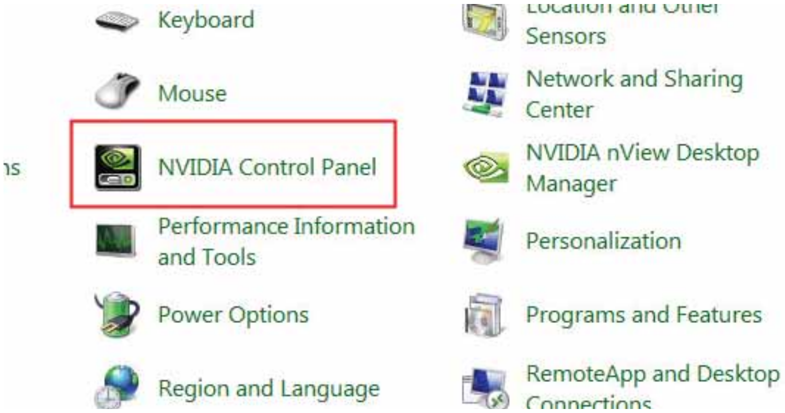
### Configure the virtual machine for an NVIDIA GRID vGPU license

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

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

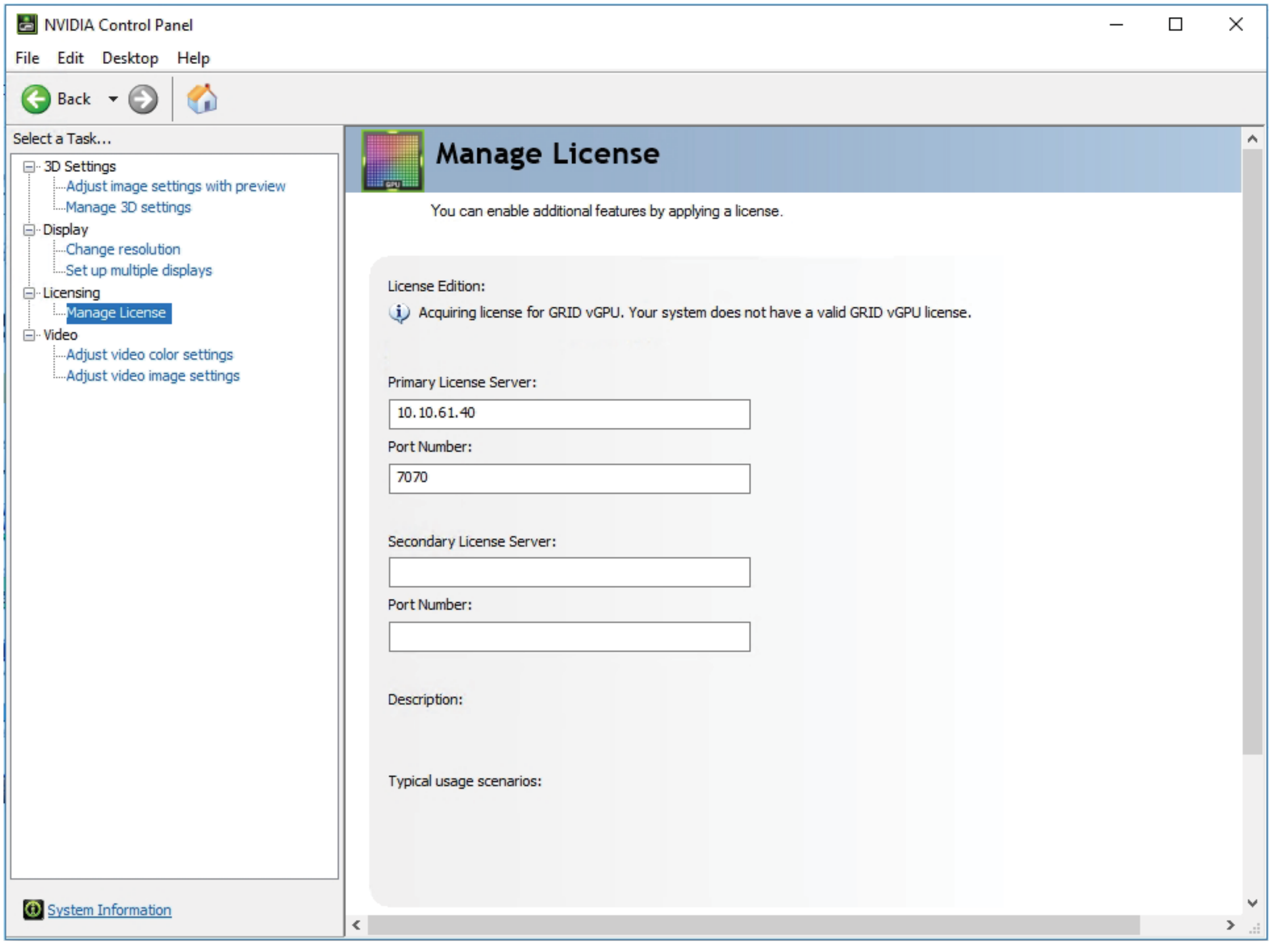
1. In the Microsoft Windows Control Panel, double-click NVIDIA Control Panel (Figure 87).

**Figure 87.** Choosing the NVIDIA control panel


ns
Keyboard
Mouse
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 (Figure 88).

**Figure 88.** Managing your license



You can enable additional features by applying a license.

License Edition:

Acquiring license for GRID vGPU. Your system does not have a valid GRID vGPU license.

Primary License Server: 10.10.61.40

Port Number: 7070

Secondary License Server:

Port Number:

Description:

Typical usage scenarios:

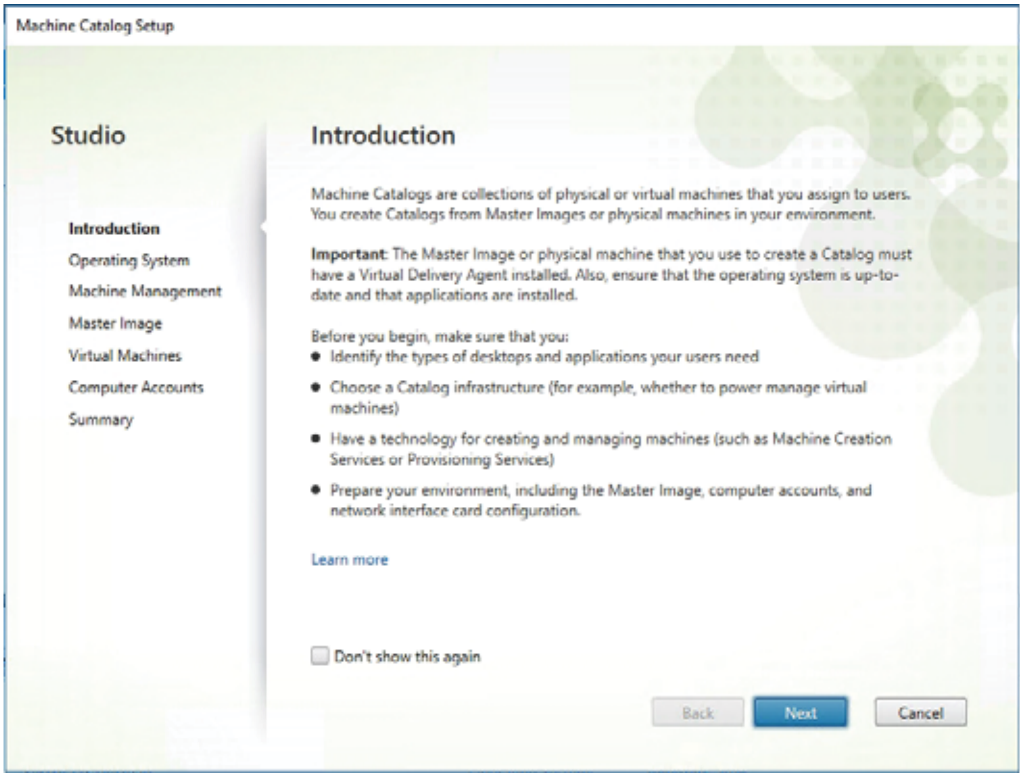
3. Click Apply.

## Deploy virtual machines with Citrix Machine Creation Services

A collection of virtual machines managed as a single entity is called a machine catalog. To create virtual machines in a catalog that have the same type of GPU using Citrix MCS, follow these steps:

1. Connect to a XenDesktop server and launch Citrix Studio.
2. Choose Create Machine Catalog from the Actions pane. Then click Next (Figure 89).

**Figure 89.** Creating a machine catalog



Machine Catalog Setup

Introduction

Machine Catalogs are collections of physical or virtual machines that you assign to users. You create Catalogs from Master Images or physical machines in your environment.

**Important**: The Master Image or physical machine that you use to create a Catalog must have a Virtual Delivery Agent installed. Also, ensure that the operating system is up-to-date and that applications are installed.

Before you begin, make sure that you:

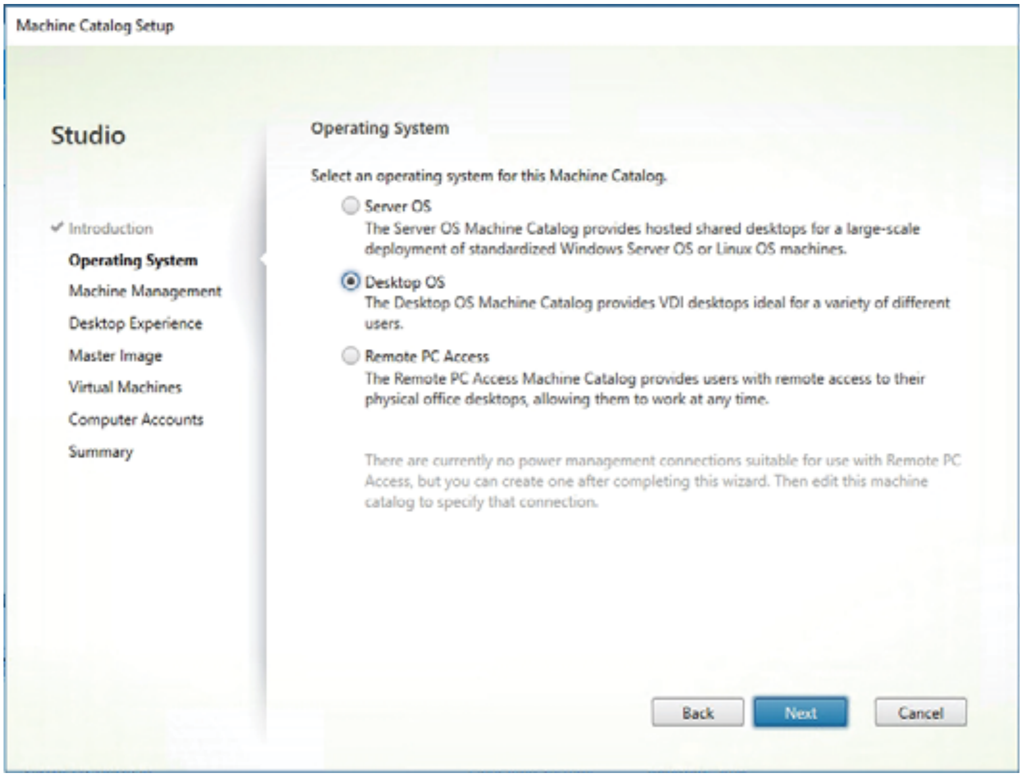
- Identify the types of desktops and applications your users need
- Choose a Catalog infrastructure (for example, whether to power manage virtual machines)
- Have a technology for creating and managing machines (such as Machine Creation Services or Provisioning Services)
- Prepare your environment, including the Master Image, computer accounts, and network interface card configuration.

Learn more

☐ Don't show this again

3. Select Desktop OS. Then click Next (Figure 90).

**Figure 90.** Selecting the OS



Machine Catalog Setup

Operating System

Select an operating system for this Machine Catalog.

○ Server OS
The Server OS Machine Catalog provides hosted shared desktops for a large-scale deployment of standardized Windows Server OS or Linux OS machines.

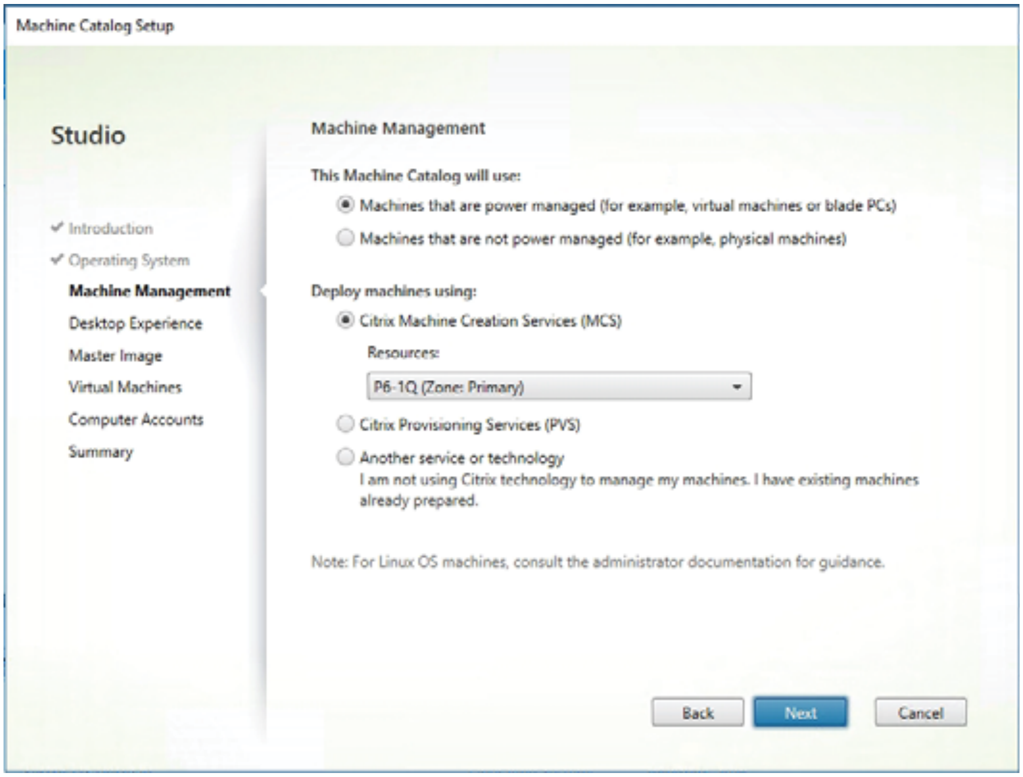
● Desktop OS
The Desktop OS Machine Catalog provides VDI desktops ideal for a variety of different users.

○ Remote PC Access
The Remote PC Access Machine Catalog provides users with remote access to their physical office desktops, allowing them to work at any time.

There are currently no power management connections suitable for use with Remote PC Access, but you can create one after completing this wizard. Then edit this machine catalog to specify that connection.

4. Select the appropriate machine management. Select the resource that will provision the virtual machine with the required GPU profile. Then click Next (Figure 91).

**Figure 91.** Selecting machine management



Machine Catalog Setup

Machine Management

This Machine Catalog will use:

● Machines that are power managed (for example, virtual machines or blade PCs)

○ Machines that are not power managed (for example, physical machines)

Deploy machines using:

● Citrix Machine Creation Services (MCS)

Resources: P6-1Q (Zone: Primary)

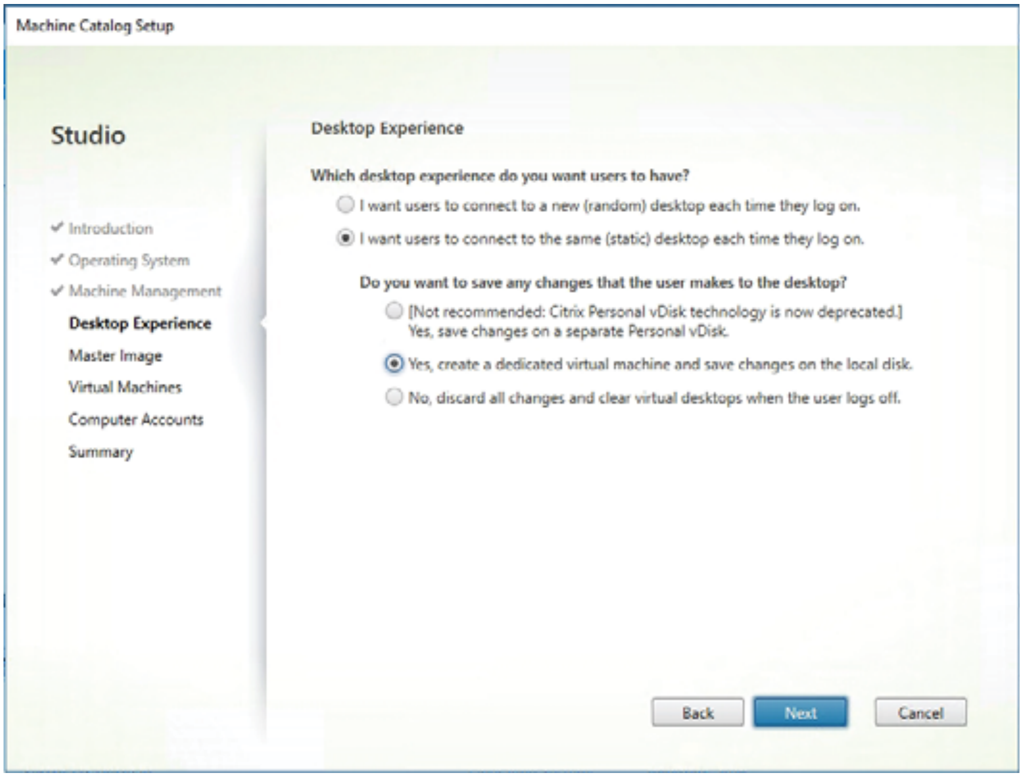
○ Citrix Provisioning Services (PVS)

○ Another service or technology
I am not using Citrix technology to manage my machines. I have existing machines already prepared.

Note: For Linux OS machines, consult the administrator documentation for guidance.

5. For Desktop Experience, select Static, Dedicated Virtual Machine. Then click Next (Figure 92).

**Figure 92.** Selecting the desktop experience



Machine Catalog Setup

Desktop Experience

Which desktop experience do you want users to have?

○ I want users to connect to a new (random) desktop each time they log on.

● I want users to connect to the same (static) desktop each time they log on.

Do you want to save any changes that the user makes to the desktop?

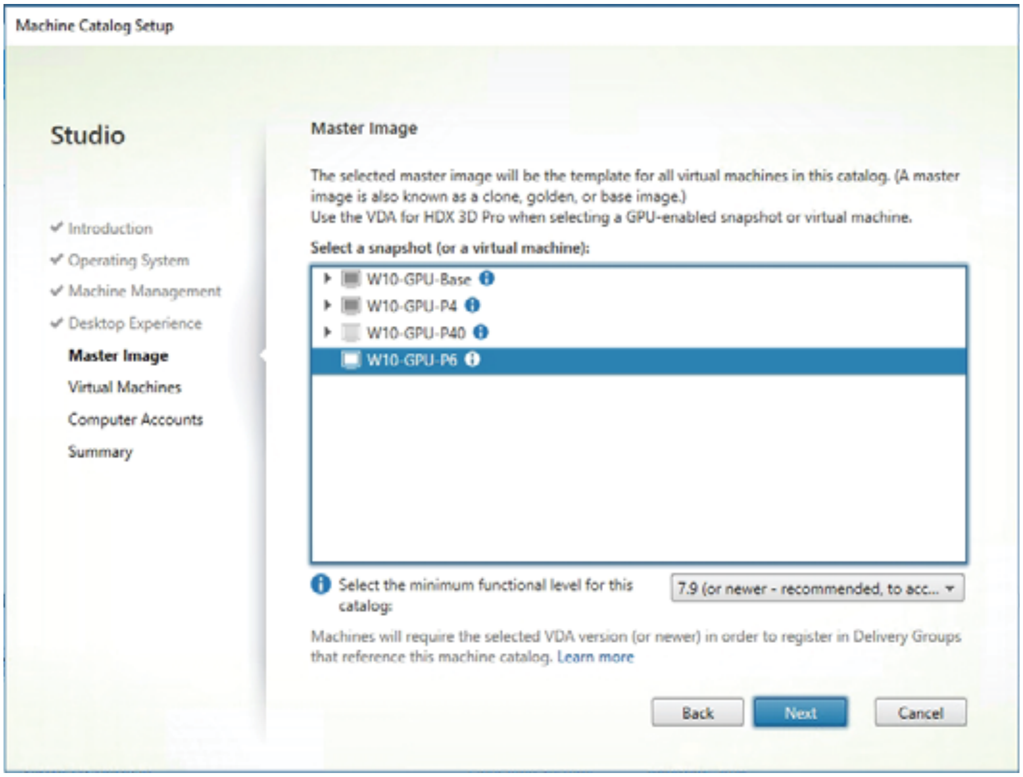
○ [Not recommended: Citrix Personal vDisk technology is now deprecated.]
Yes, save changes on a separate Personal vDisk.

● Yes, create a dedicated virtual machine and save changes on the local disk.

○ No, discard all changes and clear virtual desktops when the user logs off.

6. Select a virtual machine to be used as the catalog master image. Then click Next (Figure 93).

**Figure 93.** Selecting a virtual machine for the catalog master image



Machine Catalog Setup

Studio

- ✓ Introduction
- ✓ Operating System
- ✓ Machine Management
- ✓ Desktop Experience
- **Master Image**
- Virtual Machines
- Computer Accounts
- Summary

Master Image

The selected master image will be the template for all virtual machines in this catalog. (A master image is also known as a clone, golden, or base image.)
Use the VDA for HDX 3D Pro when selecting a GPU-enabled snapshot or virtual machine.

**Select a snapshot (or a virtual machine):**

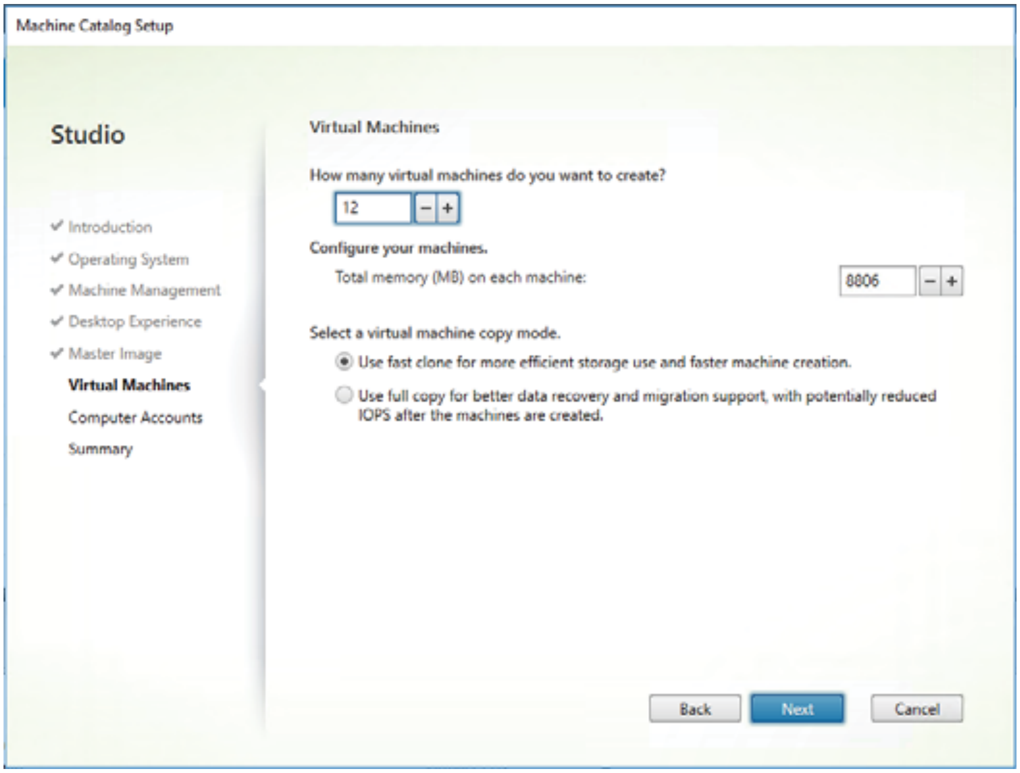
- W10-GPU-Base
- W10-GPU-P4
- W10-GPU-P40
- W10-GPU-P6

Select the minimum functional level for this catalog: 7.9 (or newer - recommended, to acc...

Machines will require the selected VDA version (or newer) in order to register in Delivery Groups that reference this machine catalog. Learn more

7. Specify the number of desktops to create and the machine configuration.
8. Set the amount of memory (in megabytes) to be used by virtual desktops.
9. Select Fast Clone as the machine copy mode.
10. Click Next (Figure 94).

**Figure 94.** Configuring the virtual machines



Virtual Machines

How many virtual machines do you want to create?

12

Configure your machines.

Total memory (MB) on each machine: 8806

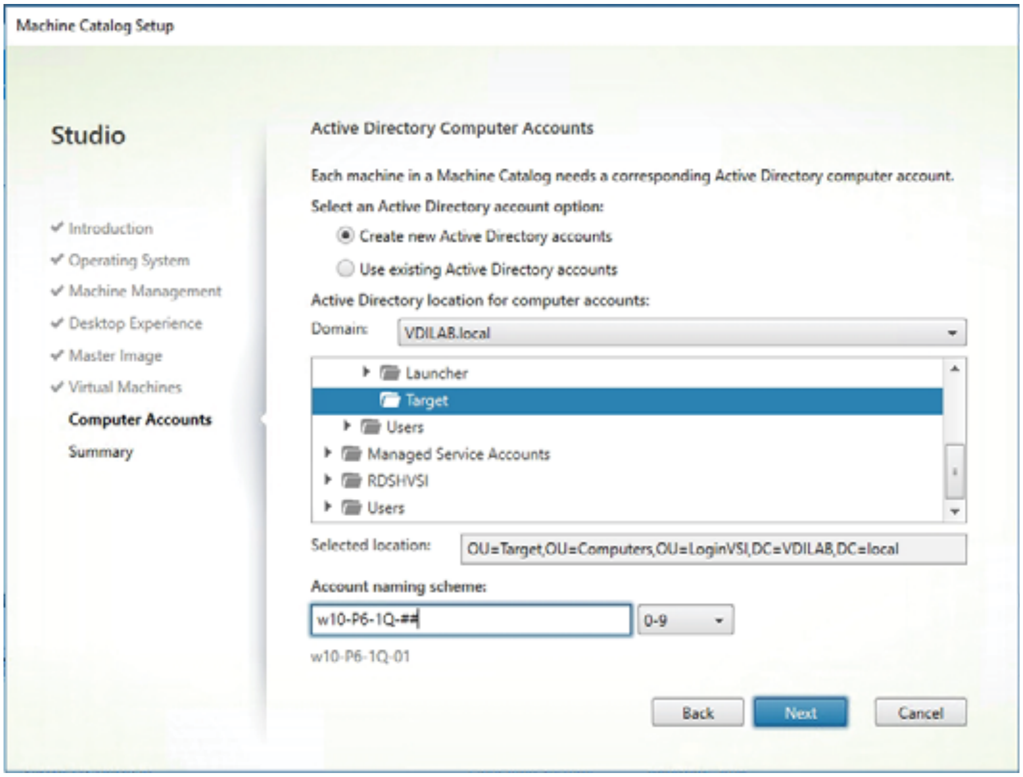
Select a virtual machine copy mode.

● Use fast clone for more efficient storage use and faster machine creation.

○ Use full copy for better data recovery and migration support, with potentially reduced IOPS after the machines are created.

11. Specify the Active Directory account naming scheme and organizational unit in which accounts will be created. Then click Next (Figure 95).

**Figure 95.** Configuring Microsoft Active Directory accounts



Machine Catalog Setup

Studio

Active Directory Computer Accounts

Each machine in a Machine Catalog needs a corresponding Active Directory computer account.

Select an Active Directory account option:

● Create new Active Directory accounts

○ Use existing Active Directory accounts

Active Directory location for computer accounts:

Domain: VDILAB.local

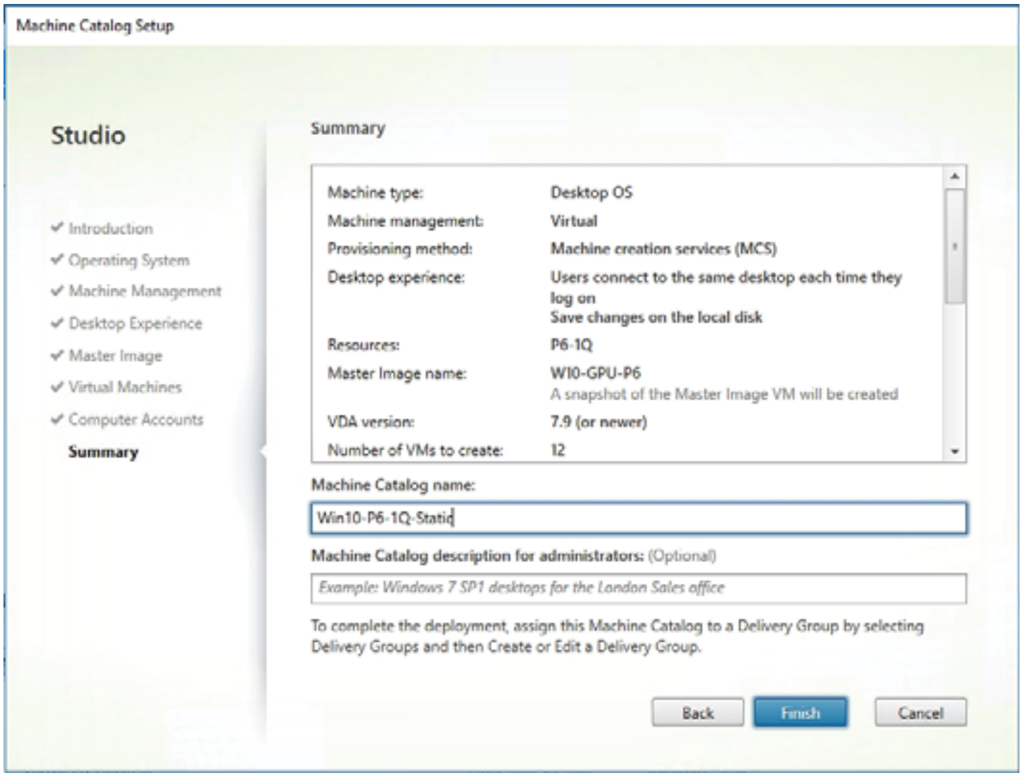
Selected location: OU=Target,OU=Computers,OU=LoginVSI,DC=VDILAB,DC=local

Account naming scheme: w10-P6-1Q-## 0-9

w10-P6-1Q-01

12. On the Summary page specify the catalog name and click Finish to start deployment (Figure 96).

**Figure 96.** Machine catalog Summary page



Summary

| | |
|---|---|
| Machine type: | Desktop OS |
| Machine management: | Virtual |
| Provisioning method: | Machine creation services (MCS) |
| Desktop experience: | Users connect to the same desktop each time they log on<br>Save changes on the local disk |
| Resources: | P6-1Q |
| Master Image name: | W10-GPU-P6<br>A snapshot of the Master Image VM will be created |
| VDA version: | 7.9 (or newer) |
| Number of VMs to create: | 12 |

Machine Catalog name: Win10-P6-1Q-Static

Machine Catalog description for administrators: (Optional) *Example: Windows 7 SP1 desktops for the London Sales office*

To complete the deployment, assign this Machine Catalog to a Delivery Group by selecting Delivery Groups and then Create or Edit a Delivery Group.

## Verify vGPU deployment

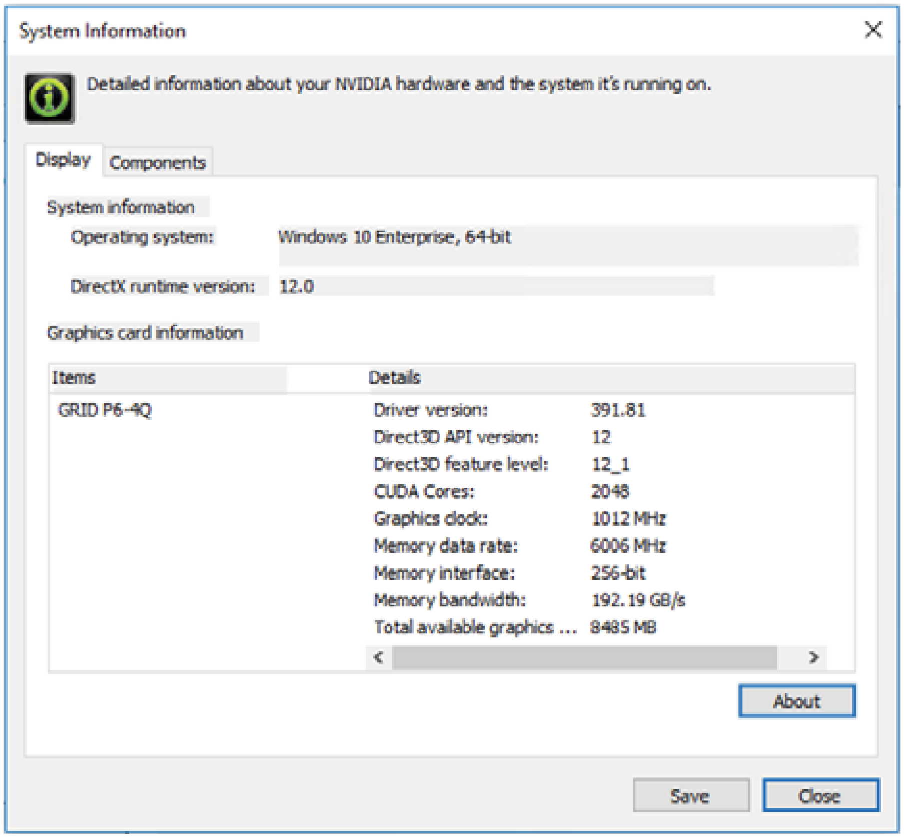
After the desktops are provisioned, use the following steps to verify vGPU deployment in the Citrix XenDesktop environment.

### 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 97).

**Figure 97.** NVIDIA control panel: System Information page



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

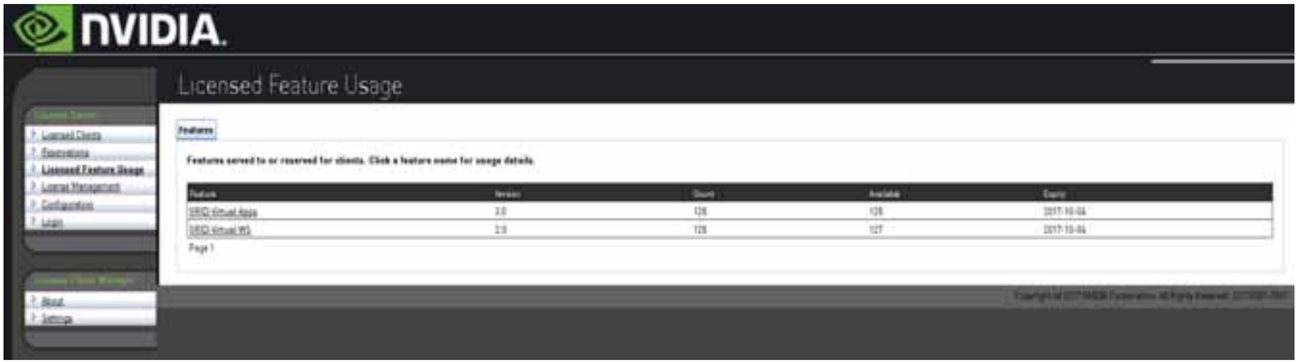
Graphics card information

| Items | Details | |
|---|---|---|
| GRID P6-4Q | Driver version: | 391.81 |
| | Direct3D API version: | 12 |
| | Direct3D feature level: | 12_1 |
| | CUDA Cores: | 2048 |
| | Graphics clock: | 1012 MHz |
| | Memory data rate: | 6006 MHz |
| | Memory interface: | 256-bit |
| | Memory bandwidth: | 192.19 GB/s |
| | Total available graphics ... | 8485 MB |

## Verify NVDIA license acquisition by desktops

A license is obtained after the virtual machine is fully booted and before the user logs on to the virtual machine (Figure 98).

**Figure 98.** NVIDIA license server: Licensed Feature Usage page



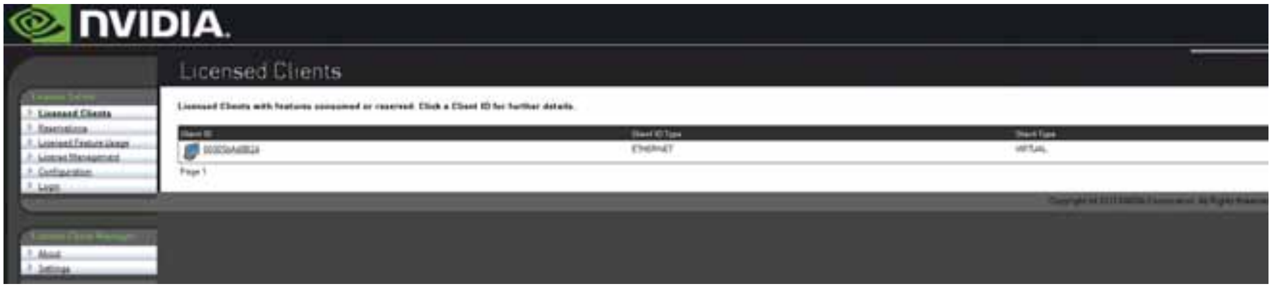
Features served to or reserved for clients. Click a feature name for usage details.

| 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

To view the details, select Licensed Clients in the left pane (Figure 99).

**Figure 99.** NVIDIA license server: Licensed Clients page



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

| Client ID | Client ID Type | Client Type |
|---|---|---|
| [illegible] | ETHERNET | VIRTUAL |

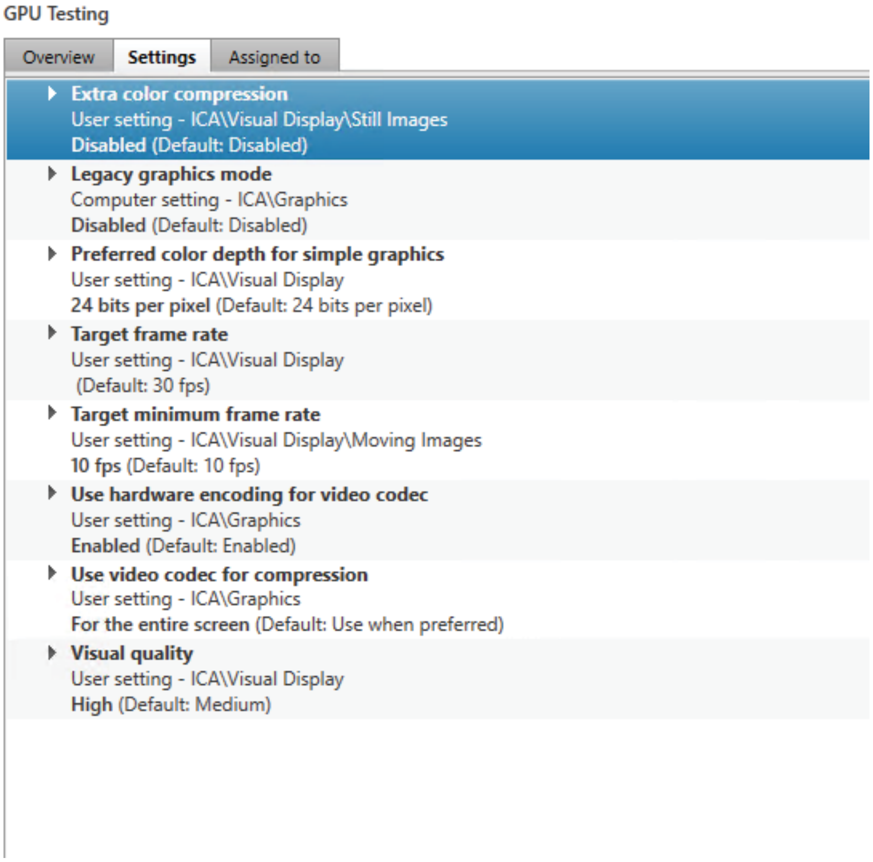
Page 1

## Citrix XenDesktop policies

Policies and profiles allow the Citrix XenDesktop environment to be easily and efficiently customized.

Citrix XenDesktop policies control user access and session environments and provide the most efficient means for controlling connection, security, and bandwidth settings. You can create policies for specific groups of users, devices, or connection types with each policy. Policies can contain multiple settings and typically are defined through Citrix Studio. (You can also use the Windows Group Policy Management Console if the network environment includes Microsoft Active Directory and permissions are set for managing group policy objects.) Figure 100 shows the policies for GPU testing used in this document.

**Figure 100.** GPU testing policies (very high-definition user experience)



**Extra color compression**
User setting - ICA\Visual Display\Still Images
Disabled (Default: Disabled)

**Legacy graphics mode**
Computer setting - ICA\Graphics
Disabled (Default: Disabled)

**Preferred color depth for simple graphics**
User setting - ICA\Visual Display
24 bits per pixel (Default: 24 bits per pixel)

**Target frame rate**
User setting - ICA\Visual Display
(Default: 30 fps)

**Target minimum frame rate**
User setting - ICA\Visual Display\Moving Images
10 fps (Default: 10 fps)

**Use hardware encoding for video codec**
User setting - ICA\Graphics
Enabled (Default: Enabled)

**Use video codec for compression**
User setting - ICA\Graphics
For the entire screen (Default: Use when preferred)

**Visual quality**
User setting - ICA\Visual Display
High (Default: Medium)

## SPECviewperf 13 benchmark results

SPECviewperf 13 is the latest version of the benchmark that measures the 3D graphics performance of systems running under the OpenGL and DirectX APIs. The benchmark's workloads, called viewsets, represent graphics content and behavior from actual applications.

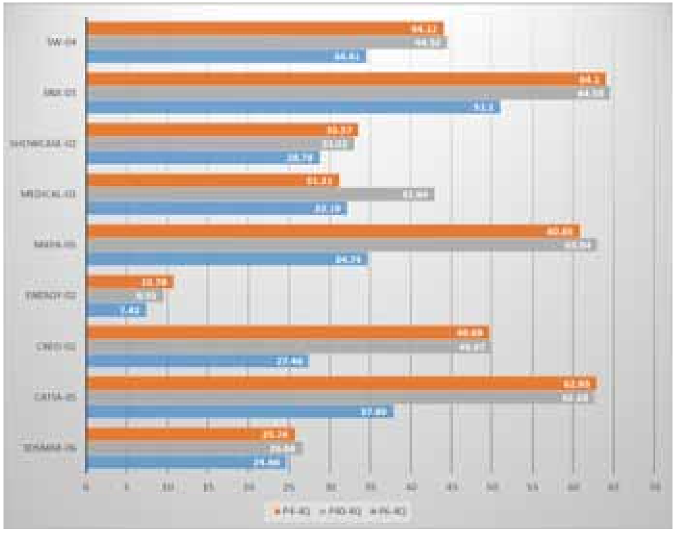
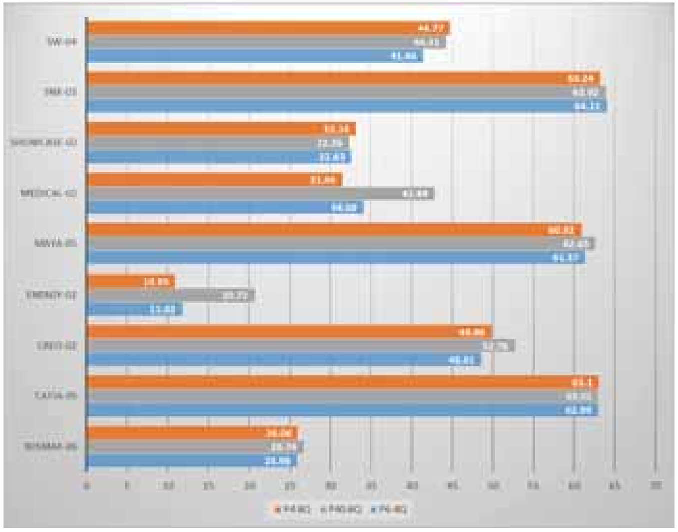
SPECviewperf 13 has the following viewsets:

- 3ds Max (3dsmax-06)
- CATIA (catia-05)
- Creo (creo-02)
- Energy (energy-02)
- Maya (maya-05)
- Medical (medical-02)
- Showcase (showcase-02)
- Siemens NX (snx-03)
- Solidworks (sw-04)

The benchmark is available for download at https://www.spec.org/gwpg/downloadindex.html#viewperf13.

Figure 101 compares all three graphics cards studied over the nine benchmark applications with NVIDIA GRID vDWS 4Q and 8Q profiles to show the impact of increased video frame buffer size by card and by application. The tests reported here used the default settings for the NVIDIA driver for the frame rate limiter (on), with the best-effort GPU scheduler. The intent was to get a clean benchmark comparison of the three Tesla cards with the default settings. In some situations and for some applications, these settings can be changed to enhance the end-user experience. The tests used the minimum 4Q profile required by the Energy application in the test suite. The 8Q profile was used for comparison testing to show the effect of additional frame buffer allocation without changing the host's CPU or memory configuration. In addition, Citrix Studio offers a frame rate policy.

**Figure 101.** SPECviewperf results for a single virtual machine comparison: P4, P6, and P40 with vDWS 4Q (left) and 8Q (right)

| 1 virtual machine 4Q on XenServer host | 1 virtual machine 8Q on XenServer host |
|---|---|
|  |  |

For the NVIDIA Tesla P4 and P40 PCIe cards running in Cisco UCS C240 M5 servers, increasing the frame buffer size from 4 GB to 8 GB had only a small effect on the composite frame rate. However, in the case of the NVIDIA Tesla P6 cards running in Cisco UCS B200 M5 servers, the 8-GB frame buffer provided a dramatic improvement in frame rate in most of the applications.

The next set of tests compared individual cards using 4-GB and 8-GB profiles with the Tesla card host running one virtual machine or the maximum number of virtual machines possible across the cards in the host.

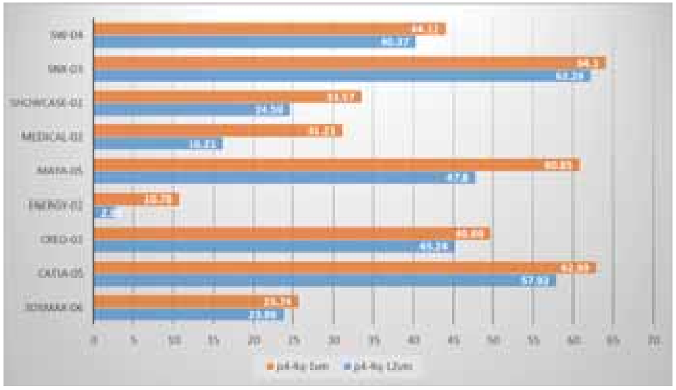
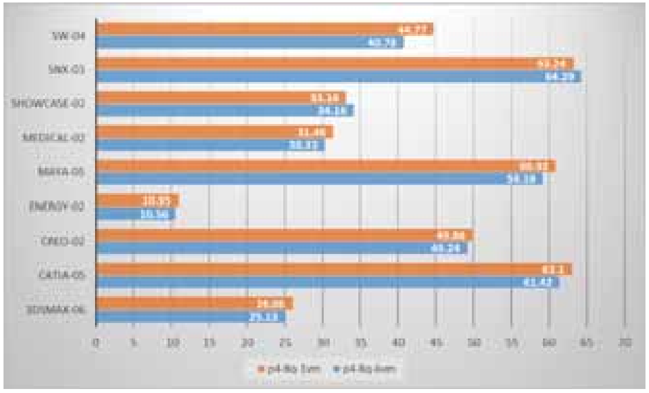
### NVIDIA Tesla P4 test results

Figure 102 shows the results for the NVIDIA Tesla P4 8-GB card running each profile size. The Cisco UCS C240 host contained six P4 cards, for a total of 48 GB of frame buffer space across the server.

The left chart shows the P4 performance with one virtual machine running a 4-GB profile (orange) and 12 virtual machines running 4-GB profiles (blue) on the same host (two test cycles).

The right chart shows the P4 performance with one virtual machine running an 8-GB profile (orange) and six virtual machine running 8-GB profiles (blue) on the same host (two test cycles).

**Figure 102.** SPECviewperf results for P4 4Q and 8Q profile tests: Single virtual machine versus maximum host density based on profile size

<table>
<tr><th>P4 4Q profile testing</th><th>P4 8Q profile testing</th></tr>
<tr><td>
</td><td>
</td></tr>
</table>

This test reveals that using a P4 card with twelve 4-GB profiles across the six cards in the server has a measurable negative impact on application frame rate. The impact varies by application, with some applications, such as Siemens NX and 3DSMAX, showing a light impact, and others, such as Medical and Energy, showing a substantial impact.

The tests of the P4 card with six virtual machines with 8-GB profiles (each virtual machine getting the full power of the P4 card) shows the expected results: near parity in performance with a single virtual machine running on a single card. The differences can be explained by the traffic on the PCI bus and increased CPU utilization.

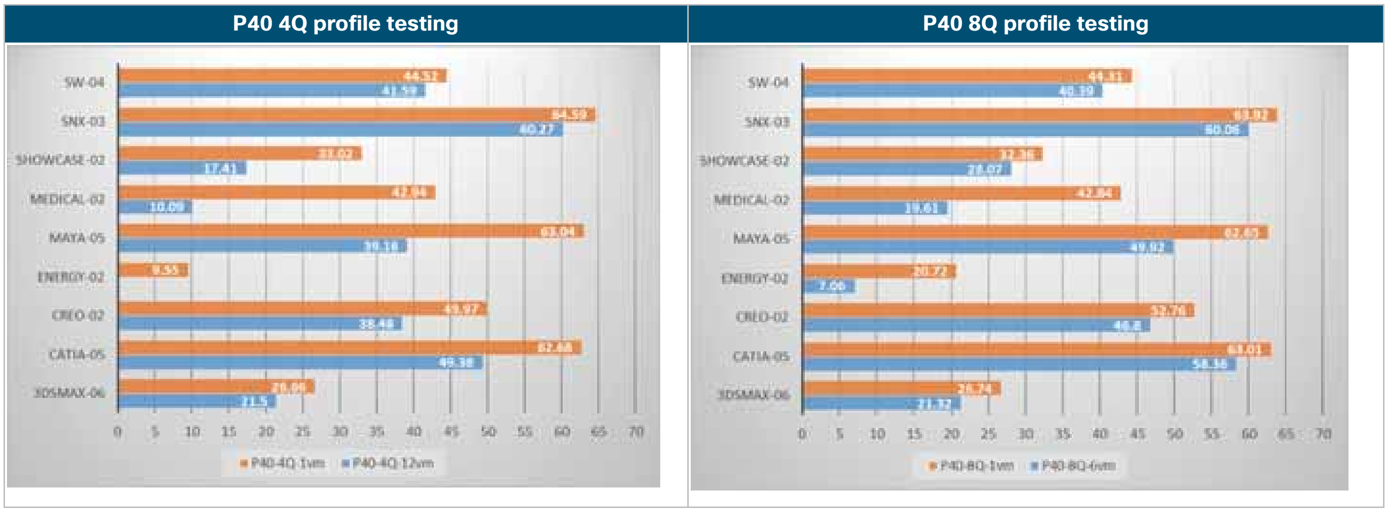
### NVIDIA Tesla P40 test results

Figure 103 shows the results for the NVIDIA Tesla P40 24-GB card running the 4-GB and 8-GB profile sizes. The Cisco UCS C240 host contained two P40 cards, for a total of 48 GB of frame buffer space across the server.

The left chart shows the P40 performance with one virtual machine running a 4-GB profile (orange) and 12 virtual machines running 4-GB profiles (blue) on the same host (two test cycles).

The right chart shows the P4 performance with one virtual machine running an 8-GB profile (orange) and six virtual machines running 8-GB profiles (blue) on the same host (two test cycles).

**Figure 103.** SPECviewperf results for P40 4Q and 8Q profile tests: Single virtual machine versus maximum host density based on profile size


P40 4Q profile testing

P40 8Q profile testing

**Note:** For the 4-GB frame buffer tests, the impact of running six virtual machines on each Tesla P40 card versus a single virtual machine on one card is pronounced. In fact, we were unable to complete benchmark test cycles with six virtual machines on each Tesla P40 because the Energy application failed on multiple virtual machines during the test. Therefore, the chart in Figure 103 was generated by running the test without that application.

For the 8-GB frame buffer tests, the variation in frame buffer with one virtual machine and with six virtual machines was not quite as pronounced as what was seen with the 4-GB buffer. The Energy application did run with the 8-GB frame buffer with multiple virtual machines per card.

P40 4Q tests with 12 virtual machines (maximum density with two NVIDIA P40 cards installed in the Cisco UCS C240 server) were performed without the Energy viewset. The Energy viewset failed with a "TRD Detected" error on all 12 virtual machines.

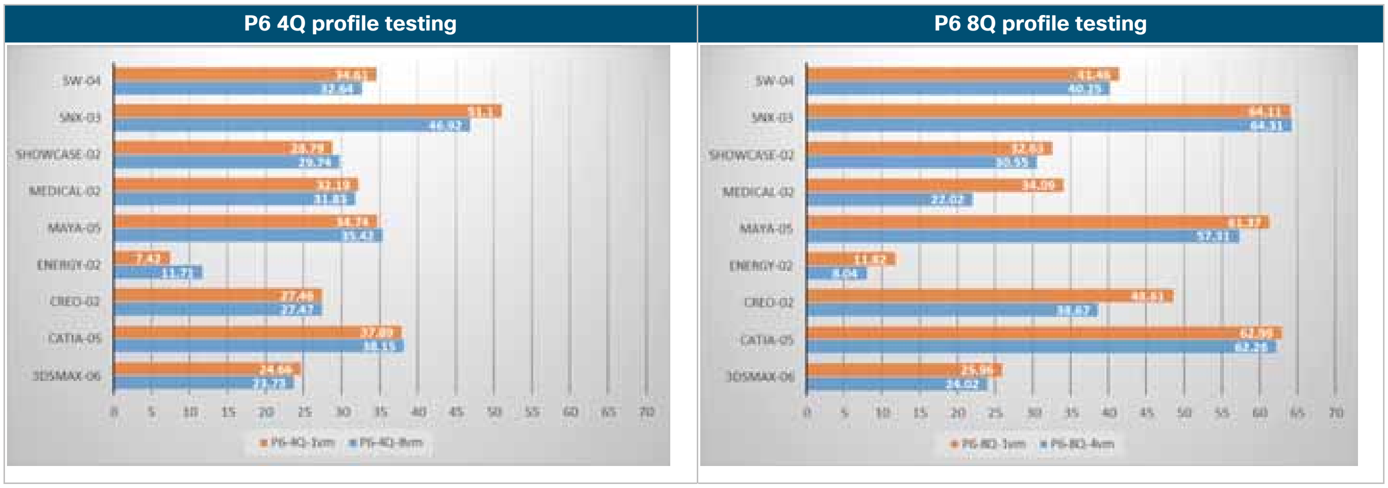
### NVIDIA Tesla P6 test results

Figure 104 shows the results for the NVIDIA Tesla P6 16-GB card running the 4-GB and 8-GB profile sizes. The Cisco B200 M5 host contained two P6 cards, for a total of 32-GB of frame buffer space across the server.

The left chart shows the P6 performance with one virtual machine running a 4-GB profile (orange) and eight virtual machines running 4-GB profiles (blue) on the same host (two test cycles).

The right chart shows the P6 performance with one virtual machine running an 8-GB profile (orange) and four virtual machines running 8-GB profiles (blue) on the same host (two test cycles).

**Figure 104.** SPECviewperf results for P6 4Q and 8Q profile tests: Single virtual machine versus maximum host density based on profile size


P6 4Q profile testing

P6 8Q profile testing

The Tesla P6 running the 4-GB profile showed the best performance at scale across the application set. The Tesla P6 with the 8-GB profile showed a significant performance improvement for the tests with one virtual machine and multiple virtual machines on most of the application sets.

## Host CPU utilization test results

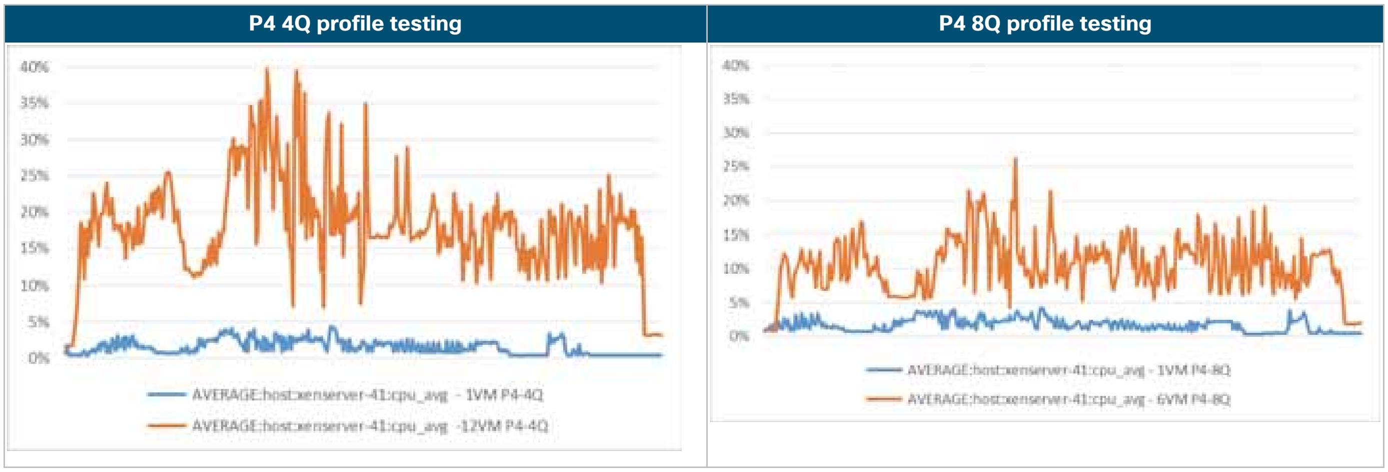
The NVIDIA Tesla GPUs work in concert with the host's Intel Xeon Scalable family processors. The following sections discuss CPU utilization during the same tests described in the preceding sections.

Figure 105 presents the data for the Tesla P4 4Q and 8Q profiles.

Figure 106 presents the data for the Tesla P40 4Q and 8Q profiles.

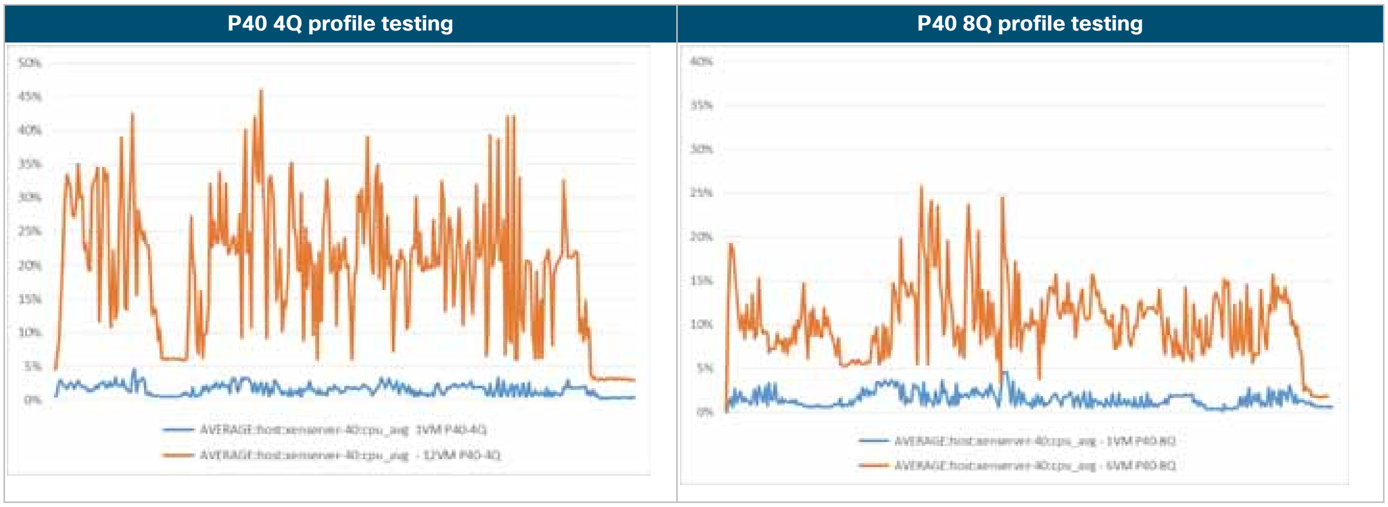
Figure 107 presents the data for the Tesla P6 4Q and 8Q profiles.

**Figure 105.** Cisco UCS C240 CPU utilization results for SPECviewperf P-4 4Q and 8Q profile tests: Single virtual machine versus maximum density


P4 4Q profile testing

P4 8Q profile testing

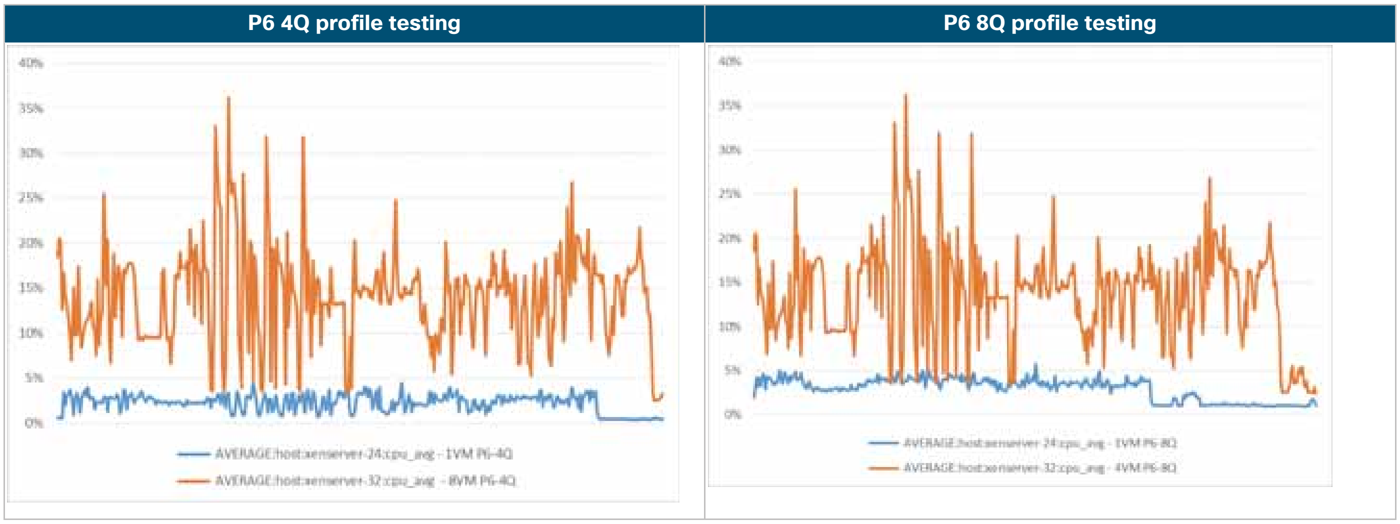
**Figure 106.** Cisco UCS C240 CPU utilization results for SPECviewperf P40 4Q and 8Q profile tests: Single virtual machine versus maximum density



P40 4Q profile testing

P40 8Q profile testing

**Figure 107.** Cisco UCS B200 CPU utilization results for SPECviewperf P6 4Q and 8Q profile tests: Single virtual machine versus maximum density



P6 4Q profile testing

P6 8Q profile testing

## Host GPU utilization test results

The NVIDIA Tesla GPUs work in concert with the host's Intel Xeon Scalable family processors. The following sections discuss GPU utilization during the same tests described in the preceding sections.

Figure 108 presents the data for the Tesla P4 4Q and 8Q profiles.

Figure 109 presents the data for the Tesla P40 4Q and 8Q profiles.

Figure 110 presents the data for the Tesla P6 4Q and 8Q profiles.

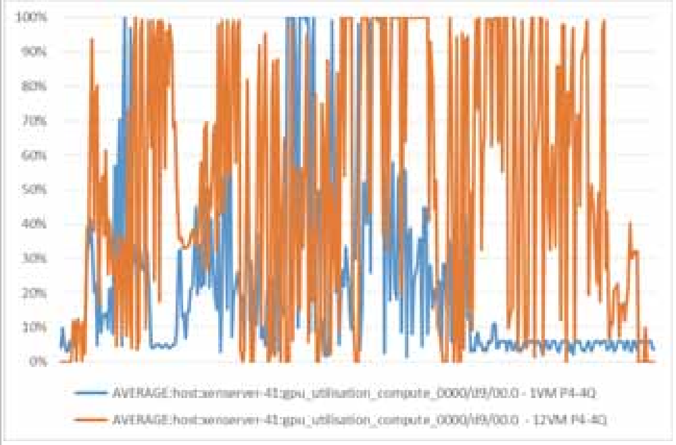
The blue plot in each graph represents GPU utilization during the tests using a single virtual machine.

The orange plot in each graph represents GPU utilization during the tests using multiple virtual machines.

**Note:** Even with a single virtual machine, there are times when the full GPU is utilized.

**Figure 108.** NVIDIA P4 results for SPECviewperf P4 4Q and 8Q profile tests: Single virtual machine versus maximum density

**P4 4Q profile testing**



**P4 8Q profile testing**

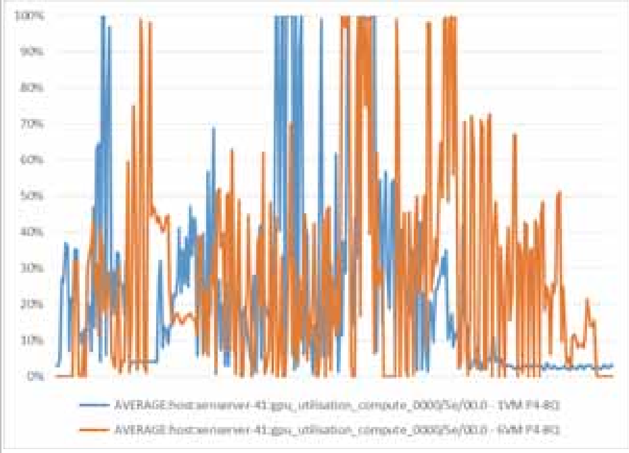
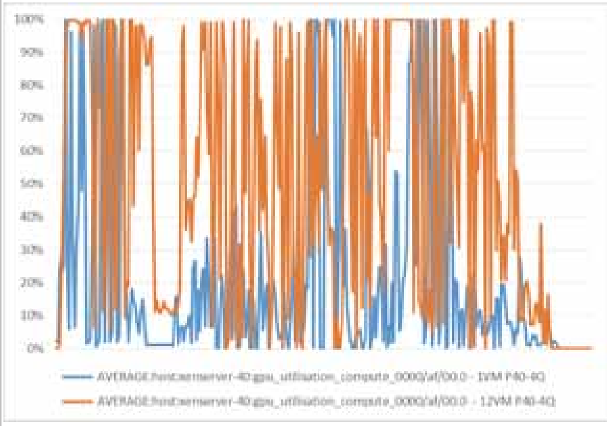


**Figure 109.** NVIDIA P40 results for SPECviewperf P40 4Q and 8Q profile tests: Single virtual machine versus maximum density

**P40 4Q profile testing**



**P40 8Q profile testing**

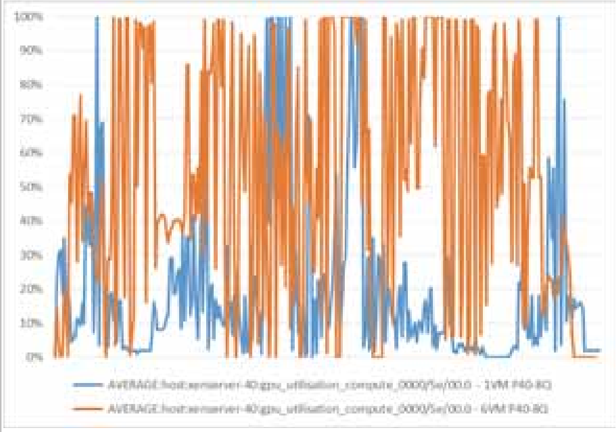
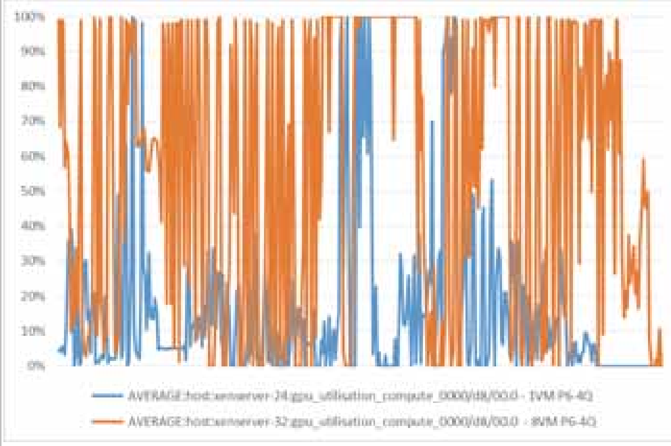
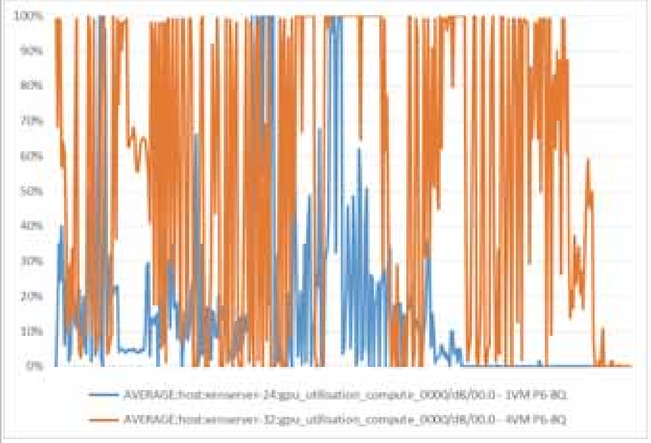


**Figure 110.** NVIDIA P6 results for SPECviewperf P6 4Q and 8Q profile tests: Single virtual machine versus maximum density

| P6 4Q profile testing | P6 8Q profile testing |
|---|---|
|  |  |

### Storage utilization test results

**Note:** Storage utilization testing was not performed for this reference architecture.

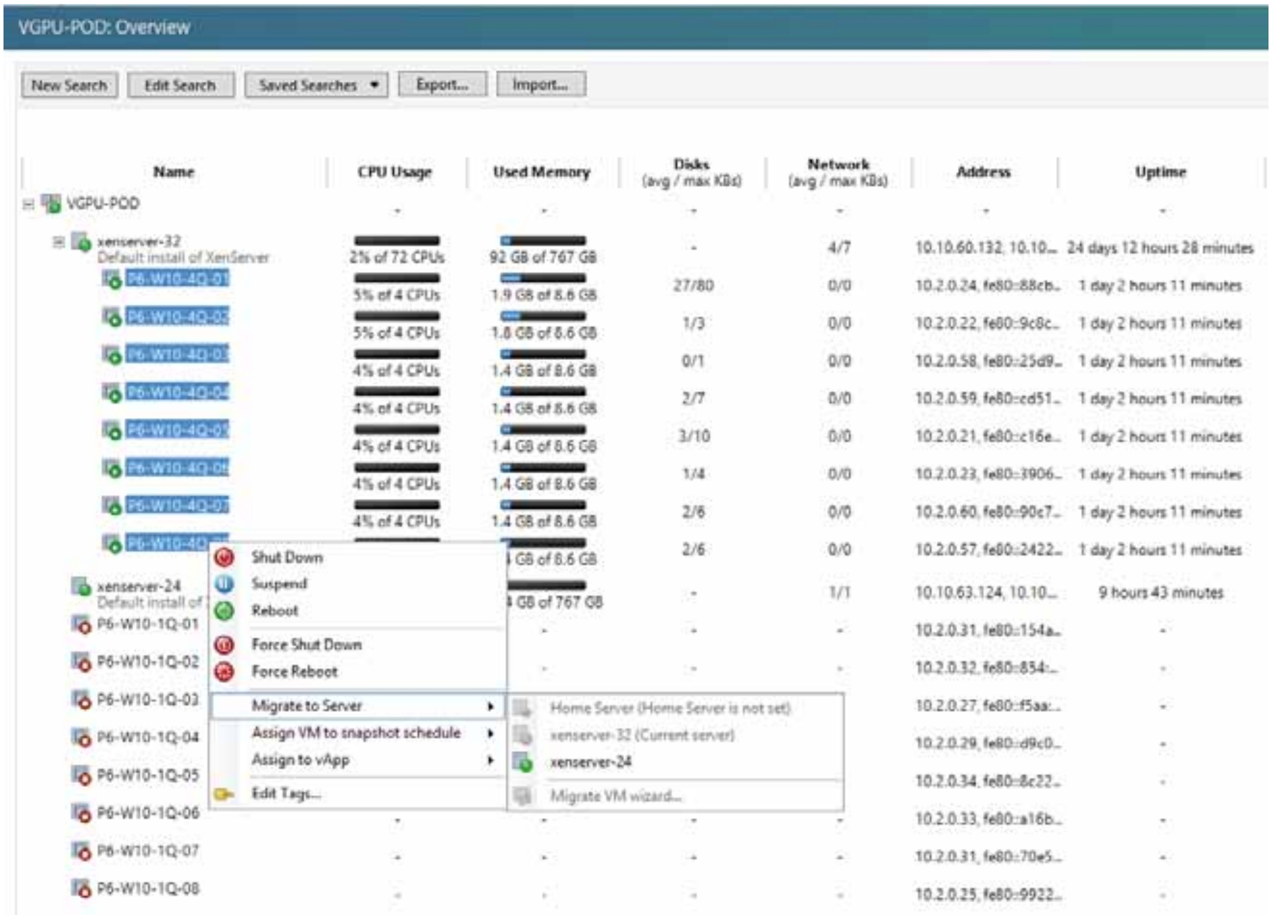
Typically, storage utilization is documented using a full-scale test with 6000 users. This testing was completed in the Cisco Validated Design that is the basis for this document. During the testing of 6000 users, storage latency was exceptional, at less than 1 millisecond. Not enough 3D graphics virtual desktops were created in the environment to perform storage utilization testing. The storage is more than capable of handling the 3D-user load with less than a 1-millisecond access time based on the test results from the Cisco Validated Design: FlexPod Datacenter with Citrix XenDesktop and XenApp 7.15 and VMware vSphere 6.5 Update 1 for 6000 Seats.

## Live vGPU-enabled virtual machine Citrix XenMotion operations

vGPU XenMotion enables a virtual machine that uses a vGPU to perform Citrix XenMotion, Storage XenMotion, and virtual machine Suspend operations. Virtual machines with vGPU XenMotion capabilities can be migrated to avoid downtime and can be included in high-availability deployments. For more information about support and restrictions, refer to Configuring Citrix XenServer 7.5.

1. To initiate XenMotion operations for vGPU-enabled virtual machines, select the virtual machines in XenCenter, right-click, and select Migrate to Server. The list will show compatible servers, as shown in Figure 111.

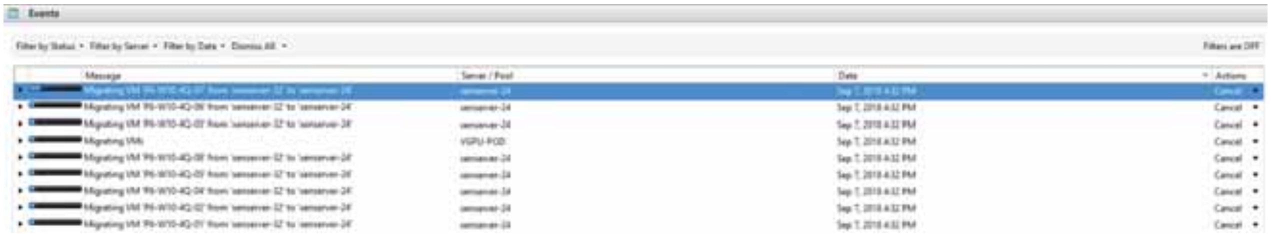
**Figure 111.** Citrix XenCenter vGPU-enabled virtual machine XenMotion operations



| Name | CPU Usage | Used Memory | Disks<br>(avg / max KBs) | Network<br>(avg / max KBs) | Address | Uptime |
|---|---|---|---|---|---|---|
| VGPU-POD | - | - | - | - | - | - |
| xenserver-32<br>Default install of XenServer | 2% of 72 CPUs | 92 GB of 767 GB | - | 4/7 | 10.10.60.132, 10.10... | 24 days 12 hours 28 minutes |
| P6-W10-4Q-01 | 5% of 4 CPUs | 1.9 GB of 8.6 GB | 27/80 | 0/0 | 10.2.0.24, fe80::88cb... | 1 day 2 hours 11 minutes |
| P6-W10-4Q-02 | 5% of 4 CPUs | 1.8 GB of 8.6 GB | 1/3 | 0/0 | 10.2.0.22, fe80::9c8c... | 1 day 2 hours 11 minutes |
| P6-W10-4Q-03 | 4% of 4 CPUs | 1.4 GB of 8.6 GB | 0/1 | 0/0 | 10.2.0.58, fe80::25d9... | 1 day 2 hours 11 minutes |
| P6-W10-4Q-04 | 4% of 4 CPUs | 1.4 GB of 8.6 GB | 2/7 | 0/0 | 10.2.0.59, fe80::cd51... | 1 day 2 hours 11 minutes |
| P6-W10-4Q-05 | 4% of 4 CPUs | 1.4 GB of 8.6 GB | 3/10 | 0/0 | 10.2.0.21, fe80::c16e... | 1 day 2 hours 11 minutes |
| P6-W10-4Q-06 | 4% of 4 CPUs | 1.4 GB of 8.6 GB | 1/4 | 0/0 | 10.2.0.23, fe80::3906... | 1 day 2 hours 11 minutes |
| P6-W10-4Q-07 | 4% of 4 CPUs | 1.4 GB of 8.6 GB | 2/6 | 0/0 | 10.2.0.60, fe80::90c7... | 1 day 2 hours 11 minutes |
| P6-W10-4Q |  | GB of 8.6 GB | 2/6 | 0/0 | 10.2.0.57, fe80::2422... | 1 day 2 hours 11 minutes |
| xenserver-24<br>Default install of |  | GB of 767 GB | - | 1/1 | 10.10.63.124, 10.10... | 9 hours 43 minutes |
| P6-W10-1Q-01 |  | - | - | - | 10.2.0.31, fe80::154a... | - |
| P6-W10-1Q-02 |  | - | - | - | 10.2.0.32, fe80::854:... | - |
| P6-W10-1Q-03 |  |  |  |  | 10.2.0.27, fe80::f5aa:... | - |
| P6-W10-1Q-04 |  |  |  |  | 10.2.0.29, fe80::d9c0... | - |
| P6-W10-1Q-05 |  |  |  |  | 10.2.0.34, fe80::8c22... | - |
| P6-W10-1Q-06 | - | - | - | - | 10.2.0.33, fe80::a16b... | - |
| P6-W10-1Q-07 | - | - | - | - | 10.2.0.31, fe80::70e5... | - |
| P6-W10-1Q-08 | - | - | - | - | 10.2.0.25, fe80::9922... | - |

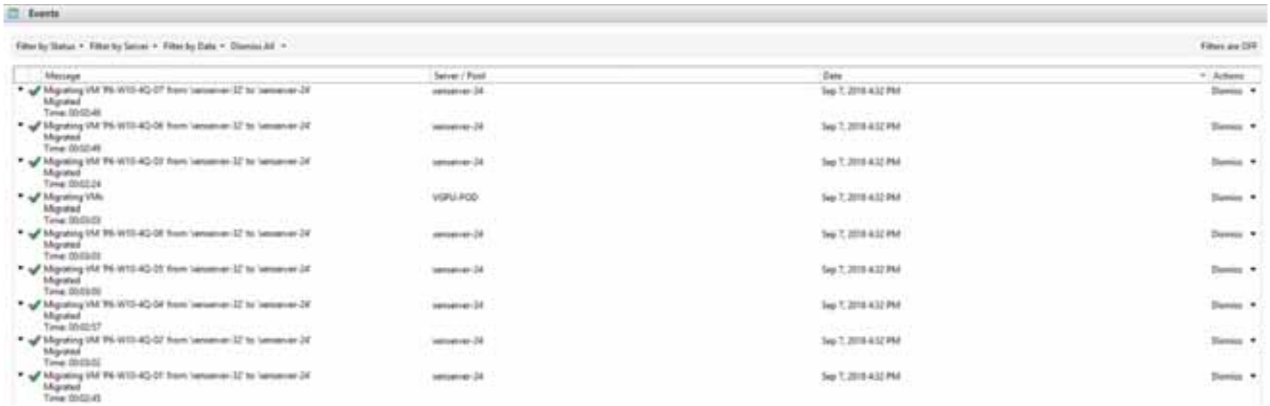
2. Selecting the server will start the virtual machine migration, as shown in Figure 112 and Figure 113.

**Figure 112.** Citrix XenCenter Events window: vGPU-enabled virtual machine XenMotion operations



| Message | Server / Pool | Date | Actions |
|---|---|---|---|
| Migrating VM 'P6-W10-4Q-07' from 'xenserver-32' to 'xenserver-24' | xenserver-24 | Sep 7, 2018 4:32 PM | Cancel |
| Migrating VM 'P6-W10-4Q-08' from 'xenserver-32' to 'xenserver-24' | xenserver-24 | Sep 7, 2018 4:32 PM | Cancel |
| Migrating VM 'P6-W10-4Q-03' from 'xenserver-32' to 'xenserver-24' | xenserver-24 | Sep 7, 2018 4:32 PM | Cancel |
| Migrating VMs | VGPU-POD | Sep 7, 2018 4:32 PM | Cancel |
| Migrating VM 'P6-W10-4Q-06' from 'xenserver-32' to 'xenserver-24' | xenserver-24 | Sep 7, 2018 4:32 PM | Cancel |
| Migrating VM 'P6-W10-4Q-05' from 'xenserver-32' to 'xenserver-24' | xenserver-24 | Sep 7, 2018 4:32 PM | Cancel |
| Migrating VM 'P6-W10-4Q-04' from 'xenserver-32' to 'xenserver-24' | xenserver-24 | Sep 7, 2018 4:32 PM | Cancel |
| Migrating VM 'P6-W10-4Q-02' from 'xenserver-32' to 'xenserver-24' | xenserver-24 | Sep 7, 2018 4:32 PM | Cancel |
| Migrating VM 'P6-W10-4Q-01' from 'xenserver-32' to 'xenserver-24' | xenserver-24 | Sep 7, 2018 4:32 PM | Cancel |

**Figure 113.** Citrix XenCenter Events window: vGPU-enabled virtual machine XenMotion operation completion



Events

Filter by Status ▾ Filter by Server ▾ Filter by Date ▾ Dismiss All ▾ Filters are OFF

| Message | Server / Pool | Date | Actions |
|---|---|---|---|
| ✓ Migrating VM 'P6-W10-4Q-07' from 'xenserver-32' to 'xenserver-24'<br>Migrated<br>Time: 00:02:48 | xenserver-24 | Sep 7, 2018 4:32 PM | Dismiss ▾ |
| ✓ Migrating VM 'P6-W10-4Q-06' from 'xenserver-32' to 'xenserver-24'<br>Migrated<br>Time: 00:02:49 | xenserver-24 | Sep 7, 2018 4:32 PM | Dismiss ▾ |
| ✓ Migrating VM 'P6-W10-4Q-03' from 'xenserver-32' to 'xenserver-24'<br>Migrated<br>Time: 00:02:24 | xenserver-24 | Sep 7, 2018 4:32 PM | Dismiss ▾ |
| ✓ Migrating VMs<br>Migrated<br>Time: 00:03:03 | VGPU-POD | Sep 7, 2018 4:32 PM | Dismiss ▾ |
| ✓ Migrating VM 'P6-W10-4Q-08' from 'xenserver-32' to 'xenserver-24'<br>Migrated<br>Time: 00:03:03 | xenserver-24 | Sep 7, 2018 4:32 PM | Dismiss ▾ |
| ✓ Migrating VM 'P6-W10-4Q-05' from 'xenserver-32' to 'xenserver-24'<br>Migrated<br>Time: 00:03:00 | xenserver-24 | Sep 7, 2018 4:32 PM | Dismiss ▾ |
| ✓ Migrating VM 'P6-W10-4Q-04' from 'xenserver-32' to 'xenserver-24'<br>Migrated<br>Time: 00:02:57 | xenserver-24 | Sep 7, 2018 4:32 PM | Dismiss ▾ |
| ✓ Migrating VM 'P6-W10-4Q-02' from 'xenserver-32' to 'xenserver-24'<br>Migrated<br>Time: 00:03:02 | xenserver-24 | Sep 7, 2018 4:32 PM | Dismiss ▾ |
| ✓ Migrating VM 'P6-W10-4Q-01' from 'xenserver-32' to 'xenserver-24'<br>Migrated<br>Time: 00:02:45 | xenserver-24 | Sep 7, 2018 4:32 PM | Dismiss ▾ |

## 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 graphical application. Doing so can cause instability for the duration of a user's session.

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

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

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

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

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

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

### Use the OpenGL Software Accelerator

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

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

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

Try the OpenGL Software Accelerator in the following cases:

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

## Conclusion

The combination of FlexPod Datacenter with Cisco UCS Manager, Cisco UCS C240 M5 Rack Servers and B200 M5 Blade Servers, NetApp AFF A300, and NVIDIA Tesla cards running NVIDIA GRID 6.2 on Citrix XenServer 7.5 and Citrix XenDesktop 7.13 provides a high-performance platform for virtualizing graphics-intensive applications.

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

## For more information

- Cisco UCS C-Series Rack Servers and B-Series Blade Servers:
  - http://www.cisco.com/en/US/products/ps10265/
- Cisco Desktop Virtualization Design Navigator
  - http://cisco.com/go/vdi-cdv
- NVIDIA:

    - http://www.nvidia.com/object/grid-technology.html
- Citrix XenApp and XenDesktop:
    - https://docs.citrix.com/en-us/xenapp-and-xendesktop/7-15-ltsr.html
    - https://docs.citrix.com/en-us/xenserver/7-5.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
- SPECviewperf 13:
    - https://www.spec.org/gwpg/gpc.static/vp13info.html
- NetApp A300 and ONTAP:
    - https://www.netapp.com/us/products/storage-systems/all-flash-array/aff-a-series.aspx#technical-specifications
    - https://www.netapp.com/us/products/data-management-software/ontap.aspx

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 C11-741599-00 11/18