

Integrate FlashStack for 3D Graphics Visualization with Citrix and NVIDIA on Cisco UCS



Last Updated: July 23, 2019

Contents

What you will learn	4
NVIDIA vGPU profiles	4
Cisco Unified Computing System	5
Cisco UCS Manager	7
Cisco UCS 6332 Fabric Interconnect	7
Cisco UCS C-Series Rack Servers	8
Cisco UCS C240 M5 Rack Server	8
Cisco UCS VIC 1387	10
Cisco UCS B200 M5 Blade Server	11
Cisco UCS VIC 1340	12
Cisco Nexus 93180YC-FX Switch	12
Cisco MDS 9132T 32-Gbps 32-Port Fibre Channel Switch	13
Purity for FlashArray (Purity//FA 5)	15
Evergreen™ Storage	16
NVIDIA GRID cards	17
NVIDIA GRID	17
NVIDIA GRID 7.2 GPU	18
NVIDIA GRID 7.2 license requirements	18
VMware vSphere 6.7	18
VMware vSphere Client	18
VMware ESXi Hypervisor 6.7	19
Graphics Acceleration in Citrix XenDesktop and XenApp	20
GPU Acceleration for Microsoft Windows Desktops	20
GPU Acceleration for Microsoft Windows Server	22
GPU Sharing for Citrix XenApp RDS Workloads	22
Citrix HDX 3D Pro Requirements	23
Solution configuration	24
Configure Cisco UCS	26
Create BIOS policy	26
Create graphics card policy	27
Install NVIDIA Tesla GPU card in Cisco UCS B200 M5	27
Installing an NVIDIA GPU card in the front of the server	28
Installing an NVIDIA GPU card in the rear of the server	31
Install NVIDIA Tesla GPU card in Cisco UCS C240 M5	33
Installing an NVIDIA Tesla T4	33
Installing a double-wide GPU card: NVIDIA Tesla P40	34
Configure the GPU card	35
Install the NVIDIA GRID vGPU Manager for VMware	37
Disable ECC memory	39
Install and configure the NVIDIA GRID license server	40
Install the GRID 7.2 license server	40
Configure the NVIDIA GRID 7.2 license server	45
NVIDIA Tesla P6, P40, and T4 profile specifications	47

- Create virtual desktops with vGPU support.....48
 - Create the Citrix XenDesktop base image48
 - Install the NVIDIA vGPU software driver51
 - Verify that the virtual machine is ready to support vGPU.....53
 - Configure the virtual machine for the NVIDIA GRID vGPU license54
- Deploy virtual machines with Citrix Machine Creation Services55
- Verify vGPU deployment63
 - Verify that the NVIDIA driver is running on the desktop.....63
 - Verify NVIDIA license acquisition by desktops64
- Create Citrix XenDesktop policies.....65
- SPECviewperf 13 benchmark results.....66**
 - SPECviewperf 13 results67
 - Host CPU utilization68
 - Host GPU utilization69
- Live vGPU-enabled virtual machine with VMware vMotion70
- Additional configurations.....74**
 - Install and upgrade NVIDIA drivers.....74
 - Use Citrix HDX Monitor.....74
 - Optimize the Citrix HDX 3D Pro user experience.....74
 - Use GPU acceleration for Microsoft Windows Server DirectX, Direct3D, and WPF rendering.....74
 - Use the OpenGL Software Accelerator74
- Conclusion75**
- For more information.....75**

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) environments, Cisco offers support for the NVIDIA GRID P6, P40, and T4 cards in the Cisco Unified Computing System™ (Cisco UCS) portfolio of PCI Express (PCIe) or mezzanine form-factor cards for the B-Series Blade Servers and C-Series Rack Servers.

With the addition of the new graphics processing capabilities, the engineering, design, imaging, and marketing departments of organizations can now experience the benefits that desktop virtualization brings to the applications they use. These new graphics capabilities help enable organizations to centralize their graphics workloads and data in the data center.

Important elements of the solution discussed in this document are the FlashStack infrastructure, Citrix support for NVIDIA virtual graphics processing unit (vGPU) technology, and the capability of VMware vSphere 6.7 Update 1 to move vGPU-enabled virtual machines using VMware vMotion to reduce user downtime.

The purpose of this document is to help our partners and customers integrate NVIDIA GRID 7.2 graphics cards, Cisco UCS B200 M5 Blade Servers, and Cisco UCS C240 M5 Rack Servers on VMware vSphere 6.7 Update 1 and Citrix XenDesktop 7.15 in vGPU mode.

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

This document presents the steps for integrating FlashStack architecture on the Cisco UCS platform using NVIDIA GRID P6, P40, and T4 cards with Citrix products so that the servers, hypervisor, and virtual desktops are ready for the installation of graphics applications.

NVIDIA vGPU profiles

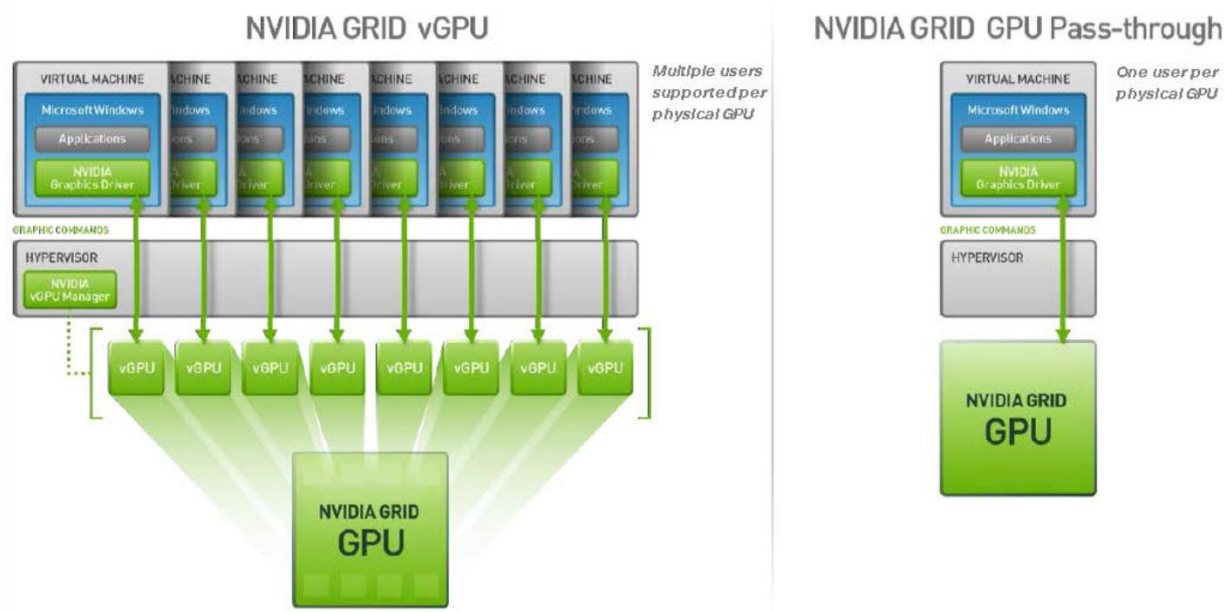
In any given enterprise, the needs of individual users vary widely. One of the main benefits of the GRID vGPU is the flexibility to use various vGPU profiles designed to serve the needs of different classes of end users.

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

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

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

Figure 1. NVIDIA GRID vGPU GPU system architecture



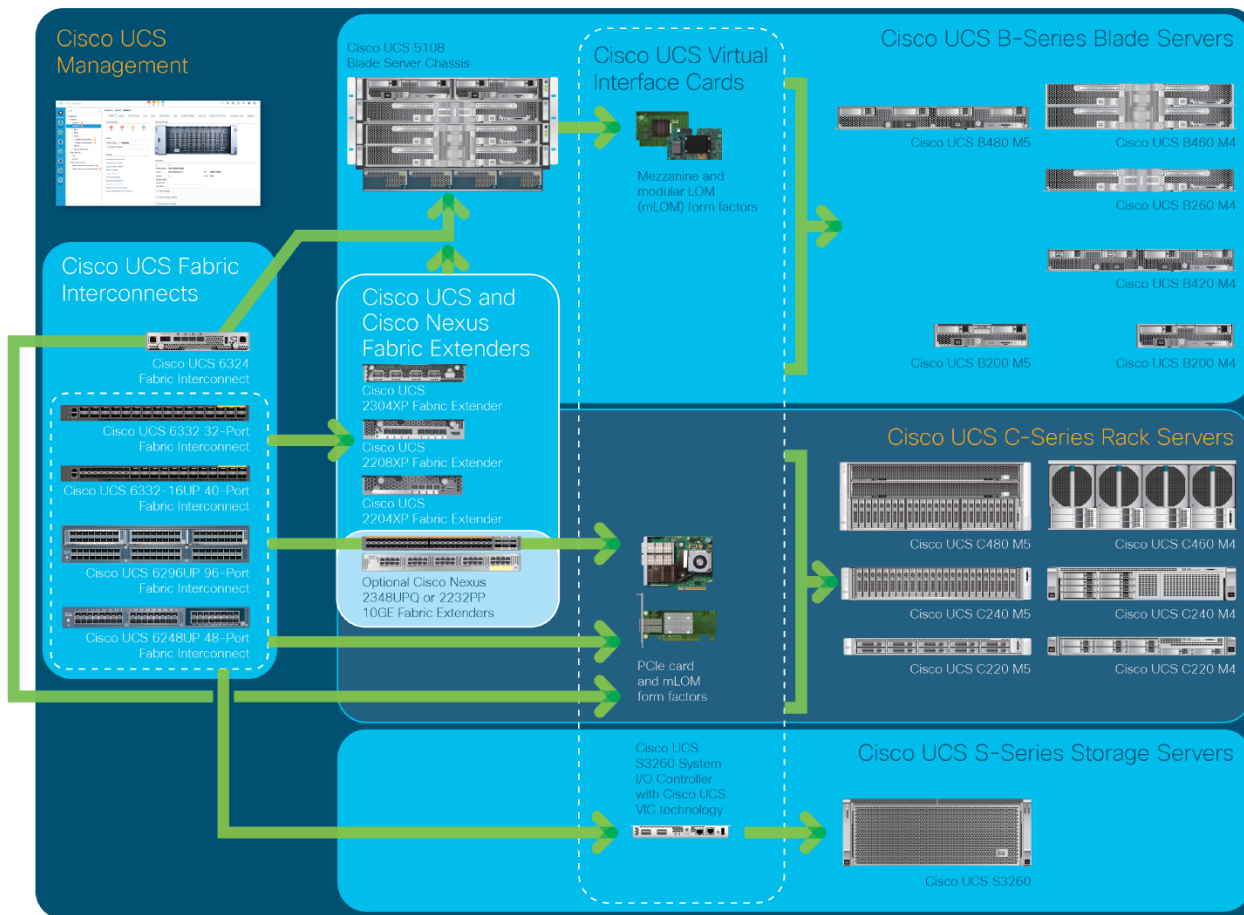
Cisco Unified Computing System

These are the main components of Cisco UCS:

- **Computing:** The system is based on an entirely new class of computing system that incorporates blade servers based on Intel® Xeon® processor E5-2600/4600 v3 and E7-2800 v3 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 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 a Cisco UCS data center deployment.

Figure 2. Cisco data center overview



Cisco UCS is designed to deliver:

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

Cisco UCS Manager provides unified, embedded management of all software and hardware components of Cisco UCS 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. Tight integration of 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 information, see <https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-b-series-blade-servers/6332-specsheet.pdf>.

The features and capabilities of the fabric interconnects include the following:

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

Figure 3. Cisco UCS 6332 Fabric Interconnect (front view)



Figure 4. Cisco UCS 6332 Fabric Interconnect (rear view)



Cisco UCS C-Series Rack Servers

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

Designed to operate both in standalone environments and as part of 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 and 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 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 processor family, 24 DIMM slots for 2666-MHz DDR4 DIMMs and up to 128-GB capacity points, up to 6 PCIe 3.0 slots, and up to 26 internal SFF drives. The C240 M5 SFF server also includes one dedicated internal slot for a 12-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 two 10G BASE-T Intel x550 LAN-on-motherboard 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 up to 2666 MHz for improved performance and lower power consumption
- One or two Intel Xeon Scalable family CPUs
- Up to six PCIe 3.0 slots (four full-height, full-length for the GPU)
- Six hot-swappable fans for front-to-rear cooling
- Twenty-four SFF front-facing SAS/SATA hard-disk drives (HDDs) or SAS/SATA solid-state disks (SSDs)
- Optionally, up to two front-facing SFF Non-Volatile Memory Express (NVMe) PCIe SSDs (replacing SAS/SATA drives); these drives must be placed in front drive bays 1 and 2 only and are controlled from Riser 2, Option C
- Optionally, up to two SFF rear-facing SAS/SATA HDDs or SSDs or up to two rear-facing SFF NVMe PCIe SSDs; rear-facing SFF NVMe drives are connected from Riser 2, Option B or C
- Support for 12-Gbps SAS drives
- Dedicated mLOM slot on the motherboard, which can flexibly accommodate the following cards:
 - Cisco VICs
 - Quad-port Intel i350 1 Gigabit Ethernet RJ45 mLOM NIC
- Two 1 Gigabit Ethernet embedded LOM ports
- Support for up to two double-wide NVIDIA GPUs, providing a robust graphics experience to more virtual users

- Excellent reliability, availability, and serviceability (RAS) features with tool-free CPU insertion, easy-to-use latching lid, and hot-swappable and hot-pluggable components
- One slot for a micro Secure Digital (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).
 - Images can be pulled from a file share (Network File System [NFS] or Common Internet File System [CIFS]) and uploaded to the cards for future use.
- A mini-storage-module connector on the motherboard supports either:
 - An SD card module with two SD card slots (mixing different-capacity SD cards is not supported)
 - An M.2 module with two SATA M.2 SSD slots (mixing different-capacity M.2 modules is not supported)

Note: SD card modules and M.2 modules cannot be mixed. M.2 does not support RAID1 with VMware. Only Microsoft Windows and Linux are supported.

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

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

The C240 M5 can be deployed as a standalone server or as part of Cisco UCS. Cisco UCS unifies computing, networking, management, virtualization, and storage access into a single integrated architecture that enables end-to-end server visibility, management, and control in both bare-metal and virtualized environments. Within a Cisco UCS deployment, the C240 M5 takes advantage of Cisco's standards-based unified computing innovations, which significantly reduce customers' total cost of ownership (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 5. Cisco UCS C240 M5 Rack Server



Figure 6. 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 7) is a dual-port SFP+ 40-Gbps Ethernet and FCoE-capable PCIe mLOM adapter installed in Cisco UCS C-Series Rack Servers. The mLOM slot can be used to install a Cisco VIC without consuming a PCIe slot, which provides greater I/O expandability. It incorporates next-generation converged network adapter (CNA) technology from Cisco, providing investment protection for future feature releases. The card enables a policy-based, stateless, agile server infrastructure that can present more than 256 PCIe standards-compliant interfaces to the host that can be dynamically configured as either NICs or host bus adapters (HBAs). The personality of the card is determined dynamically at boot time using the service profile associated with the server. The number, type (NIC or HBA), identity (MAC address and World Wide Name [WWN]), failover policy, bandwidth, and quality-of-service (QoS) policies of the PCIe interfaces are all determined using the service profile.

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

Figure 7. Cisco UCS VIC 1387



Cisco UCS B200 M5 Blade Server

Delivering performance, versatility and density without compromise, the Cisco UCS B200 M5 Blade Server (Figure 8) addresses a broad set of workloads: IT and web infrastructure, distributed databases, and more. 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 Cisco UCS B200 M5 harnesses the power of the latest Intel Xeon Scalable family CPUs with up to 3072 GB of RAM (using 128-GB DIMMs), two SSDs or HDDs, and up to 80 Gbps of throughput.

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

The 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. Alternatively, the same rear mezzanine adapter slot can be configured with an NVIDIA P6 GPU.

The UCS B200 M5 has one front mezzanine slot. The UCS B200 M5 can be ordered with or without the front mezzanine card. The Front Mezzanine Card can accommodate 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 8. Cisco UCS B200 M5 Blade Server (front view)



Cisco UCS VIC 1340

The Cisco UCS VIC 1340 (Figure 9) is a 2-port 40-Gbps Ethernet or dual 4 x 10-Gbps Ethernet, 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 9. Cisco UCS VIC 1340



Cisco Nexus 93180YC-FX Switch

The Cisco Nexus[®] 93180YC-EX Switch (Figure 10) provides a flexible line-rate Layer 2 and Layer 3 feature set in a compact form factor. Designed with Cisco Cloud Scale technology, it supports highly scalable cloud architectures. With the option to operate in Cisco NX-OS or Cisco Application Centric Infrastructure (Cisco ACI™) mode, it can be deployed across enterprise, service provider, and Web 2.0 data centers.

The 93180YC-EX offers these features:

- Architectural flexibility
 - Leaf-node support for Cisco ACI architecture with flexible port configuration.
 - Seamless convergence thanks to 48 downlink ports that can work as 1/10/25-Gbps Ethernet or FCoE ports or as 8/16/32-Gbps Fibre Channel ports.
 - Easy migration with 6 uplink ports that can be configured as 40/100-Gbps Ethernet or FCoE ports.

- Rich features
 - Automated policy-based systems management with Cisco ACI.
 - Open APIs enable third-party integration with our partners.
 - Better management of speed mismatch between access and uplink ports with 40 MB of shared buffer space.
 - Support for Fibre Channel interfaces for back-end storage connectivity.
- Top-notch security
 - Whitelist model, policy enforcement and application security with Cisco ACI microsegmentation
 - Wire-rate MACsec encryption on all ports
- Real-time visibility and telemetry
 - Built-in Cisco Tetration sensors for rich traffic-flow telemetry and line-rate data collection
 - Get actionable insights in less than 1 second
 - Get visibility into everything in your data center
- Highly available and efficient design
 - High-performance, nonblocking architecture.
 - Easily deployed into either a hot-aisle or a cold-aisle configuration.
 - Easily deployed into either a hot-aisle or a cold-aisle configuration.
- Simplified operations
 - Automate IT work flows and shorten app deployment from weeks to minutes
- Investment protection
 - Flexible migration options with support for 10-Gbps and 25-Gbps access connectivity and 40-Gbps and 100-Gbps uplinks.
 - Cisco's 40-Gbps [bidirectional transceiver](#) allows for reuse of an existing 10 Gigabit Ethernet multimode cabling plant for 40 Gigabit Ethernet.

Figure 10. Cisco Nexus 93180YC-EX Switch



Cisco MDS 9132T 32-Gbps 32-Port Fibre Channel Switch

The next-generation Cisco MDS 9132T 32-Gbps 32-Port Fibre Channel Switch (Figure 11) provides high-speed Fibre Channel connectivity from the server rack to the SAN core. It empowers small, midsize, and large enterprises that are rapidly deploying cloud-scale applications using extremely dense virtualized servers, providing the dual benefits of greater bandwidth and consolidation.

Small-scale SAN architectures can be built from the foundation using this low-cost, low-power, nonblocking, line-rate, low-latency, fixed, standalone SAN switch. The switch supports bidirectional airflow and connects both storage and host ports.

Medium-size to large-scale SAN architectures built with SAN core directors can extend 32-Gbps connectivity to the server rack using these switches either in switch mode or Network Port Virtualization (NPV) mode.

Additionally, investing in this switch for a lower-speed (4, 8, or 16 Gbps) server rack gives users the option to upgrade to 32-Gbps server connectivity in the future using a 32-Gbps HBA available today. The Cisco® MDS 9132T 32-Gbps 32-Port Fibre Channel switch

also provides exceptional flexibility through a unique port expansion module (Figure 11) that provides a robust cost-effective, field-swappable, port-upgrade option.

This switch also offers state-of-the-art SAN analytics and telemetry capabilities, which are built in to this next-generation hardware platform. This new technology couples a next-generation port application-specific integrated circuit (ASIC) with a fully dedicated Network Processing Unit designed to complete analytics calculations in real time. The telemetry data extracted from the inspection of the frame headers are calculated on board (within the switch) and, using an industry-leading open format, can be streamed to any analytics-visualization platform. This switch also includes a dedicated 10/100/1000BASE-T telemetry port to maximize data delivery to any telemetry receiver including Cisco Data Center Network Manager.

Figure 11. Cisco MDS 9132T 32-Gbps Fibre Channel Switch



Figure 12. Cisco MDS 9132T 32-Gbps 16-port Fibre Channel Port Expansion Module



The 9132T switch offers these main features:

- High performance: The MDS 9132T architecture, with chip-integrated nonblocking arbitration, provides consistent 32-Gbps low-latency performance across all traffic conditions for every Fibre Channel port on the switch.
- Capital Expenditure (CapEx) savings: The 32-Gbps ports allow users to deploy them on existing 16- or 8-Gbps transceivers, reducing initial CapEx with an option to upgrade to 32-Gbps transceivers and adapters in the future.
- High availability: The MDS 9132T switches continue to provide the same outstanding availability and reliability as the previous-generation Cisco MDS 9000 Family switches by providing optional redundancy on all major components such as the power supply and fan. Dual power supplies also facilitate redundant power grids.
- Pay-as-you-grow: The MDS 9132T Fibre Channel switch provides an option to deploy as few as eight 32-Gbps Fibre Channel ports in the entry-level variant, which can grow by 8 ports to 16 ports, and thereafter, with a port expansion module with sixteen 32-Gbps ports, to up to 32 ports. This approach results in lower initial investment and power consumption for entry-level configurations of up to 16 ports compared to a fully loaded switch. Upgrading through an expansion module also reduces the overhead of managing multiple instances of port activation licenses on the switch. This unique combination of port upgrade options allows four possible configurations: of 8 ports, 16 ports, 24 ports, and 32 ports.
- Next-generation Application-Specific Integrated Circuit (ASIC): The MDS 9132T Fibre Channel switch is powered by the same high-performance 32-Gbps Cisco ASIC with an integrated network processor that powers the Cisco MDS 9700 48-Port 32-Gbps Fibre Channel Switching Module. Among all the advanced features that this ASIC enables, one of the most notable is inspection of Fibre Channel and Small Computer System Interface (SCSI) headers at wire speed on every flow in the smallest form-factor Fibre Channel switch without the need for any external taps or appliances. The recorded flows can be analyzed on the switch and also exported using a dedicated 10/100/1000BASE-T port for telemetry and analytics purposes.

- Intelligent network services: Slow-drain detection and isolation, VSAN technology, Access Control Lists (ACLs) for hardware-based intelligent frame processing, smart zoning and fabricwide Quality of Service (QoS) enable migration from SAN islands to enterprise wide storage networks. Traffic encryption is optionally available to meet stringent security requirements.
- Sophisticated diagnostics: The MDS 9132T provides intelligent diagnostics tools such as Inter-Switch Link (ISL) diagnostics, read diagnostic parameters, protocol decoding, network analysis tools, and integrated Cisco Call Home capability for greater reliability, faster problem resolution, and reduced service costs.
- Virtual machine awareness: The MDS 9132T provides visibility into all virtual machines logged into the fabric. This feature is available through HBAs capable of priority tagging the Virtual Machine Identifier (VMID) on every FC frame. Virtual machine awareness can be extended to intelligent fabric services such as analytics to visualize the performance of every flow originating from each virtual machine in the fabric.
- Programmable fabric: The MDS 9132T provides powerful Representational State Transfer (REST) and Cisco NX-API capabilities to enable flexible and rapid programming of utilities for the SAN as well as polling point-in-time telemetry data from any external tool.
- Single-pane management: The MDS 9132T can be provisioned, managed, monitored, and troubleshot using Cisco Data Center Network Manager (DCNM), which currently manages the entire suite of Cisco data center products.
- Self-contained advanced anticounterfeiting technology: The MDS 9132T uses on-board hardware that protects the entire system from malicious attacks by securing access to critical components such as the bootloader, system image loader and Joint Test Action Group (JTAG) interface.

Purity for FlashArray (Purity//FA 5)

At the heart of every FlashArray is Purity Operating Environment software. Purity//FA5 implements advanced data reduction, storage management, and flash management features, enabling organizations to enjoy Tier 1 data services for all workloads, proven 99.9999% availability over two years (inclusive of maintenance and generational upgrades), completely non-disruptive operations, 2X better data reduction versus alternative all-flash solutions, and – with FlashArray//X – the power and efficiency of DirectFlash™. Moreover, Purity includes enterprise-grade data security, comprehensive data protection options, and complete business continuity through ActiveCluster multi-site stretch cluster. All these features are included with every array.

FlashArray//X Specifications

	CAPACITY	PHYSICAL	//X CONNECTIVITY
//X10	Up to 55 TB / 53.5 TiB effective capacity** Up to 20 TB / 18.6 TiB raw capacity	3U 490 – 600 Watts (nominal – peak) 95 lbs (43.1 kg) fully loaded 5.12" x 18.94" x 29.72" chassis	Onboard Ports (per controller) <ul style="list-style-type: none"> • 2 x 1/10/25 Gb Ethernet • 2 x 1/10/25 Gb Ethernet Replication • 2 x 1Gb Management Ports
//X20	Up to 275 TB / 251.8 TiB effective capacity** Up to 87 TB / 80.3 TiB raw capacity**	3U 620 – 688 Watts (nominal – peak) 95 lbs (43.1 kg) fully loaded 5.12" x 18.94" x 29.72" chassis	Host I/O Cards (3 slots/controller) <ul style="list-style-type: none"> • 2-port 10GBase-T Ethernet • 2-port 1/10/25 Gb Ethernet • 2-port 40 Gb Ethernet • 2 Port 50Gb Ethernet (NVMe-oF Ready)*** • 2-port 16/32 Gb Fibre Channel (NVMe-oF Ready) • 4-port 16/32 Gb Fibre Channel (NVMe-oF Ready)
//X50	Up to 650 TB / 602.8 TiB effective capacity** Up to 183 TB / 171 TiB raw capacity†	3U 620 – 760 Watts (nominal – peak) 95 lbs (43.1 kg) fully loaded 5.12" x 18.94" x 29.72" chassis	
//X70	Up to 1.3 PB / 1238.5 TiB effective capacity** Up to 366 TB / 320.1 TiB raw capacity†	3U 915 – 1345 Watts (nominal – peak) 97 lbs (44.0 kg) fully loaded 5.12" x 18.94" x 29.72" chassis	
//X90	Up to 3 PB / 3003.1 TiB effective capacity** Up to 878 TB / 768.3 TiB raw capacity†	3U – 6U 1100 – 1570 Watts (nominal – peak) 97 lbs (44 kg) fully loaded 5.12" x 18.94" x 29.72" chassis	
DIRECT FLASH SHELF	Up to 1.9 PB effective capacity** Up to 512 TB / 448.2 TiB raw capacity	3U 460 - 500 Watts (nominal – peak) 87.7 lbs (39.8kg) fully loaded 5.12" x 18.94" x 29.72" chassis	

* Stated //X specifications are applicable to //X R2 versions.

** Effective capacity assumes HA, RAID, and metadata overhead, GB-to-GiB conversion, and includes the benefit of data reduction with always-on inline deduplication, compression, and pattern removal. Average data reduction is calculated at 5-to-1 and does not include thin provisioning or snapshots.

*** FlashArray //X currently supports NVMe-oF through RoCEv2 with a roadmap for FC-NVMe and TCP-NVMe.

Evergreen™ Storage

Customers can deploy storage once and enjoy a subscription to continuous innovation through Pure's Evergreen Storage ownership model: expand and improve performance, capacity, density, and/or features for 10 years or more – all without downtime, performance impact, or data migrations. Pure has disrupted the industry's 3-5-year rip-and-replace cycle by engineering compatibility for future technologies right into its products, notably with the NVMe-Ready Guarantee for //M and online upgrade from any //M to //X.

7 YEARS OF NON-DISRUPTIVE **EVERGREEN** IMPROVEMENTS

176x Capacity Increase **146x** Density Increase **13x** IOPs **5x** Throughput **7** HW Generations **\$0** Wasted Investment



NVIDIA GRID cards

For desktop virtualization applications, the NVIDIA Tesla P6, P4, and P40 cards are an optimal choice for high-performance graphics (Table 2).

Table 2. Technical specifications for NVIDIA GRID cards

	P6	T4	P40
Number of GPUs	Single Pascal	Single Turing	Single Pascal
Cores	2048 CUDA cores	2560 CUDA cores 320 Turing Tensor cores 40 RT cores	3840 CUDA cores
Memory size	16-GB GDDR5	16-GB GDDR6	24-GB GDDR5
Maximum number of vGPU instances	16 (1-GB profile)	16 (1-GB profile)	24 (1-GB profile)
Power	90 watts (W)	70W	250W
Form factor	MXM (blade servers), x16 lanes	PCIe 3.0 single slot (low profile)	PCIe 3.0 dual slot (rack servers), x16 lanes
Cooling solution	Bare board	Passive	Passive
H.264 1080p30 streams	24	38	24
Maximum number of users per board	16 (1-GB profile)	16 (1-GB profile)	24 (with 1-GB profile)
Virtualization use case	Blade optimized	Performance optimized	Performance optimized

NVIDIA GRID

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

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

NVIDIA GRID 7.2 GPU

The NVIDIA GRID solution runs on Tesla GPUs based on NVIDIA Volta, NVIDIA Pascal, NVIDIA Maxwell, and NVIDIA Turing architectures. These GPUs come in two server form factors: the NVIDIA Tesla [P6](#) for blade servers and converged infrastructure, and the NVIDIA Tesla [T4](#) and [P40](#) for rack servers.

NVIDIA GRID 7.2 license requirements

GRID 7.2 requires concurrent user licenses and an on-premises NVIDIA license server to manage the licenses. When the guest OS boots up, 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 7.2 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:

- Quadro Virtual Data Center Workstation (vDWS)
- Virtual PC
- Virtual App

For complete details about GRID 7.2 license requirements, see the [NVIDIA documentation](#).

VMware vSphere 6.7

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

VMware vSphere 6.7 introduces many enhancements to vSphere Hypervisor, VMware virtual machines, vCenter Server, virtual storage, and virtual networking, further extending the core capabilities of the vSphere platform.

The vSphere 6.7 release offers an especially robust new feature set for vSphere. The vCenter Server Appliance is particularly prominent, with several new features. For instance, the installer has a new modern look and feel. In addition, the installer is now supported on both the Linux and Mac OS platforms as well as Microsoft Windows. The vCenter Server Appliance also now includes exclusive features such as the following:

- Migration capabilities
- Improved Appliance Management
- VMware Update Manager
- Native High Availability
- Built-in Backup / Restore

VMware vSphere Client

VMware vSphere 6.7 includes a fully supported version of the HTML5-based vSphere Client that runs alongside the vSphere Web Client. The vSphere Client is built in to vCenter Server 6.7 (both Windows and Appliance versions) and is enabled by default. Although the HTML-5 based vSphere Client does not have full feature parity, the team has prioritized many of the day-to-day tasks of administrators and continues to seek feedback on items that will enable customers to use it full time. The vSphere Web Client continues to be accessible through http://<vcenter_fqdn>/vsphere-client, and the vSphere Client is accessible through http://<vcenter_fqdn>/ui. VMware is periodically updating the vSphere Client outside the normal vCenter Server release cycle. To help ensure that customers can easily stay up-to-date, the vSphere Client can be updated without any effects on the rest of vCenter Server.

Here are some of the benefits of the new vSphere Client:

- Clean, consistent user interface built on VMware's new Clarity user interface standards (being adopted across the VMware portfolio)
- Built on HTML5 so it is truly a cross-browser and cross-platform application
- No browser plug-ins to install or manage
- Integrated into vCenter Server 6.7 and fully supported
- Full support for Enhanced Linked Mode

Users of the Fling have been extremely positive about the product's performance.

VMware ESXi Hypervisor 6.7

VMware vSphere 6.7 introduces the following new features in the hypervisor:

- **Scalability improvements**
 - ESXi 6.7 dramatically increases the scalability of the platform. With vSphere Hypervisor 6.0, clusters can scale to as many as 64 hosts, up from 32 in previous releases. With 64 hosts in a cluster, vSphere 6.0 can support 8000 virtual machines in a single cluster. This capability enables greater consolidation ratios, more efficient use of VMware vSphere Distributed Resource Scheduler (DRS), and fewer clusters that must be separately managed. Each vSphere Hypervisor 6.7 instance can support up to 480 logical CPUs, 12 terabytes (TB) of RAM, and 1024 virtual machines. By using the newest hardware advances, ESXi 6.7 enables the virtualization of applications that previously had been thought to be nonvirtualizable.
- **ESXi 6.7 security enhancements**
 - Account management: ESXi 6.7 enables management of local accounts on the ESXi server using new ESXi CLI commands. The capability to add, list, remove, and modify accounts across all hosts in a cluster can be centrally managed using a vCenter Server system. Previously, the account and permission management functions for ESXi hosts were available only for direct host connections. The setup, removal, and listing of local permissions on ESXi servers can also be centrally managed.
 - Account lockout: ESXi Host Advanced System Settings has two new options for the management of failed local account login attempts and account lockout duration. These parameters affect Secure Shell (SSH) and vSphere Web Services connections, but not ESXi direct console user interface (DCUI) or console shell access.
 - Password complexity rules: In previous versions of ESXi, password complexity changes had to be made by manually editing the `/etc/pam.d/passwd` file on each ESXi host. In vSphere 6.0, an entry in Host Advanced System Settings enables changes to be centrally managed for all hosts in a cluster.
 - Improved auditability of ESXi administrator actions: Prior to vSphere 6.0, actions at the vCenter Server level by a named user appeared in ESXi logs with the vpxuser user name: for example, `[user=vpxuser]`. In vSphere 6.7, all actions at the vCenter Server level for an ESXi server appear in the ESXi logs with the vCenter Server user name: for example, `[user=vpxuser:`

DOMAIN\User]. This approach provides a better audit trail for actions run on a vCenter Server instance that conducted corresponding tasks on the ESXi hosts.

- Flexible lockdown modes: Prior to vSphere 6.7, only one lockdown mode was available. Feedback from customers indicated that this lockdown mode was inflexible in some use cases. With vSphere 6.7, two lockdown modes are available:
 - In normal lockdown mode, DCUI access is not stopped, and users on the DCUI access list can access the DCUI.
 - In strict lockdown mode, the DCUI is stopped.
- Exception users: vSphere 6.0 offers a new function called exception users. Exception users are local accounts or Microsoft Active Directory accounts with permissions defined locally on the host to which these users have host access. These exception users are not recommended for general user accounts, but they are recommended for use by third-party applications—for service accounts, for example—that need host access when either normal or strict lockdown mode is enabled. Permissions on these accounts should be set to the bare minimum required for the application to perform its task and with an account that needs only read-only permissions on the ESXi host.
- Smart card authentication to DCUI: This function is for U.S. federal customers only. It enables DCUI login access using a Common Access Card (CAC) and Personal Identity Verification (PIV). The ESXi host must be part of an Active Directory domain.

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 the following:

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

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

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

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

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

GPU Acceleration for Microsoft Windows Desktops

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

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

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

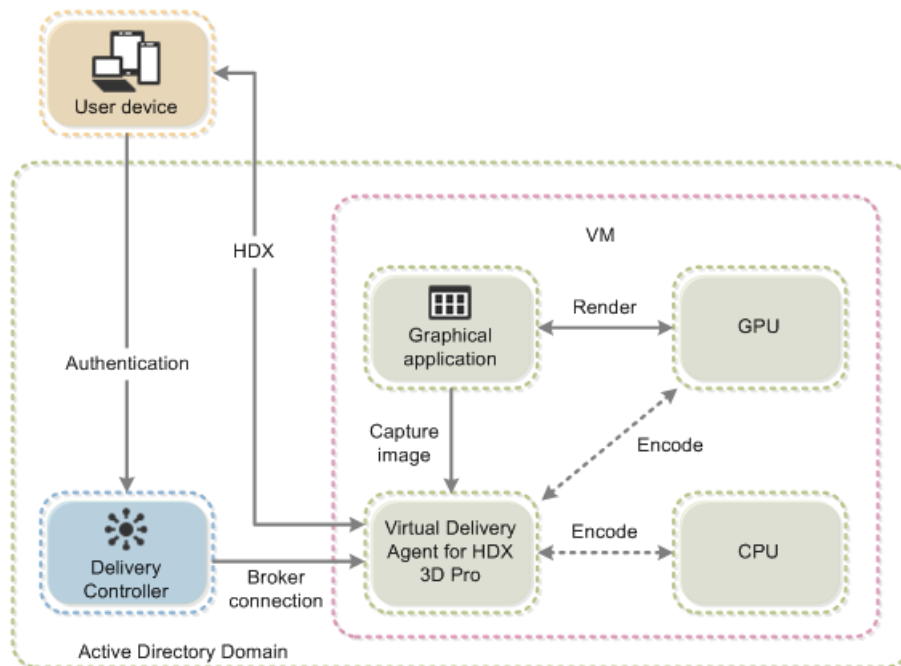
HDX 3D Pro offers the following features:

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

As shown in Figure 13:

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

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 VMware vSphere 6. When Citrix HDX 3D Pro is used with GPU pass-through, each GPU in the server supports one multiuser virtual machine.

Scalability using RDS GPU sharing depends on several factors:

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

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

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

Citrix HDX 3D Pro Requirements

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

- GPU pass-through is available with Citrix XenServer; VMware vSphere and ESX, where it is referred to as virtual direct graphics acceleration, or vDGA; and Microsoft Hyper-V in Microsoft Windows Server 2016, where it is referred to as discrete device assignment (DDA).
- vGPU is available with Citrix XenServer and VMware vSphere; see <https://www.citrix.com/products/xenapp-xendesktop/hdx-3d-pro.html>.
- Citrix recommends that the host computer have at least 4 GB of RAM and four virtual CPUs with a clock speed of 2.3 GHz or higher.

The requirements for the GPU are as follows:

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

The requirements for the user device are as follows:

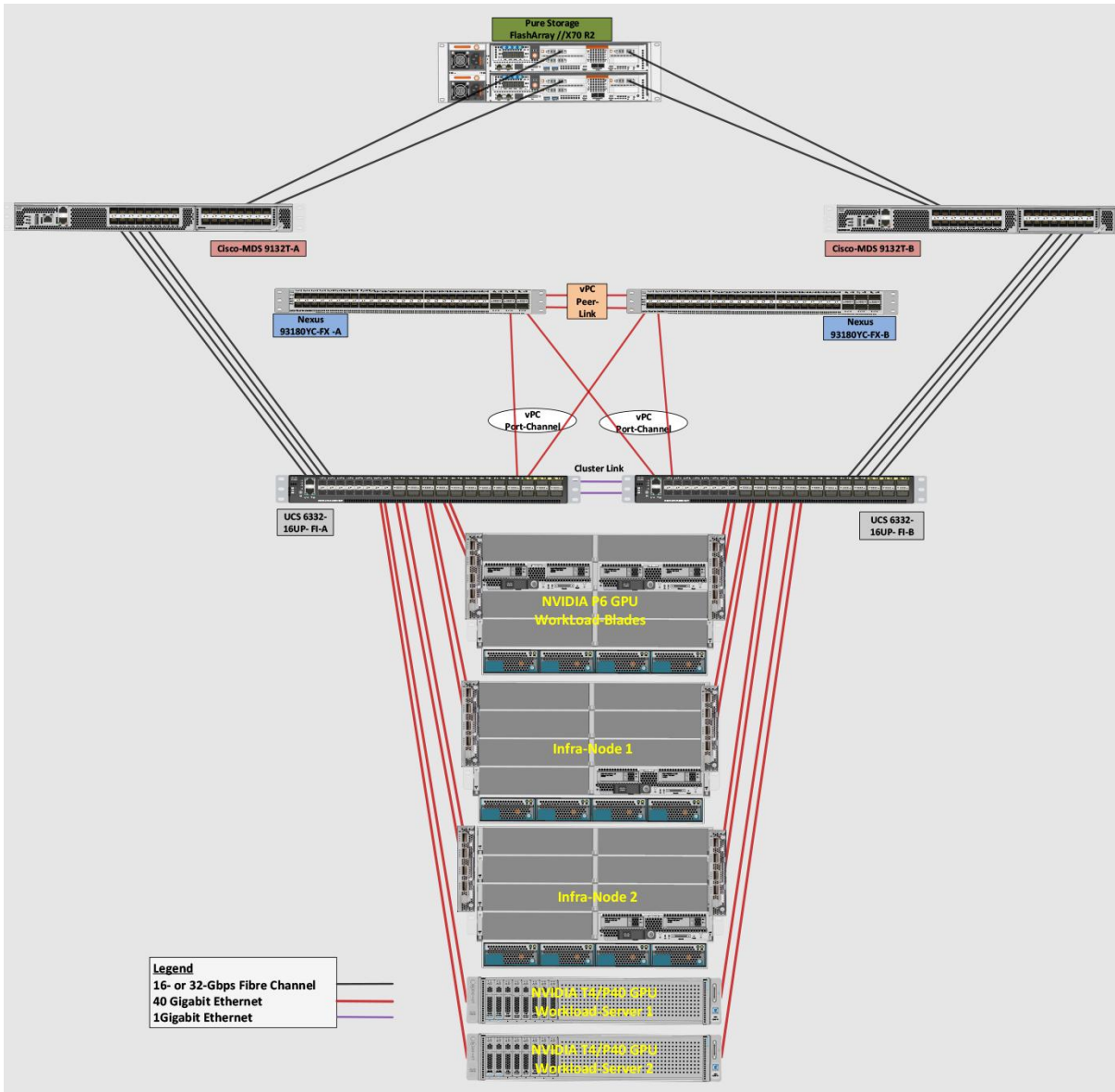
- HDX 3D Pro supports all monitor resolutions that are supported by the GPU on the host computer. However, for optimal performance with the minimum recommended user device and GPU specifications, Citrix recommends a maximum monitor resolution for user devices of 1920 x 1200 pixels for LAN connections, and 1280 x 1024 pixels for WAN connections.
- Citrix recommends that user devices have at least 1 GB of RAM and a CPU with a clock speed of 1.6 GHz or higher. Use of the default deep compression codec, which is required on low-bandwidth connections, requires a more powerful CPU unless the decoding is performed in hardware. For 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 13 provides an overview of the physical connectivity configuration of the FlashStack solution. The solution is described in a great detail in the Cisco Validated Design [FlashStack Data Center with Citrix XenDesktop 7.15 and VMware vSphere 6.7 U1 with Cisco UCS Manager 4.0 for 6000 Seats](#). This architecture was used to validate Tesla NVIDIA graphic cards using SPECviewperf 13 and Citrix XenDesktop HDX 3D Pro, reported in this document.

Figure 13. Cabling diagram for a FlashStack data center with Cisco UCS



The following hardware components were used in the solution:

- Cisco UCS B200 M5 Blade Servers with Intel Xeon Gold 6140 2.30-GHz 18-core processors, with 768 GB of 2666-MHz RAM for infrastructure.
- Cisco UCS B200 M5 Blade Servers with Intel Xeon Gold 6140 2.30-GHz 18-core processors, with 768 GB of 2666-MHz RAM and NVIDIA P6 GPUs for graphics accelerated VCC workloads
- Cisco UCS C240 M5 Rack Servers with Intel Xeon Gold 6154 3.00-GHz 18-core processors, with 768 GB of 2666-MHz RAM and NVIDIA T4 or P40 GPUs for graphics accelerated VCC workloads
- Four Cisco UCS 5108 Blade Server Chassis with two Cisco UCS 2304 Fabric Extender I/O modules
- Cisco UCS VIC 1387 mLOM (Cisco UCS C240 M5)
- Cisco UCS VIC 1340 mLOM (Cisco UCS B200 M5)
- Pure Storage FlashArray //x70 R2, used for all data
- Cisco Nexus 93180YC-FX Switches, used in NX-OS mode for Layer 2 communications
- Cisco MDS 9132T 32-Gbps or 16-Gbps Fibre Channel switches for Fibre Channel connectivity

The following software components were used in the solution:

- Cisco UCS Firmware Release 4.0(4b)
- VMware vSphere ESXi 6.7 Update 1 (U1)
- Citrix XenDesktop 7.15 Long-Term Service Release (LTSR) Cumulative Update 3 (CU3)
- Microsoft Windows 10 64-bit
- Microsoft Server 2016
- SPECviewperf 13 software and commercial license
- NVIDIA GRID 7.2 software and licenses:
 - NVIDIA-VMware_ESXi_6.7_Host_Driver-410.107-1OEM.670.0.0.8169922.x86_64.vib
 - 412.31_grid_win10_server2016_64bit_international.exe

Configure Cisco UCS

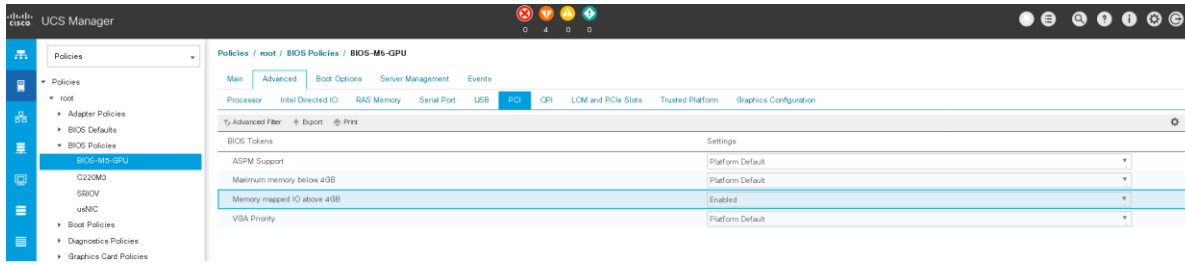
This section describes the Cisco UCS configuration.

Create BIOS policy

Create a new BIOS policy.

1. Right-click BIOS Policy.
2. On the Advanced tab for the new BIOS policy, click PCI. Select settings as shown in Figure 14:
 - Memory mapped IO above 4GB: Enabled

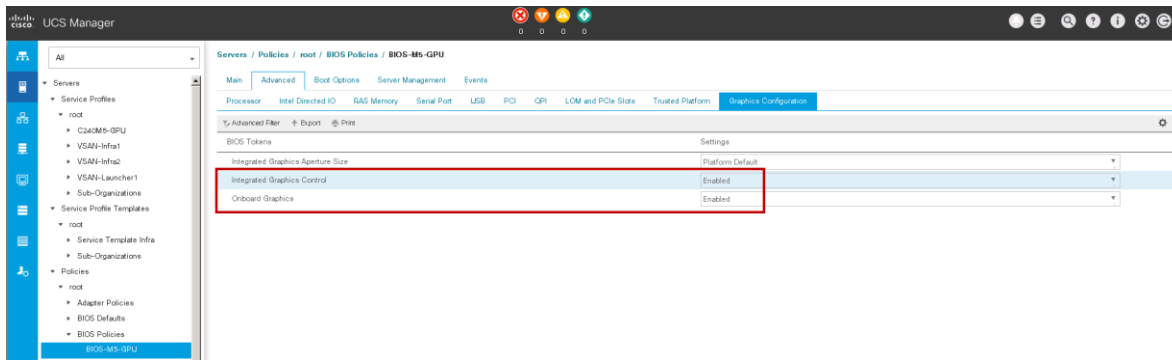
Figure 14. PCI setting for BIOS policy: Enable Memory mapped IO (MMIO) above 4 GB



3. Click Graphics Configuration and select BIOS policy settings as shown in Figure 15:

- Integrated Graphics Control: Enabled
- Onboard Graphics: Enabled

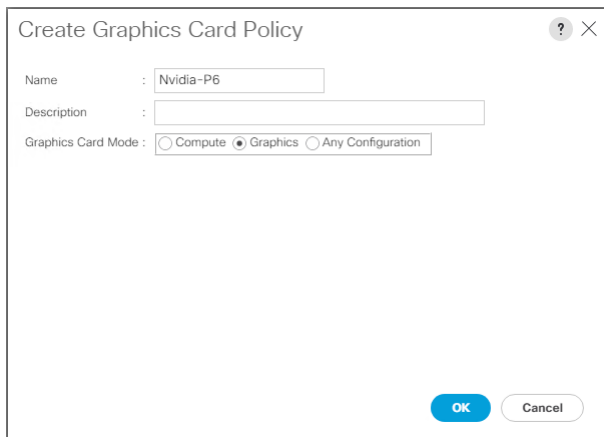
Figure 15. PCI BIOS policy configuration



Create graphics card policy

Create a new graphics card policy with the desired graphics card mode. For VDI deployment, graphics mode is used here (Figure 16).

Figure 16. Graphics card policy



Install NVIDIA Tesla GPU card in Cisco UCS B200 M5

Install the Tesla GPU card in the Cisco UCS B200 M5 Blade 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:

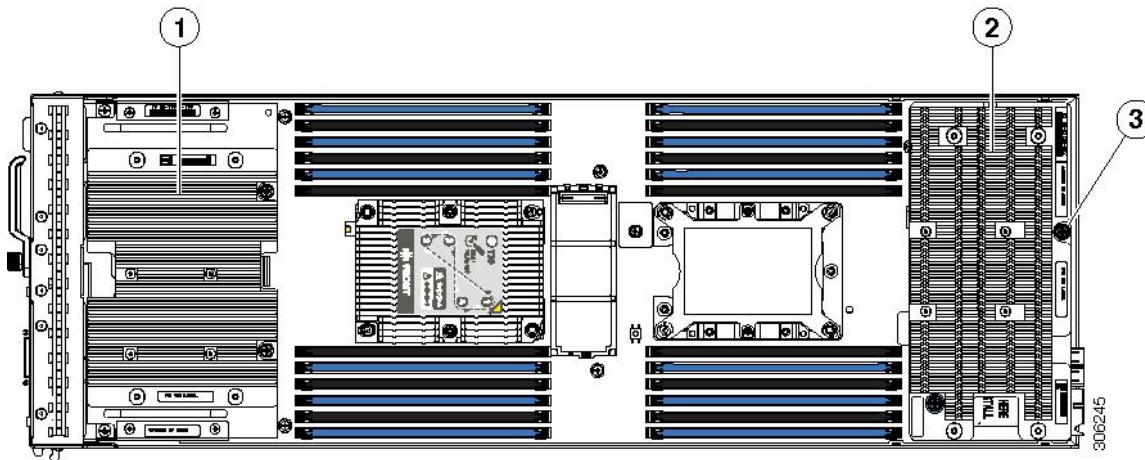
1. The UCSB-GPU-P6-F can be installed only in the front mezzanine slot of the server.

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

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

Figure 17 shows the NVIDIA P6 GPU installed in the front and rear mezzanine slots.

Figure 17. NVIDIA GPU installed in front and rear mezzanine slots

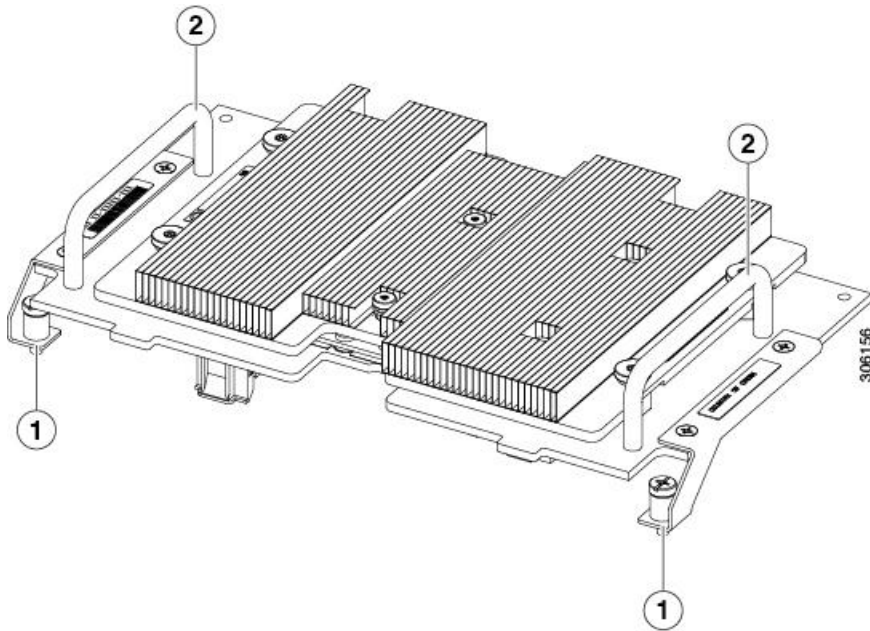


1	Front GPU	2	Rear GPU
3	Custom standoff screw		

Installing an NVIDIA GPU card in the front of the server

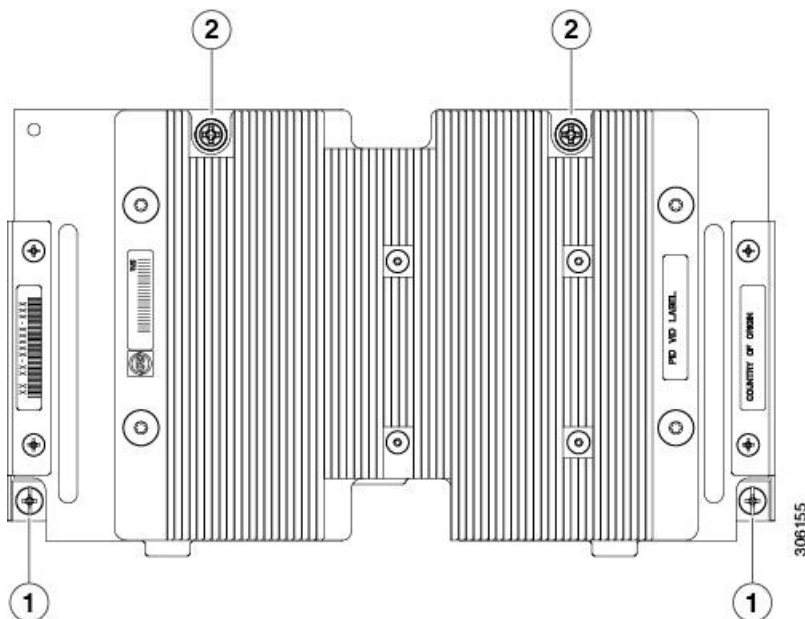
Figure 19 and Figure 20 show the front NVIDIA P6 GPU (UCSB-GPU-P6-F).

Figure 18. NVIDIA P6 GPU that installs in the front of the server



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

Figure 19. Top view of the NVIDIA P6 GPU for the front of the server



1 Leg with thumbscrew that attaches to the server motherboard	2 Thumbscrew that attaches to a standoff below
---	--

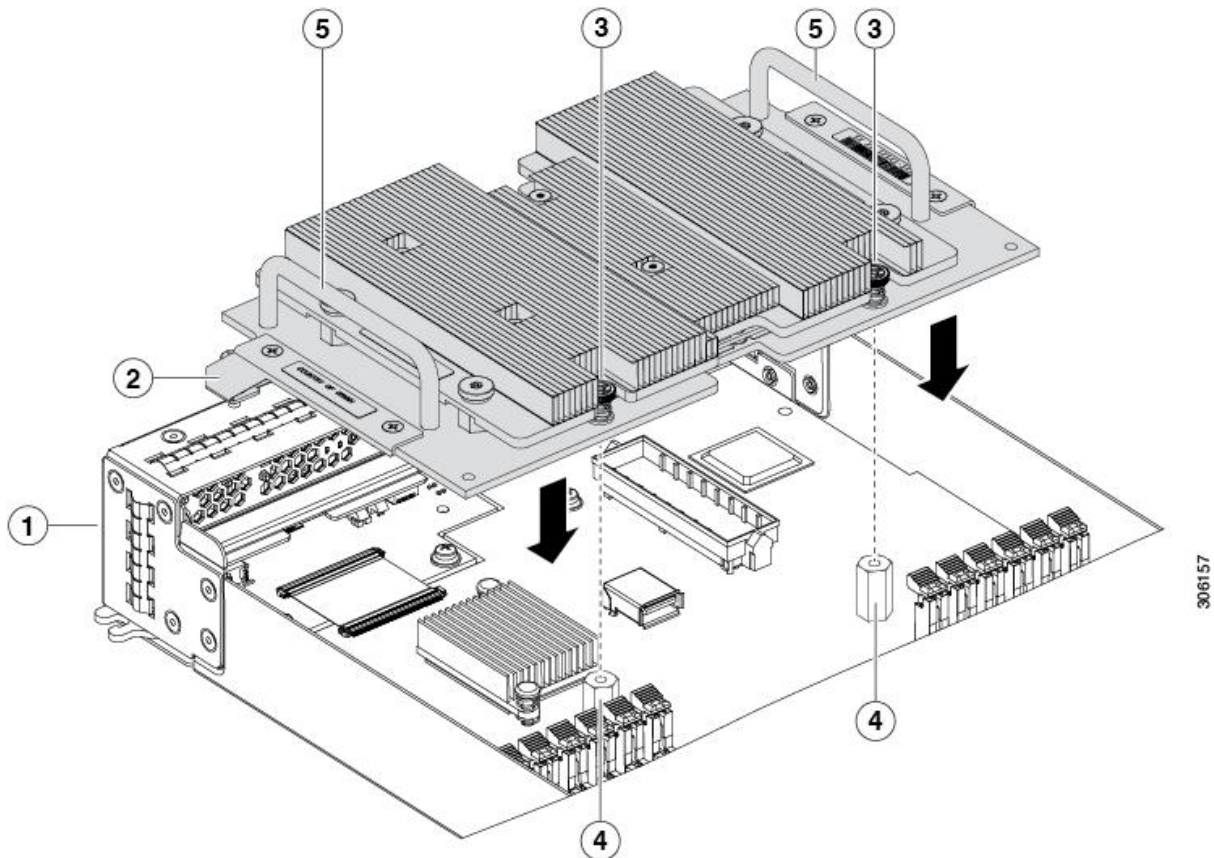
Note: Before installing the NVIDIA P6 GPU (UCSB-GPU-P6-F) in the front mezzanine slot, do the following:

- 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 at the following URL for information about supported hardware: <http://www.cisco.com/c/en/us/support/servers-unified-computing/ucs-manager/products-release-notes-list.html>.
- 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.

To install the NVIDIA P6 GPU, follow these steps:

1. Position the GPU in the correct orientation to the front of the server (callout 1) as shown in Figure 21.
2. Install the GPU into the server. Press down on the handles (callout 5) to firmly secure the GPU.
3. Tighten the thumbscrews (callout 3) at the back of the GPU with the standoffs (callout 4) on the motherboard.
4. Tighten the thumbscrews on the legs (callout 2) to the motherboard.
5. Install the drive blanking panels.

Figure 20. Installing the NVIDIA GPU in the front of the server

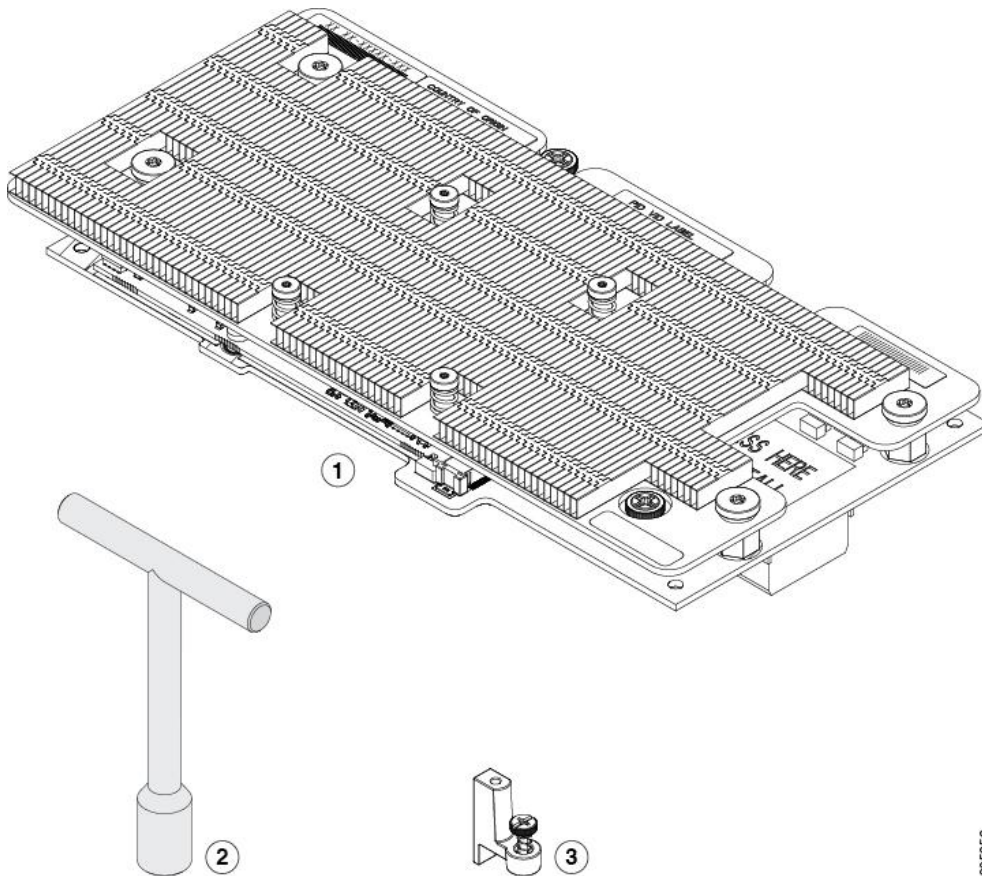


1	Front of the server	2	Leg with thumbscrew that attaches to the motherboard
3	Thumbscrew to attach to standoff below	4	Standoff on the motherboard
5	Handle to press down on to firmly install the GPU		

Installing an NVIDIA GPU card in the rear of the server

If you are installing the UCSB-GPU-P6-R in 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 22 shows the three components of the option kit.

Figure 21. NVIDIA P6 GPU (UCSB-GPU-P6-R) option kit



1	NVIDIA P6 GPU (CPU and heat sink)	2	T-shaped wrench
3	Custom standoff	-	

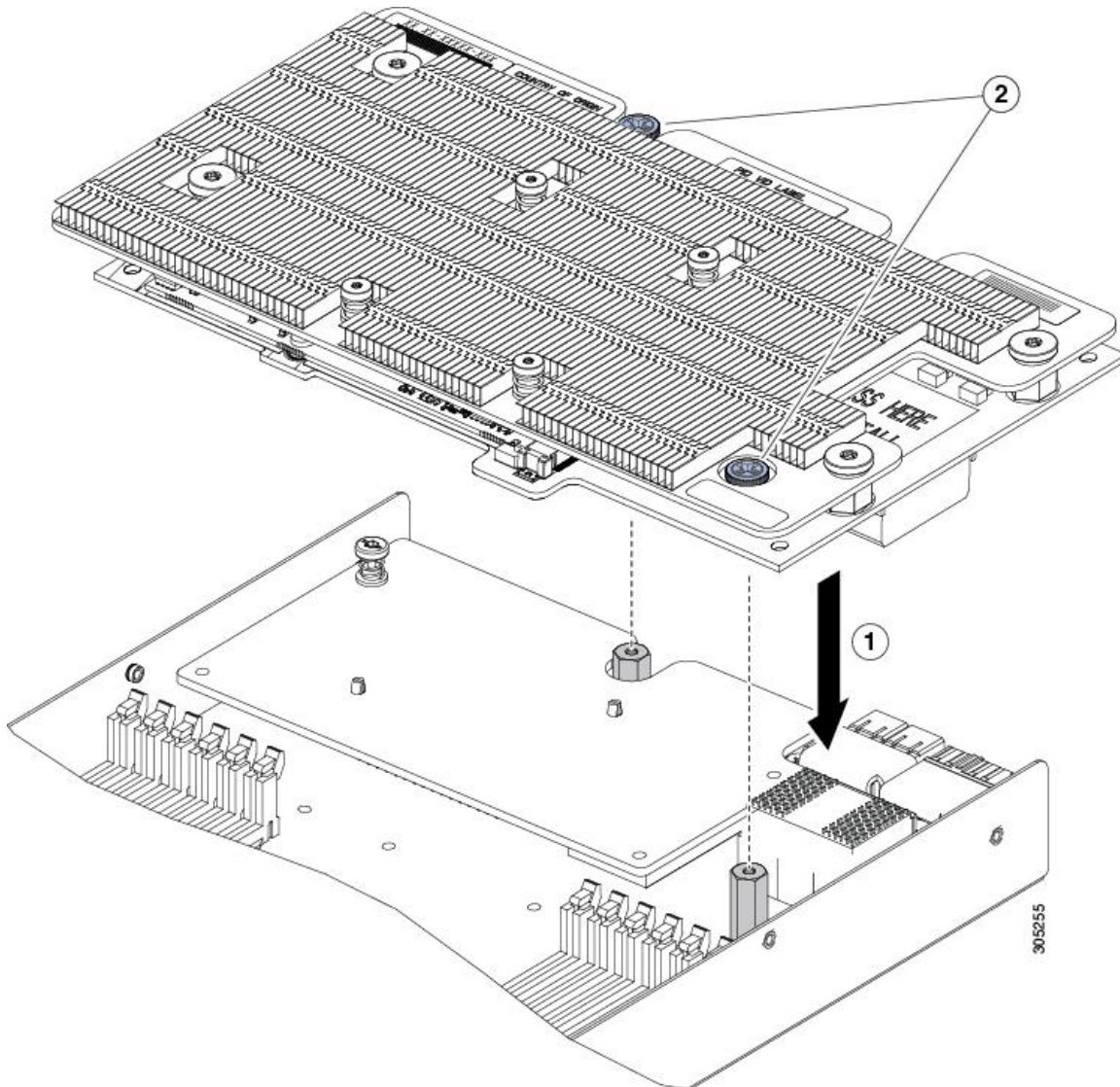
Note: Before installing the NVIDIA P6 GPU (UCSB-GPU-P6-R) in the rear mezzanine slot, do the following:

- 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 at the following URL for information about supported hardware: <http://www.cisco.com/c/en/us/support/servers-unified-computing/ucs-manager/products-release-notes-list.html>.
- 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.

Follow these steps to install the card (Figure 23):

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.
3. Position the GPU over the connector on the motherboard and align all the captive screws to the standoff posts (callout 1).
4. Tighten the captive screws (callout 2).

Figure 22. Installing the NVIDIA P6 GPU in the rear mezzanine slot



Install NVIDIA Tesla GPU card in Cisco UCS C240 M5

Install the Tesla GPU card in the Cisco UCS C240 M5 Rack Server.

Installing an NVIDIA Tesla T4

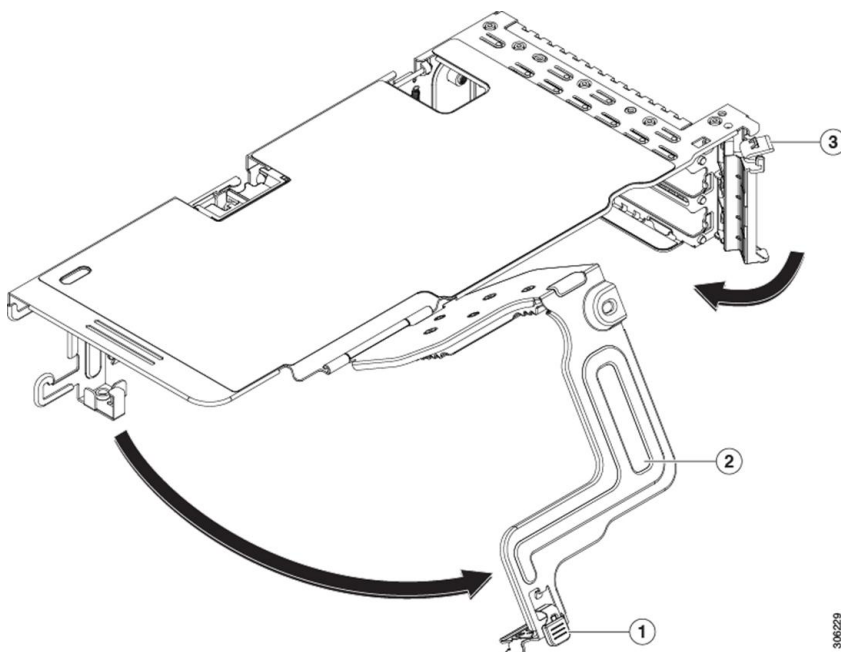
Use the following procedure to install the NVIDIA Tesla T4:

Note: This server can support up to six single-wide NVIDIA Tesla T4 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:

Note: Up to six single-wide GPU cards are supported in all 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. Ensure 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. Ensure that the clip on the plate clicks into the locked position (Figure 24).
 - 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 23. 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: NVIDIA Tesla P40

Use the following procedure to install an NVIDIA Tesla P40:

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](#).

a. 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.

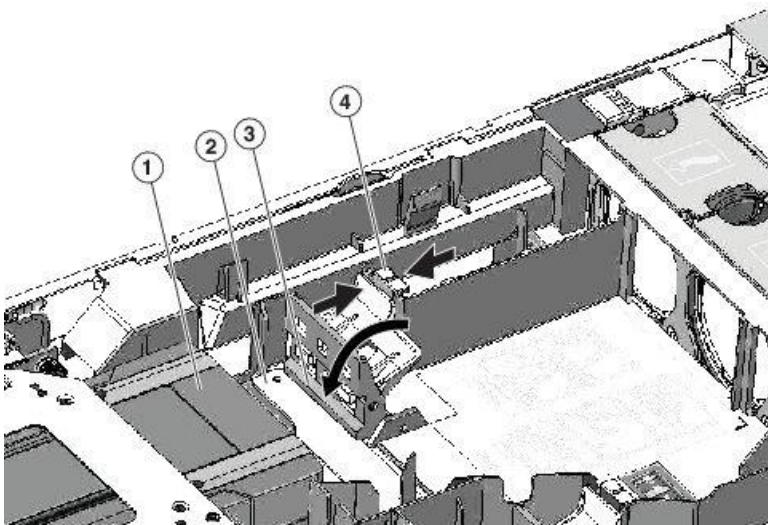
b. Connect the GPU power cable. The straight power cable connectors are color-coded. Connect the cable's black connector into the black connector on the GPU card and the cable's white connector into the white GPU POWER connector on the PCIe riser.

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.

- c. Close the card-tab retainer over the end of the card.
- d. Swing the hinged securing plate closed on the bottom of the riser. Ensure 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.
- g. 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.
- h. Insert the GPU front support bracket into the latch that is on the air baffle (Figure 25):
 - 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.
 - Ensure that the latch release tab clicks and locks the latch in place.

Figure 24. 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

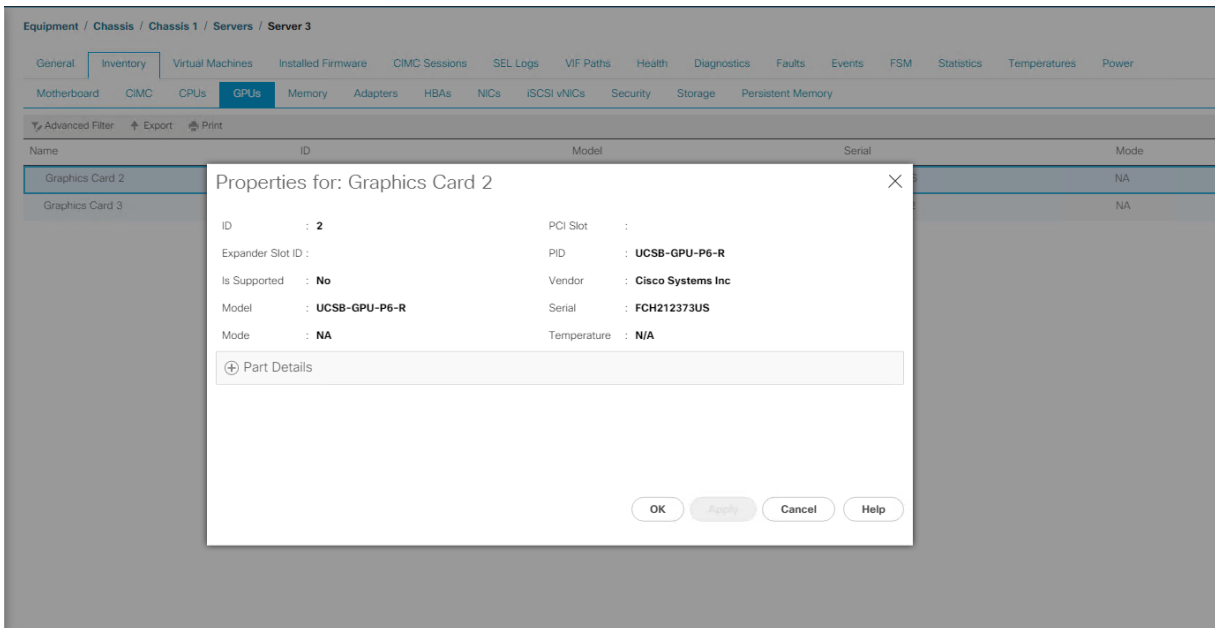
5. Replace the top cover to the server.
6. Replace the server in the rack, replace cables, and then fully power on the server by pressing the Power button.

Configure the GPU card

Follow these steps to configure the GPU card.

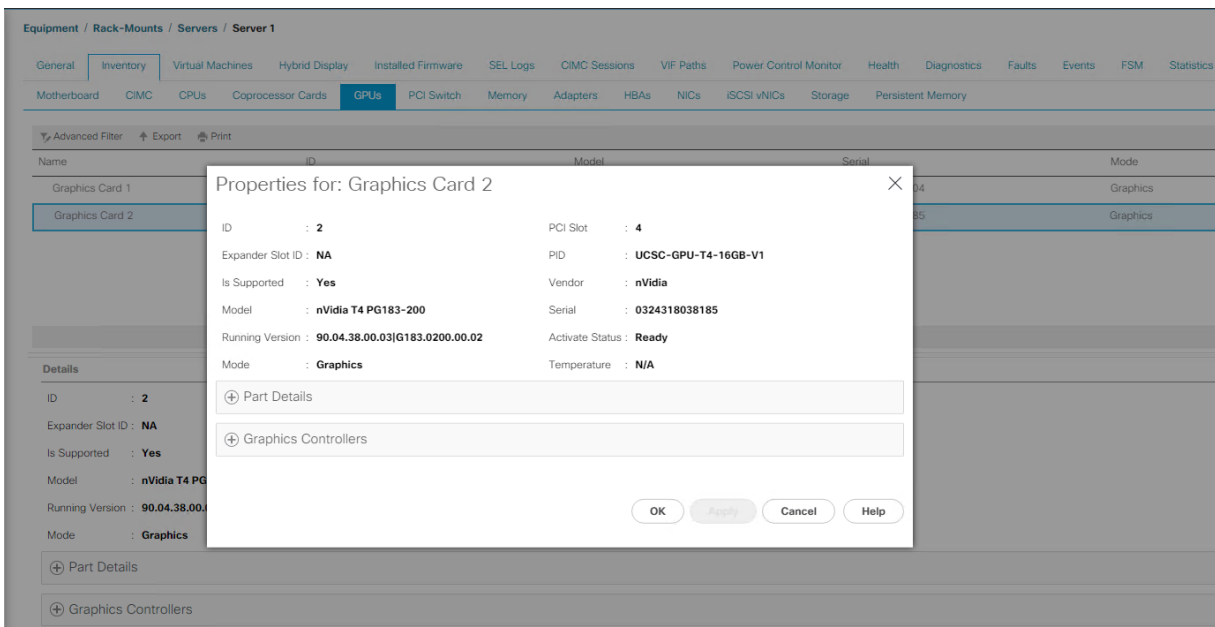
1. After the NVIDIA P6 GPU cards are physically installed and the Cisco UCS B200 M5 Blade Server is discovered in Cisco UCS Manager, select the server and choose Inventory > GPUs. As shown in Figure 26, PCIe slots 2 and 3 are used with two GRID P6 cards.

Figure 25. NVIDIA GRID P6 card inventory displayed in Cisco UCS Manager



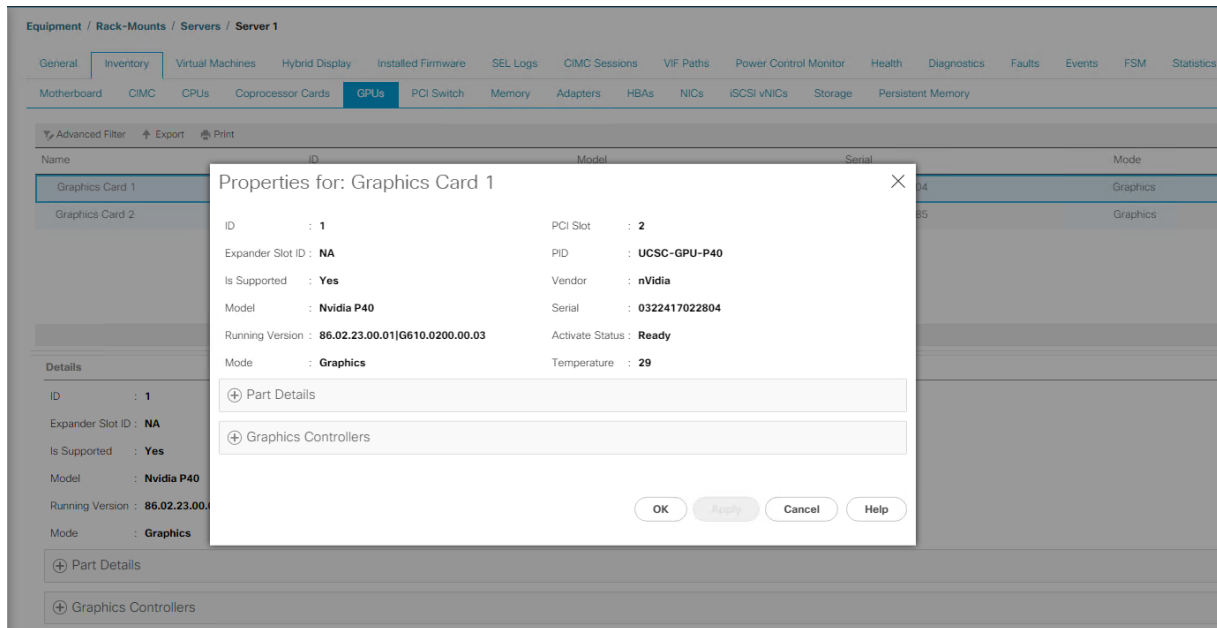
2. After the NVIDIA T4 GPU cards are physically installed and the Cisco UCS C240 M5 Rack Server is discovered in Cisco UCS Manager, select the server and choose Inventory > GPUs. As shown in Figure 27 , PCIe slots 2 and 5 are used with two GRID P4 cards.

Figure 26. NVIDIA GRID T4 card inventory displayed in Cisco UCS Manager



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 28, PCIe slots 2 and 5 are used with the two GRID P40 cards.

Figure 27. NVIDIA GRID P40 card inventory displayed in Cisco UCS Manager



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

Install the NVIDIA GRID vGPU Manager for VMware

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

Customers should install the latest NVIDIA GRID vGPU software. Instructions for obtaining the software are available from [NVIDIA](https://www.nvidia.com/en-us/grid/).

To install the vGPU Manager vSphere Installation Bundle (VIB), you need to access the ESXi host through the ESXi shell or SSH, and the VIB must be reachable from the ESXi host.

To install vGPU Manager, follow these steps:

Note: Before proceeding with the vGPU Manager installation, verify that the ESXi host is in maintenance mode.

1. Use the `esxcli` command to install the vGPU Manager package:

```
~] esxcli software vib update -v directory/NVIDIA-VMware_ESXi_Host_Driver-410.107-10EM.670.0.0.8169922.x86_64.vib
```

```
[root@CHI-B3:~] esxcli software vib install -v /vmfs/volumes/ESXTOP/ISO/NVIDIA-VMware_ESXi_6_7_Host_Driver-410.107-10EM.670.0.0.8169922.x86_64.vib
Installation Result
Message: Operation finished successfully.
Reboot Required: false
VIBs Installed: NVIDIA_bootbank_NVIDIA-VMware_ESXi_6_7_Host_Driver_410.107-10EM.670.0.0.8169922
VIBs Removed:
VIBs Skipped:
[root@CHI-B3:~] █
```

2. Reboot the ESXi host.
3. After the ESXi host reboots, verify that the GRID package has been installed and loaded correctly by checking for the NVIDIA kernel driver in the list of kernel loaded modules:

```
~] vmkload_mod -l | grep nvidia
```

```
[root@CH1-B3:~] vmkload_mod -l | grep nvidia
nvidia                204  16008
[root@CH1-B3:~]
```

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 in your platform similar to the following:

```
[root@CH1-B3:~] nvidia-smi
Mon Jun 24 22:39:11 2019

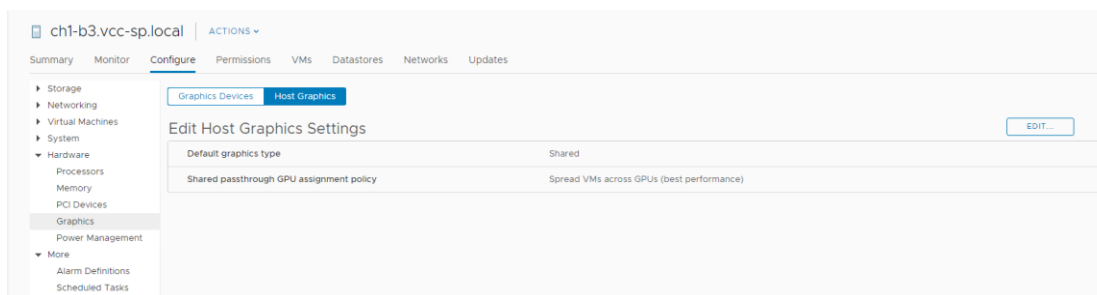
+-----+
| NVIDIA-SMI 410.107      Driver Version: 410.107      CUDA Version: N/A      |
+-----+
| GPU  Name                Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf    Pwr:Usage/Cap|  Memory-Usage | GPU-Util  Compute M. |
+-----+-----+
|   0   Tesla P6             On         | 00000000:18:00.0 Off  |          Off          |
| N/A   31C   P8     9W /  90W |  42MiB / 16383MiB |      0%   Default     |
+-----+-----+
|   1   Tesla P6             On         | 00000000:D8:00.0 Off  |          Off          |
| N/A   44C   P8    10W /  90W |  42MiB / 16383MiB |      0%   Default     |
+-----+-----+

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

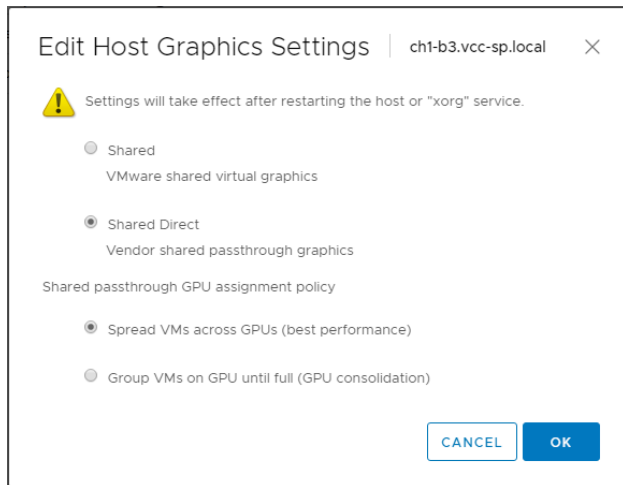
[root@CH1-B3:~]
```

5. Repeat the process for all hosts in the pool.
6. After the GRID vGPU Manager has been installed, configure Host Graphics in vCenter on all hosts in the Resource Pool.
 - a. Select the ESXi host and click the Configure tab. From the list of options at the left, select Graphics. Click Edit Host Graphics Settings (Figure 28).

Figure 28. Editing host graphics settings



- b. Select the following settings (Figure 29):
 - Shared Direct (Vendor shared passthrough graphics)
 - Spread VMs across GPUs (best performance)

Figure 29. Selecting host graphics settings

7. Reboot the ESXi host to make the changes take effect.

Disable ECC memory

GPUs based on the Pascal GPU architecture support error correcting code (ECC) memory for improved data integrity.

However, the NVIDIA vGPU does not support ECC memory. If ECC memory is enabled, vGPU fails to start. Therefore, you must ensure that ECC memory is disabled on all GPUs if you are using NVIDIA vGPU.

1. Use the `nvidia-smi -q` command to list the status of all GPUs and verify that ECC is being enabled:

```
[root@CH1-B3:~] nvidia-smi -q
=====NVSMI LOG=====

Timestamp                : Wed May 29 20:56:28 2019
Driver Version           : 410.107
CUDA Version             : Not Found

Attached GPUs            : 2
GPU 00000000:18:00.0
  Product Name           : Tesla P6
  Product Brand          : Tesla
  Display Mode           : Enabled
  Display Active         : Disabled
  Persistence Mode       : Enabled
  Accounting Mode        : Enabled
  Accounting Mode Buffer Size : 4000
```

2. Disable ECC on your host by running this command:

```
~] nvidia-smi -e 0
```

```
[root@CH1-B3:~] nvidia-smi -e 0
Disabled ECC support for GPU 00000000:18:00.0.
Disabled ECC support for GPU 00000000:D8:00.0.
All done.
Reboot required.
```

Note: You can use option `-i` to target a specific GPU. If two cards are installed in a server, run the command twice, using 0 and then 1 to represent the two GPU cards:

```
nvidia-smi -i 0 -e 0
```

Install and configure the NVIDIA GRID license server

This section summarizes the installation and configuration process for the GRID 7.2 license server.

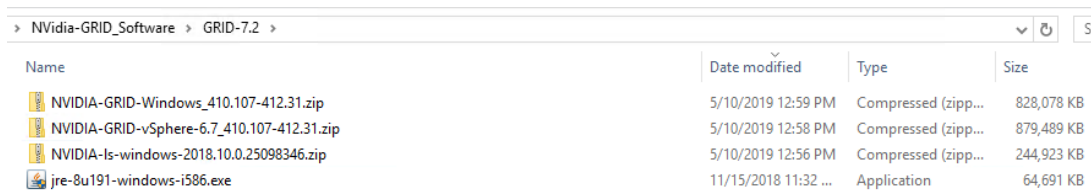
The NVIDIA GRID vGPU is a licensed feature on Tesla P6, P40, and T4 cards. A software license is required to use the full vGPU features on a guest virtual machine. A GRID license server with the appropriate licenses is required.

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 required for the Citrix environment setup are in the software folder (Figure 30):

- The GRID License Server installer
- The GRID Manager software installed on the ESXi host
- The NVIDIA drivers and software that are installed in Microsoft Windows

Figure 30. Software required for NVIDIA GRID 7.2 setup on the VMware ESXi host



Name	Date modified	Type	Size
NVIDIA-GRID-Windows_410.107-412.31.zip	5/10/2019 12:59 PM	Compressed (zipp...	828,078 KB
NVIDIA-GRID-vSphere-6.7_410.107-412.31.zip	5/10/2019 12:58 PM	Compressed (zipp...	879,489 KB
NVIDIA-ls-windows-2018.10.0.25098346.zip	5/10/2019 12:56 PM	Compressed (zipp...	244,923 KB
jre-8u191-windows-i586.exe	11/15/2018 11:32 ...	Application	64,691 KB

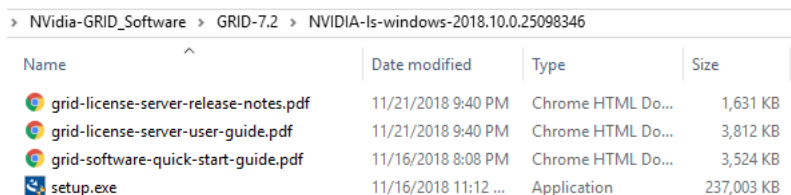
Install the GRID 7.2 license server

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 7.2 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 31).

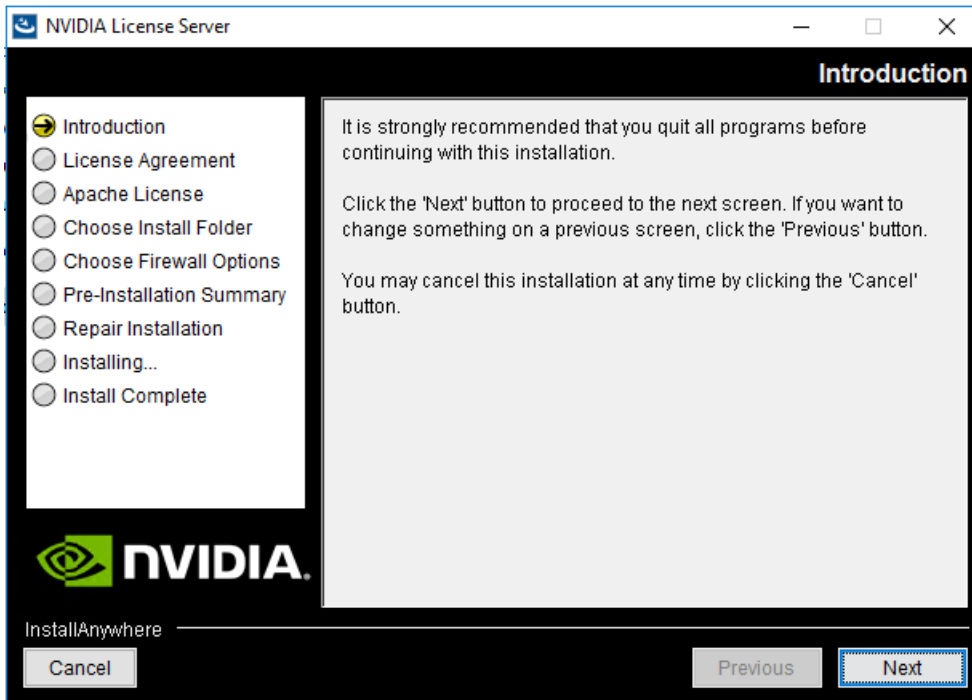
Figure 31. Run setup.exe



Name	Date modified	Type	Size
grid-license-server-release-notes.pdf	11/21/2018 9:40 PM	Chrome HTML Do...	1,631 KB
grid-license-server-user-guide.pdf	11/21/2018 9:40 PM	Chrome HTML Do...	3,812 KB
grid-software-quick-start-guide.pdf	11/16/2018 8:08 PM	Chrome HTML Do...	3,524 KB
setup.exe	11/16/2018 11:12 ...	Application	237,003 KB

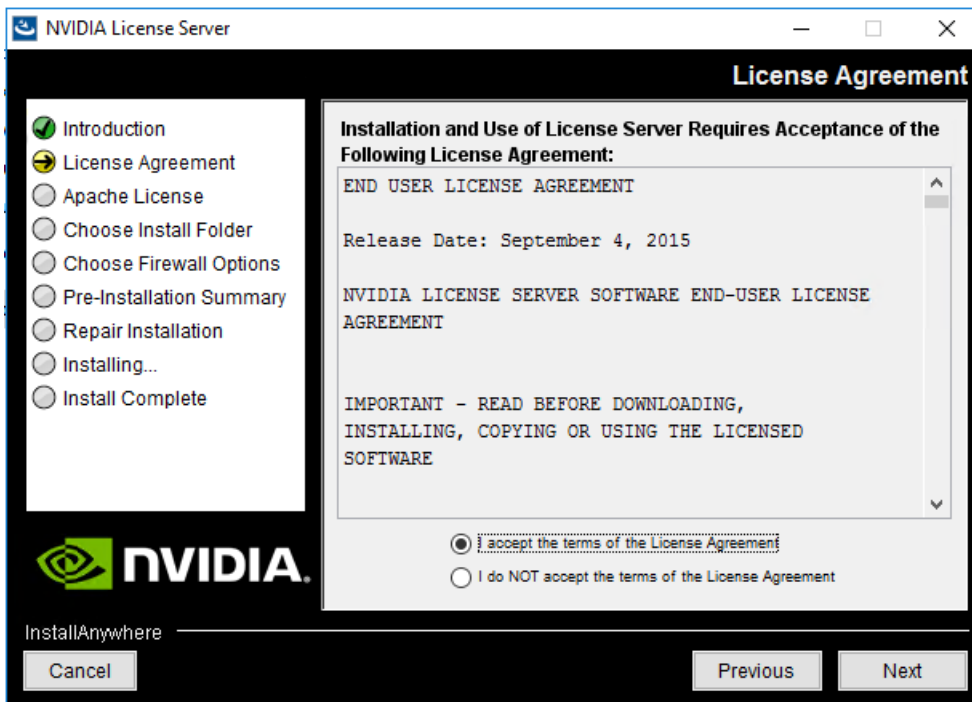
2. Click Next (Figure 32).

Figure 32. NVIDIA license server



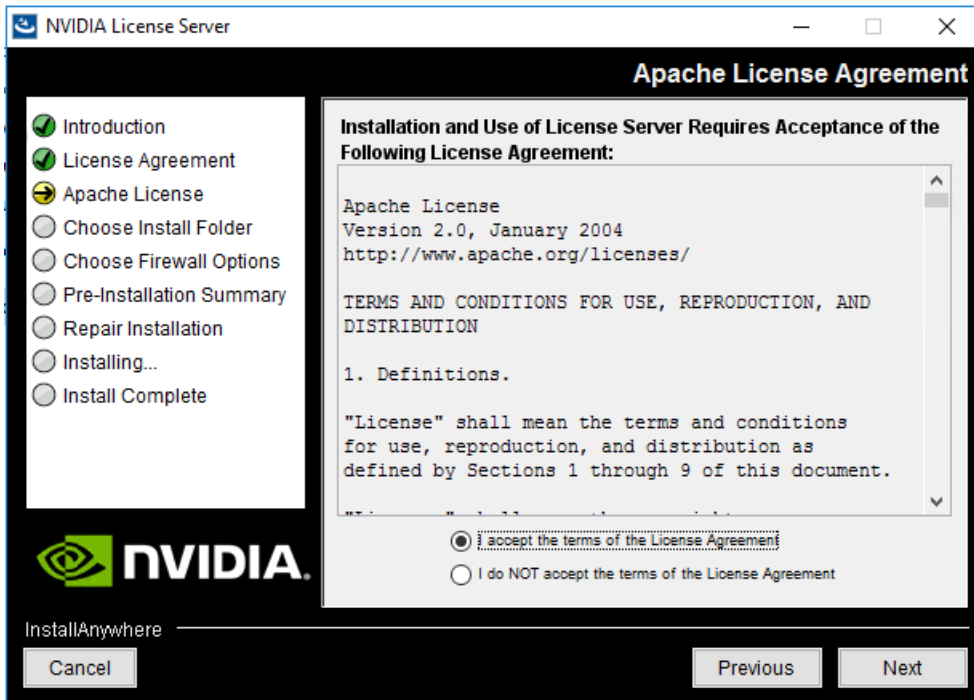
3. Accept the license agreement and click Next (Figure 33).

Figure 33. NVIDIA license agreement



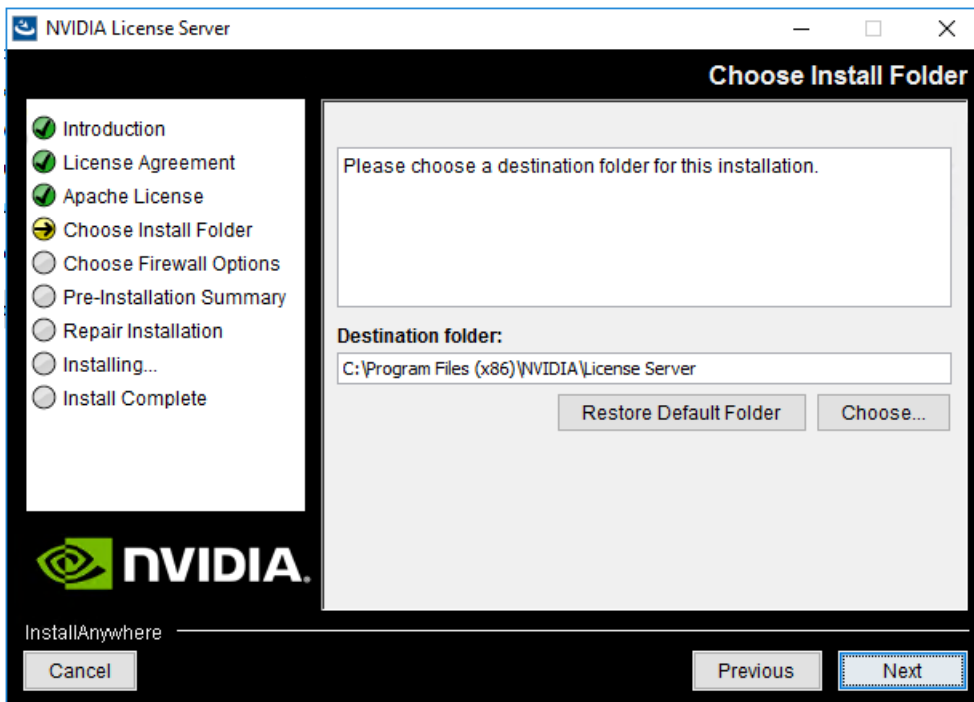
4. Accept the Apache license agreement and click Next (Figure 34).

Figure 34. Apache license agreement



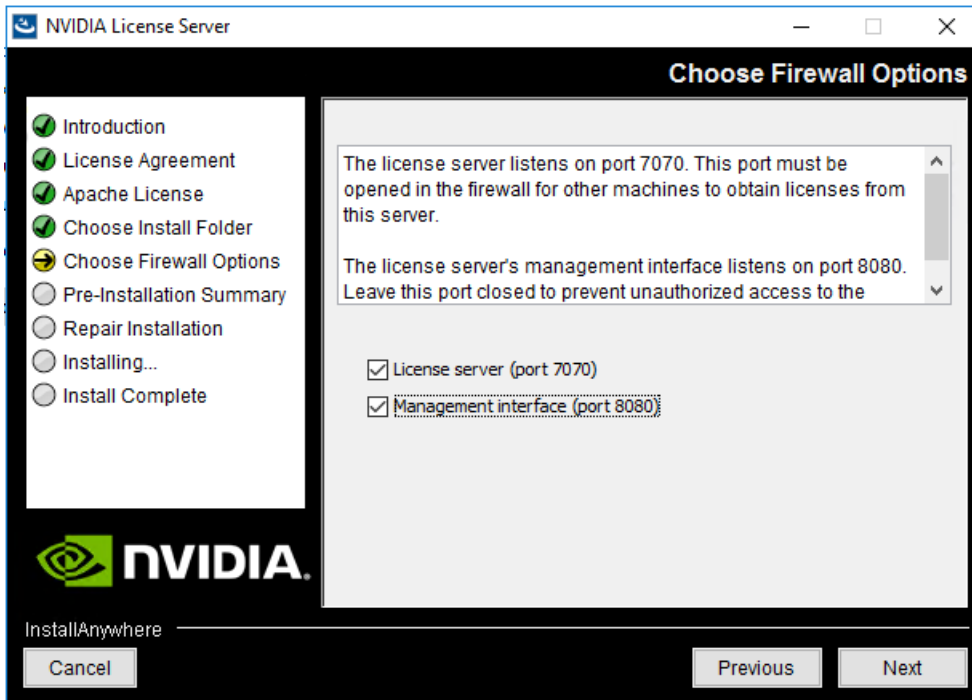
5. Choose the desired installation folder and click Next (Figure 35).

Figure 35. Choosing a destination folder

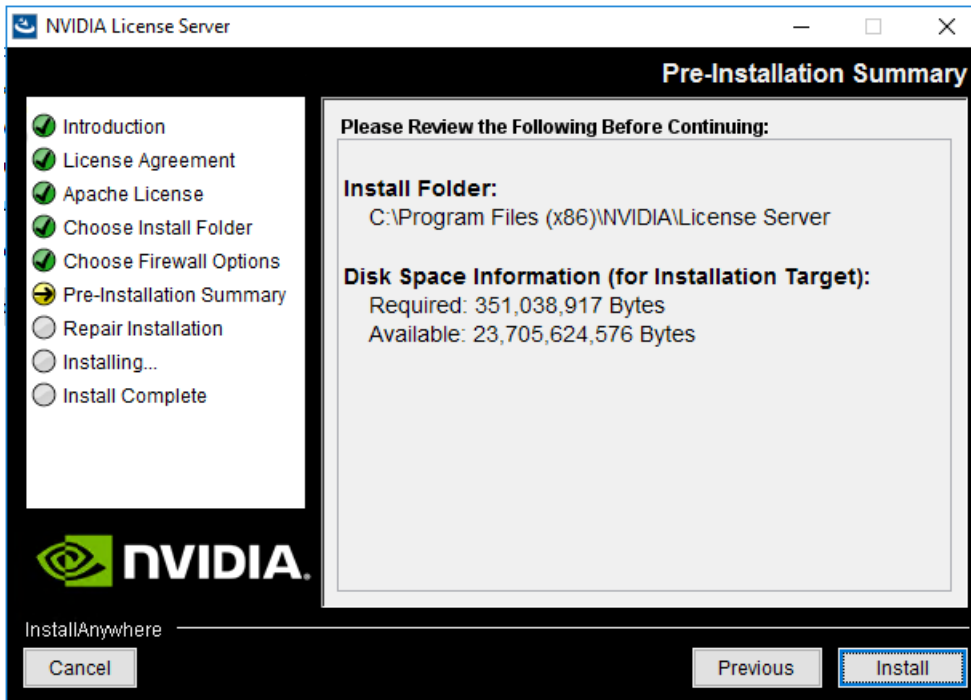


6. The license server listens on port 7070. This port must be opened in the firewall for other machines to obtain licenses from this server. Select the “License server (port 7070)” option.
7. The license server’s management interface listens on port 8080. If you want the administration page accessible from other machines, you will need to open up port 8080. Select the “Management interface (port 8080)” option.
8. Click Next (Figure 36).

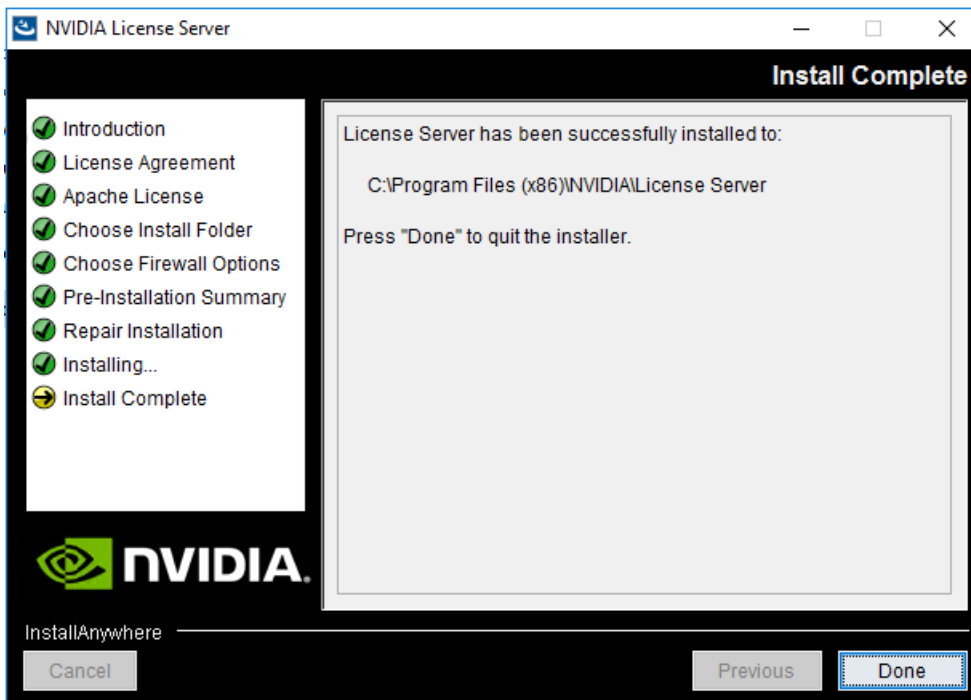
Figure 36. Setting firewall options



9. The Pre-installation Summary and Repair Installation options automatically progress without user input (Figure 37).

Figure 37. Installing the license server

10. When the installation process is complete, click Done (Figure 38).

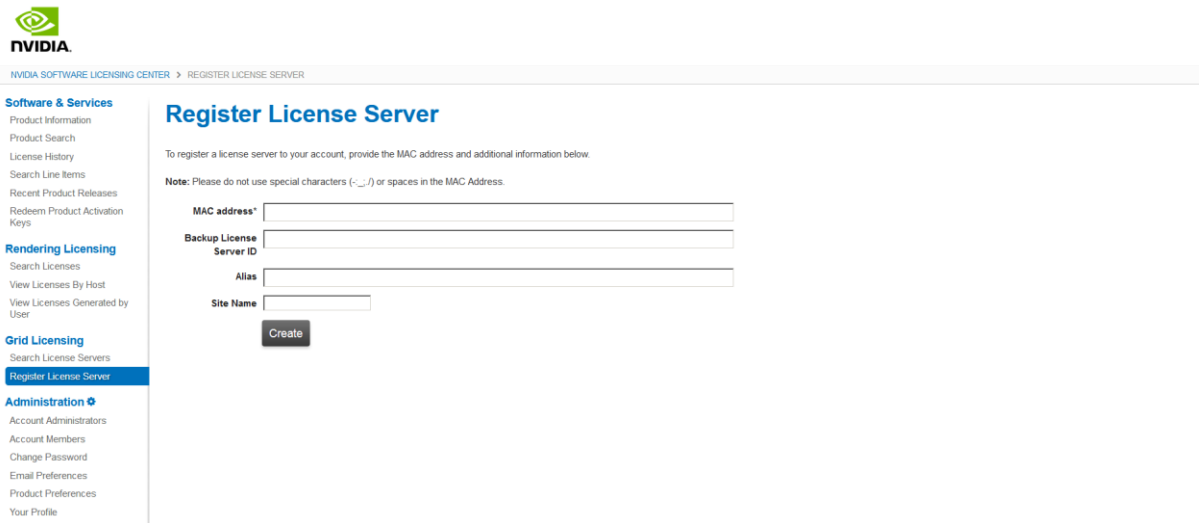
Figure 38. Installation complete

Configure the NVIDIA GRID 7.2 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 39. In the License Server ID field, enter the MAC address of your local license server's NIC. Leave the ID Type set to Ethernet. For the Alias and Site Name, choose user-friendly names. Then click Create.

Figure 39. Registering the license server



NVIDIA
NVIDIA SOFTWARE LICENSING CENTER > REGISTER 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:

Software & Services
Product Information
Product Search
License History
Search Line Items
Recent Product Releases
Redeem Product Activation Keys

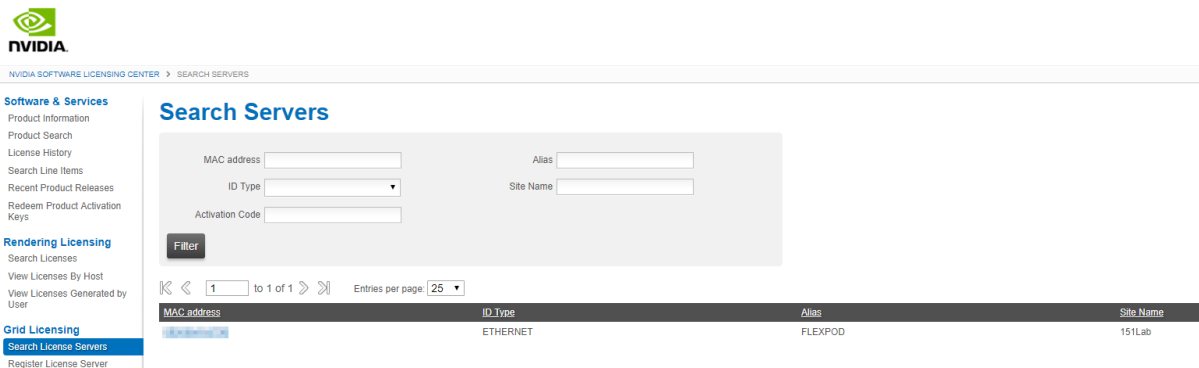
Rendering Licensing
Search Licenses
View Licenses By Host
View Licenses Generated by User

Grid Licensing
Search License Servers
Register License Server

Administration
Account Administrators
Account Members
Change Password
Email Preferences
Product Preferences
Your Profile

4. Click the Search License Servers node.
5. Click your license server ID (Figure 40).

Figure 40. Selecting the license server ID



NVIDIA
NVIDIA SOFTWARE LICENSING CENTER > SEARCH SERVERS

Search Servers

MAC address: Alias:

ID Type: Site Name:

Activation Code:

1 of 1 Entries per page: 25

MAC address	ID Type	Alias	Site Name
	ETHERNET	FLEXPOD	151Lab

Software & Services
Product Information
Product Search
License History
Search Line Items
Recent Product Releases
Redeem Product Activation Keys

Rendering Licensing
Search Licenses
View Licenses By Host
View Licenses Generated by User

Grid Licensing
Search License Servers
Register License Server

6. Click Map Add-Ons and choose the number of license units out of your total pool to allocate to this license server (Figure 41). Then select Map Add-Ons (Figure 42).

Figure 41. Choosing the number of license units

Map Add-Ons

Search Add-Ons for Server

Activation Code Entitlement ID

Add-On Name Feature Name

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	XXXXXXXXXXXXXXXXXXXX	XXXXXXXXXXXXXXXXXXXX	Perpetual	64	64	<input type="text"/>
Quadro vDWS Edition, Subscription License, 1 CCU, NFR	XXXXXXXXXXXXXXXXXXXX	XXXXXXXXXXXXXXXXXXXX	Exp. 01/01/2020	40	144	<input type="text"/>

Figure 42. Mapped add-ons after successful mapping

View Server

MAC address

ID Type ETHERNET

Alias

Site Name 151Lab

[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	XXXXXXXXXXXXXXXXXXXX	48	

7. Click Download License File and save the .bin file to your license server (Figure 43).

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 43. Saving the .bin file

View Server

MAC address

ID Type ETHERNET

Alias

Site Name 151Lab

[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	XXXXXXXXXXXXXXXXXXXX	48	

- On the local license server, browse to <http://<FQDN>:8080/licserver> to display the License Server Configuration page.
- Click License Management in the left pane.
- Click Browse to locate your recently download .bin license file. Select the .bin file and click OK.
- Click Upload. The message “Successfully applied license file to license server” should appear on the screen (Figure 44). The features are available (Figure 45).

Figure 44. License file successfully applied

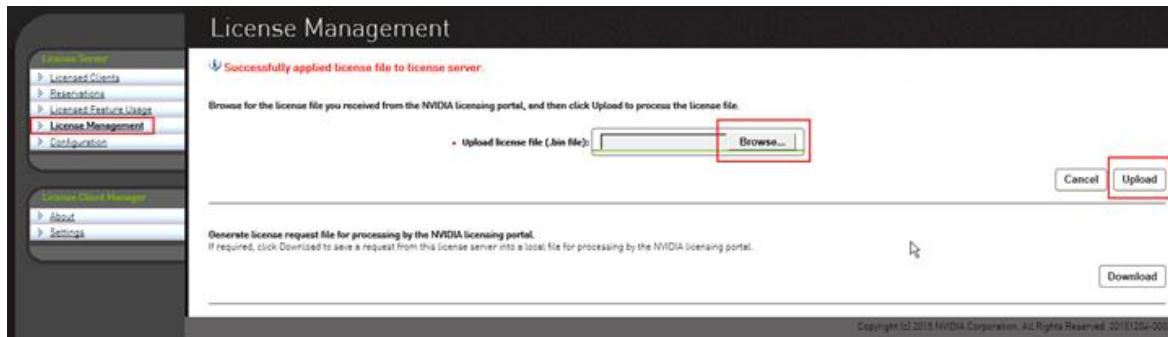
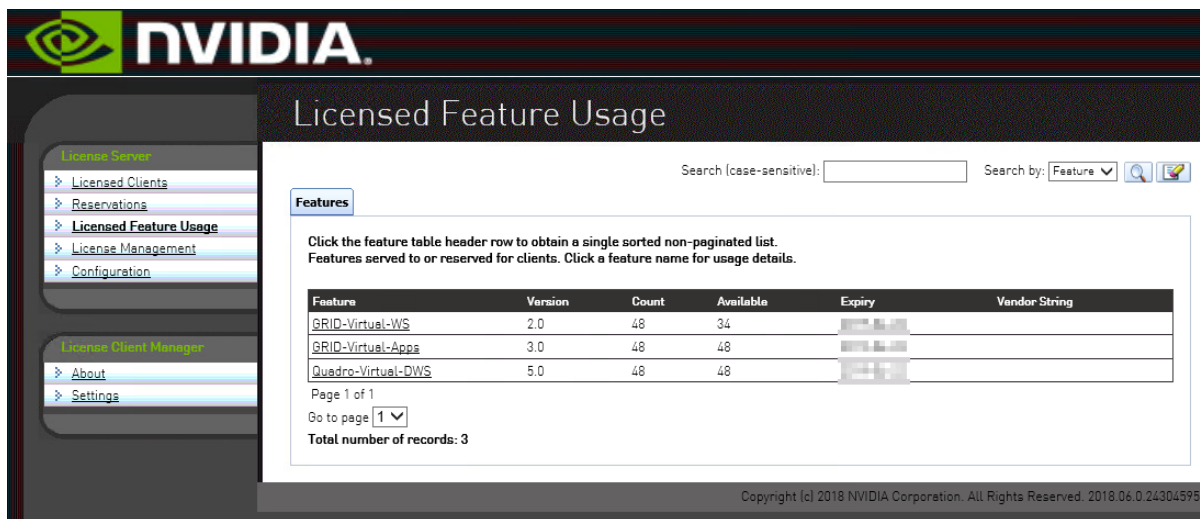


Figure 45. NVIDIA license server with licensed features ready for use



NVIDIA Tesla P6, P40, and T4 profile specifications

The Tesla P6, T4 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 3 lists the vGPU types supported by GRID GPUs.

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

Table 3. User profile specifications for NVIDIA Tesla cards

End user and GRID options			
End-user profile	GRID Virtual App Profiles	GRID Virtual PC Profiles	Quadro vDWS profiles
1 GB	P6-1A T4-1A P40-1A	P6-1B T4-1B P40-1B	P6-1Q T4-1Q P40-1Q
2 GB	P6-2A T4-2A P40-2A	P6-2B T4-2B T4-2B4 P40-2B	P6-2Q T4-2Q P40-2Q
3 GB	P40-3A	–	P40-3Q
4 GB	P6-4A T4-4A P40-4A	–	P6-4Q T4-4Q P40-4Q
6GB	P40-6A	–	P40-6Q
8 GB	P6-8A T4-8A P40-8A	–	P6-8Q T4-8Q P40-8Q
12 GB	P40-12A	–	P40-12Q
16 GB	P6-16A T4-16A	–	P6-16Q T4-16Q
24 GB	P40-24A	–	P40-24Q

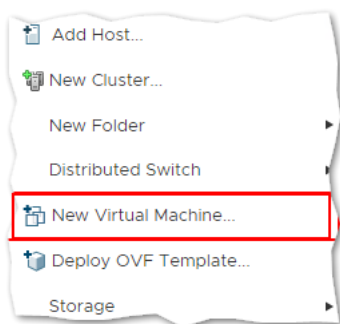
Create virtual desktops with vGPU support

Use the procedures in this section to create virtual desktops with vGPU support.

Create the Citrix XenDesktop base image

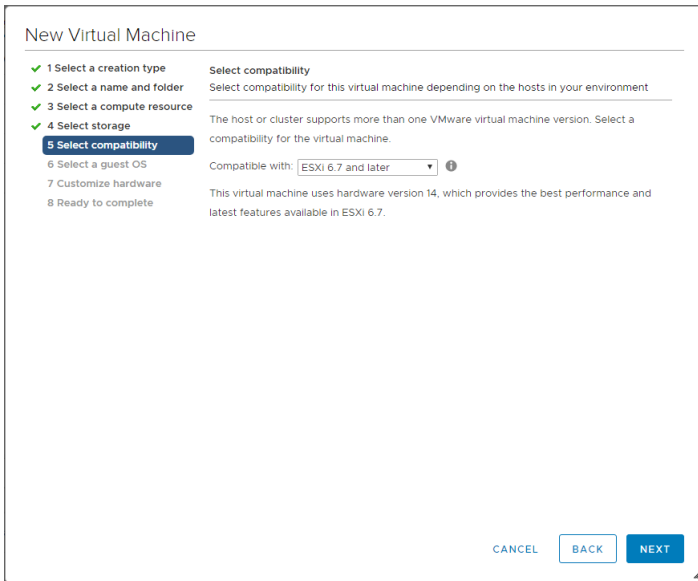
Create the virtual machine that will later be used as the virtual desktop base image.

- Using vCenter, create a new virtual machine. To do this, right-click a host or cluster and choose New Virtual Machine. Work through the New Virtual Machine wizard (Figure 46).

Figure 46. Creating a new virtual machine in VMware vSphere Web Client

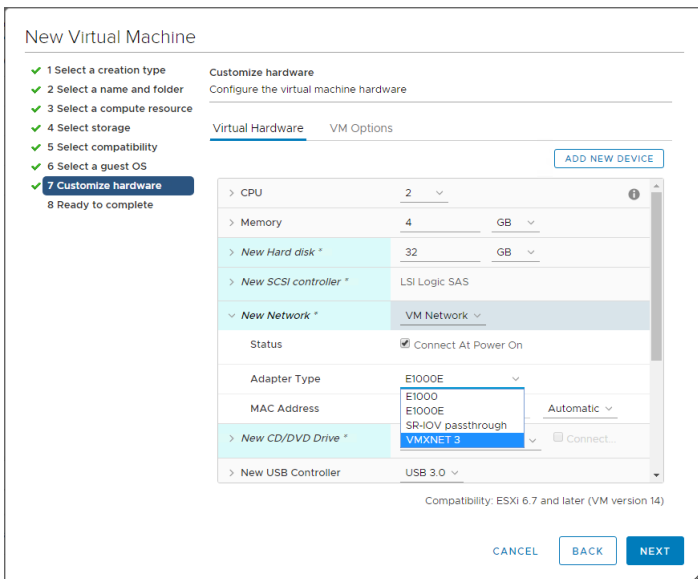
- Choose “ESXi 6.7 and later” from the “Compatible with” drop-down menu to use the latest features, including the mapping of shared PCI devices, which is required for the vGPU feature (Figure 47).

Figure 47. Selecting the virtual machine version and compatibility



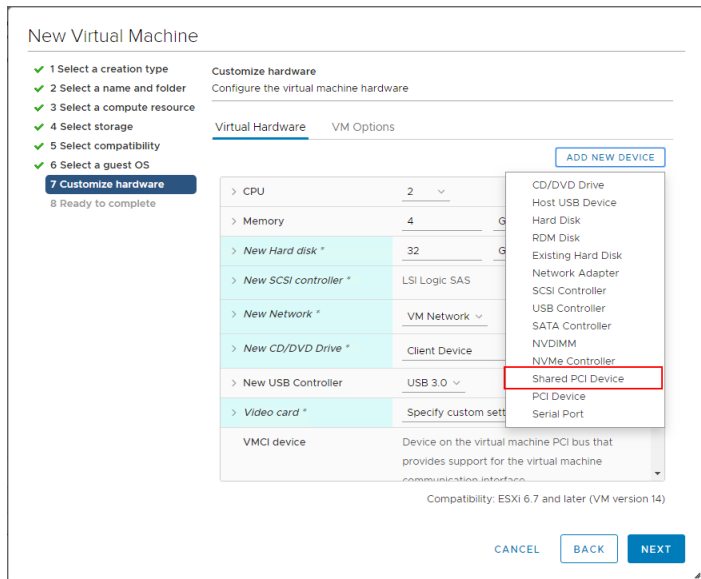
3. In the “Customize hardware” step of the New Virtual Machine wizard, select New Network and choose VMXNET 3 as the adapter type for your virtual machine (Figure 48).

Figure 48. Selecting the network adapter type



4. Click Add New Device and choose Shared PCI Device (Figure 49).

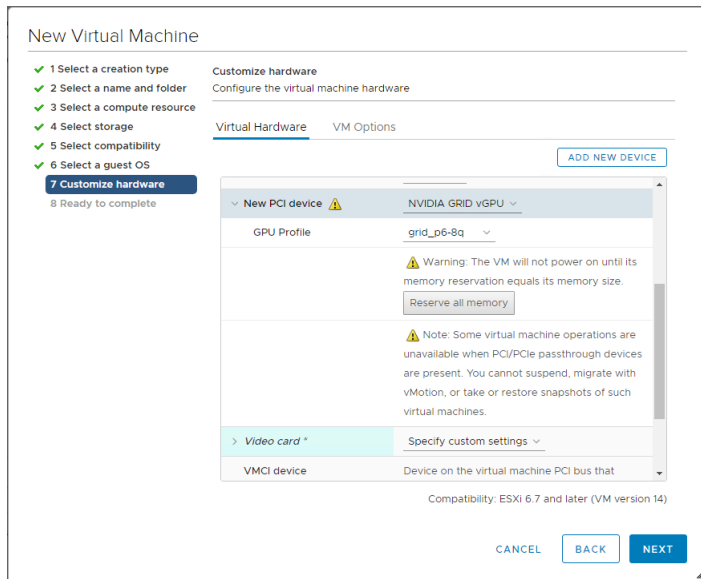
Figure 49. Adding a shared PCI device to the virtual machine to attach the GPU profile



5. Select the appropriate GPU profile and reserve all virtual machine memory (Figure 50).

Note: Allocating vGPU to a virtual machine requires you to reserve all guest memory.

Figure 50. GPU profile selection and memory reservation



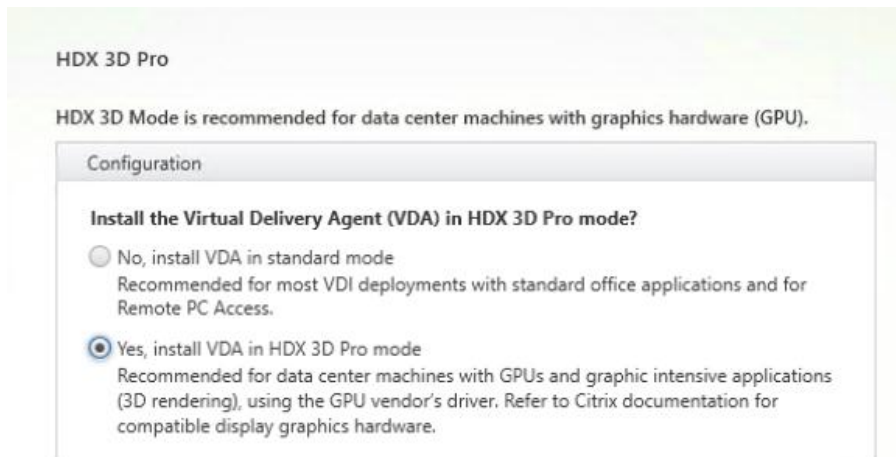
6. Install and configure Microsoft Windows on the virtual machine:

- a. Install VMware Tools.
- b. Install SPECviewperf 13.
- 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 51).

- 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 51. Installing upgrade Citrix HDX 3D Pro VDA



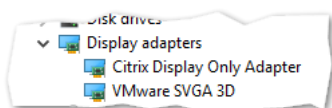
e. Optimize the Windows OS. [Citrix Optimizer](#), the optimization tool, includes customizable templates to enable or disable Windows system services. Because most Windows system services are enabled by default, the optimization tool can be used to easily disable unnecessary services and features to improve performance.

Install the NVIDIA 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.

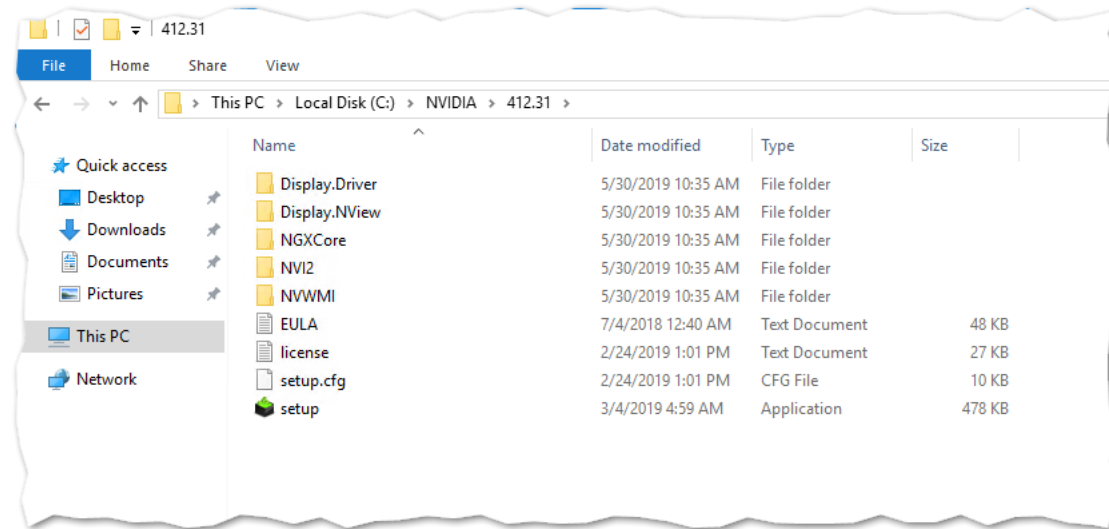
Before the NVIDIA driver is installed on the guest virtual machine, the Device Manager shows the VMware and Citrix display adapters (Figure 52).

Figure 52. 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 64-bit NVIDIA Windows driver from the vGPU driver pack to the desktop virtual machine and run `setup.exe` (Figure 53).

Figure 53. NVIDIA driver pack



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

3. Install the graphics drivers using the Express Option (Figure 54). After the installation has been completed successfully, choose Close (Figure 55) and restart the virtual machine.

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

Figure 54. Select the Express or Custom installation option

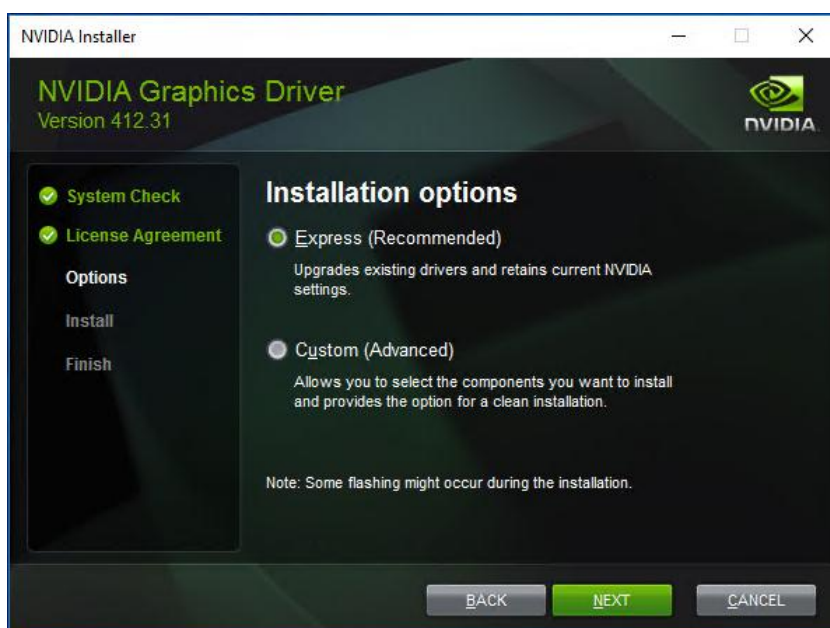
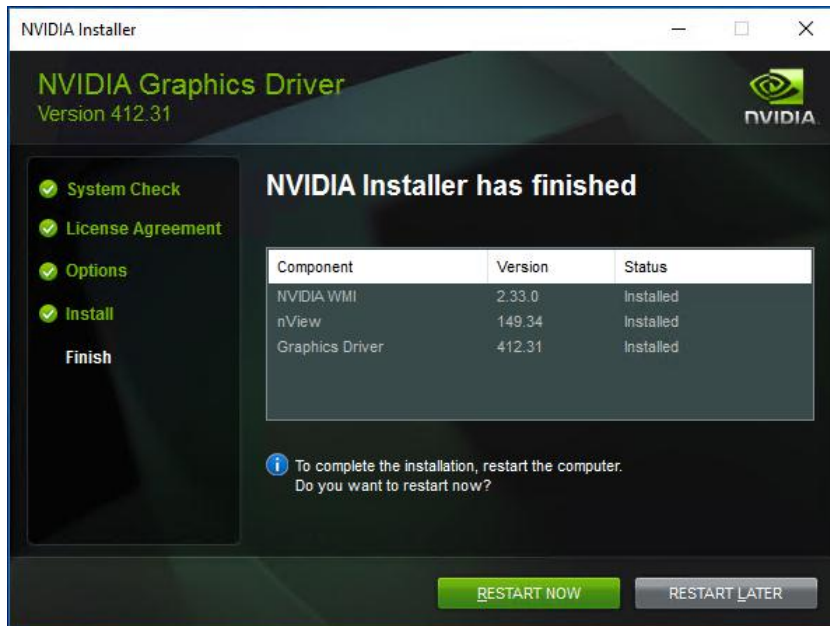


Figure 55. Components installed during the NVIDIA graphics driver installation process

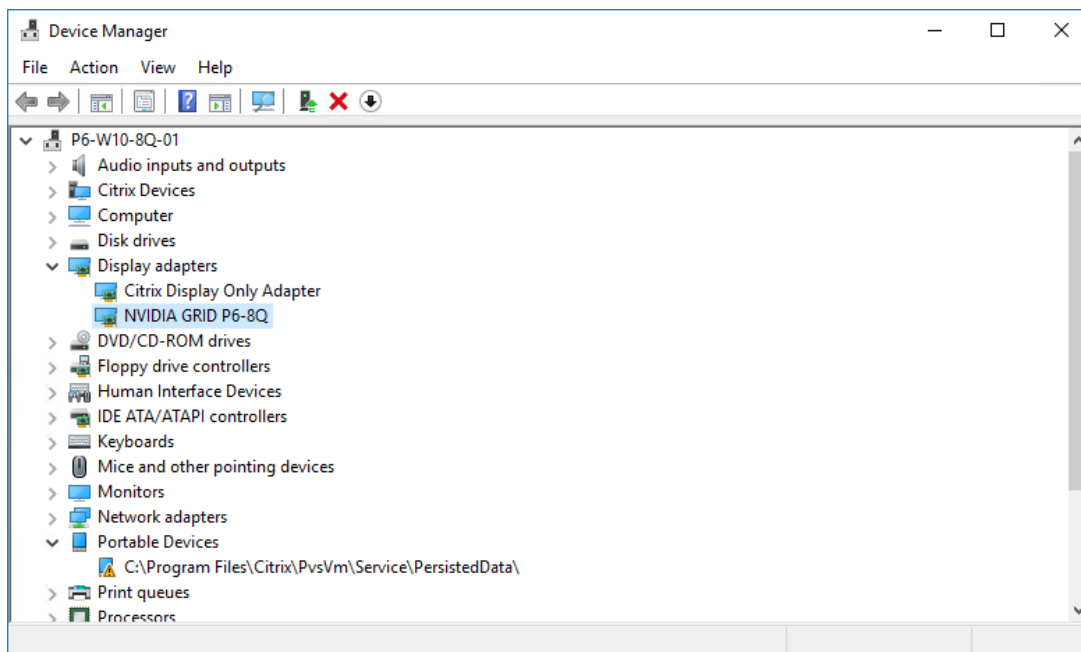


Verify that the virtual machine is ready to support vGPU

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

1. Open the Windows Device Manager and expand the Display Adapter section. The device will reflect the chosen profile (Figure 56).

Figure 56. Validating the driver installation with Device Manager



2. Verify that NVIDIA GRID is enabled by using the NVFBCEnable tool provided by NVIDIA (Figure 57).

Figure 57. Validating the driver installation with the NVFBCEnable tool

```
Administrator: Windows PowerShell
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 58, enable it with NVFBCEnable tool as shown in Figure 59 and reboot the virtual machine.

Figure 58. Validating the driver installation with the NVFBCEnable tool: NvFBC disabled

```
Administrator: Windows PowerShell
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 59. Validating the driver installation with the NVFBCEnable tool: Enabling NvFBC

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

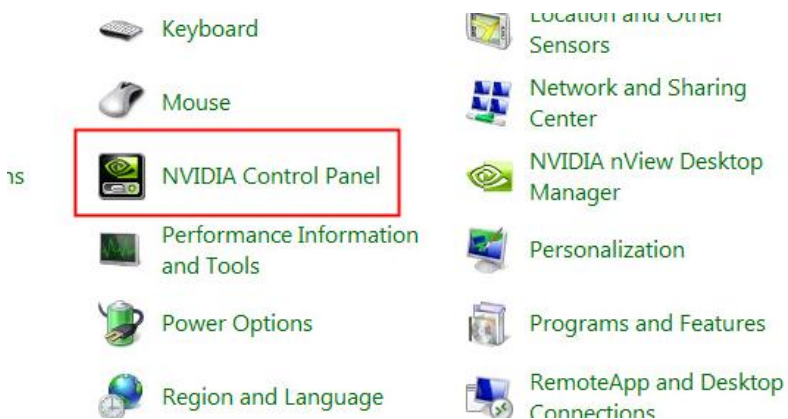
Configure the virtual machine for the NVIDIA GRID vGPU license

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

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

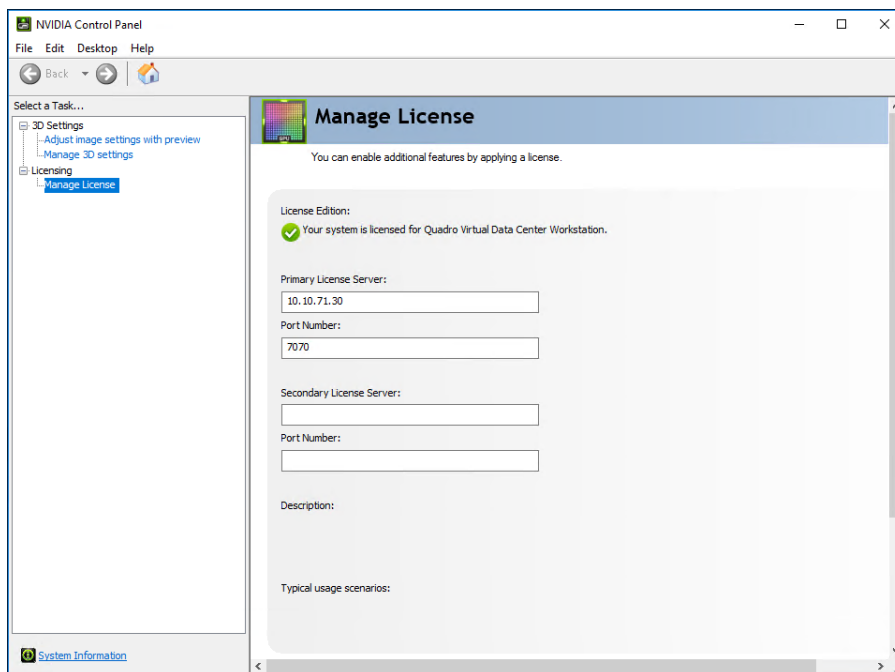
1. In the Microsoft Windows Control Panel, double-click NVIDIA Control Panel (Figure 60).

Figure 60. Choosing the NVIDIA Control Panel



2. Select Manage License from the left pane and enter your license server address and port number (Figure 61).

Figure 61. Managing your license

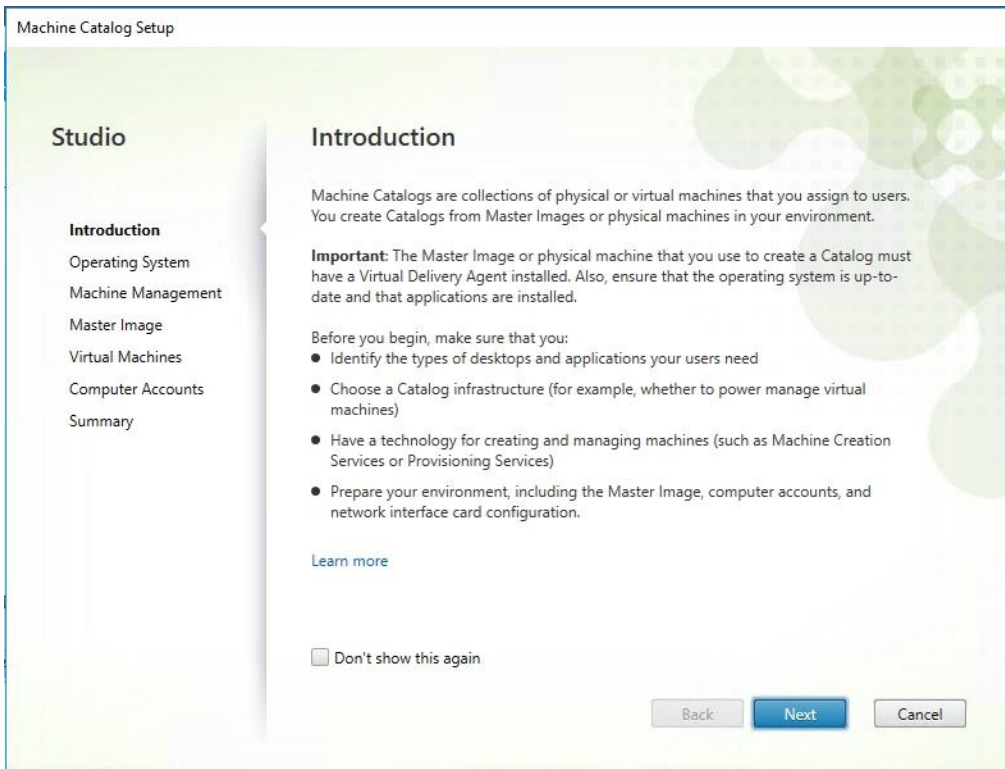


3. Click Apply.

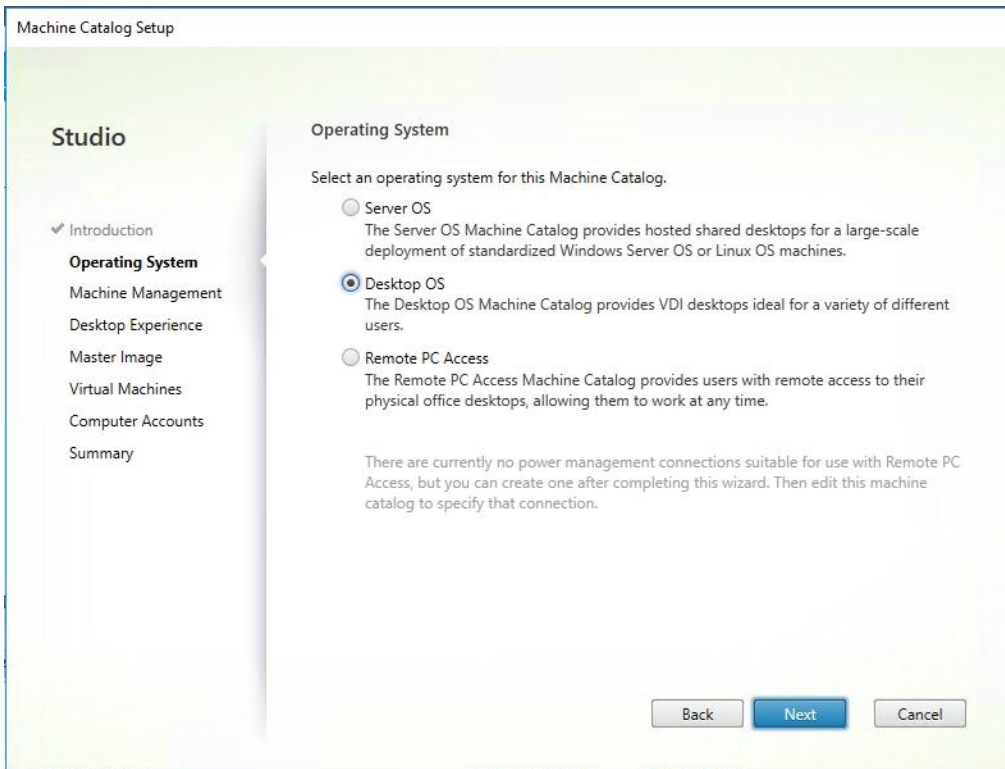
Deploy virtual machines with Citrix Machine Creation Services

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

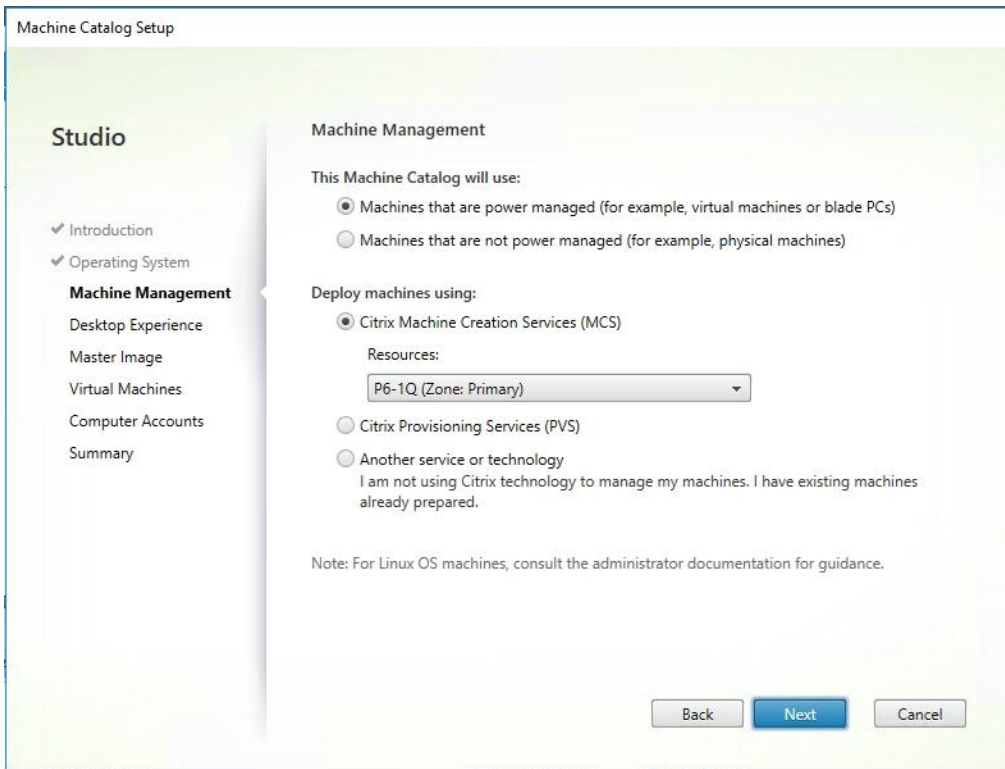
1. Connect to a XenDesktop server and launch Citrix Studio.
2. From the Actions pane, choose Create Machine Catalog. Click Next (Figure 62).

Figure 62. Create a machine catalog

3. Select Desktop OS. Click Next (Figure 63).

Figure 63. Selecting Desktop OS

4. Select the appropriate machine management. Select the resource that will provision the virtual machine with the required GPU profile. Then click Next (Figure 64).

Figure 64. Setting up virtual machine management

The screenshot shows the 'Machine Catalog Setup' wizard in the 'Studio' environment. The left sidebar lists the steps: Introduction, Operating System, Machine Management (selected), Desktop Experience, Master Image, Virtual Machines, Computer Accounts, and Summary. The main content area is titled 'Machine Management' and contains the following options:

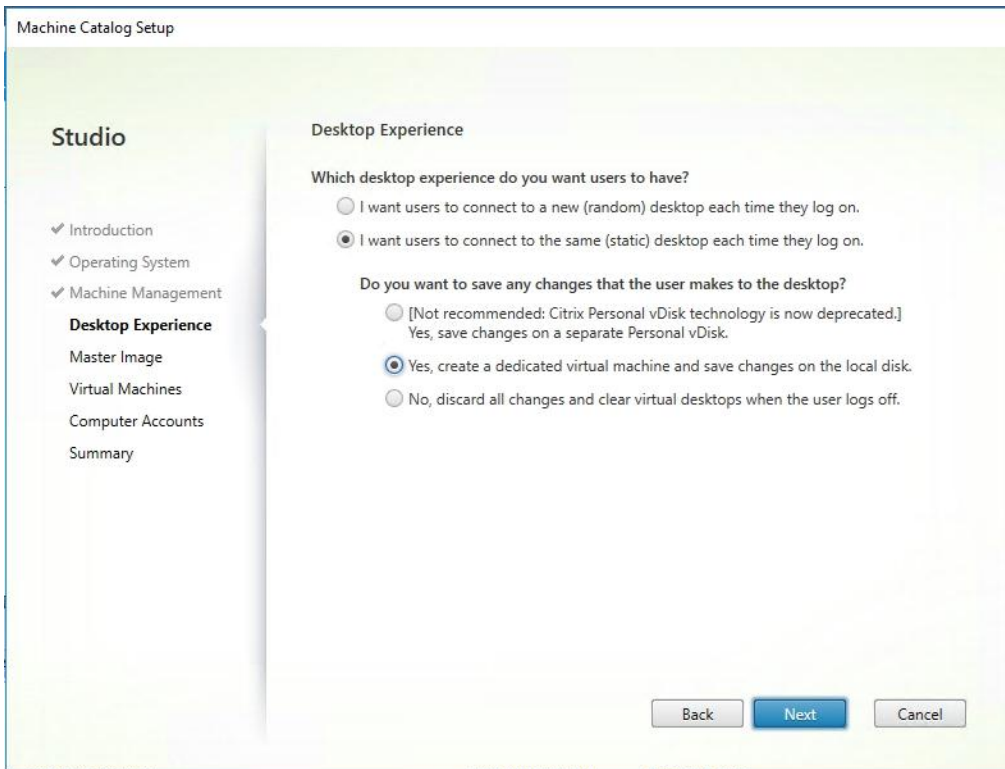
- This Machine Catalog will use:**
 - Machines that are power managed (for example, virtual machines or blade PCs)
 - Machines that are not power managed (for example, physical machines)
- Deploy machines using:**
 - Citrix Machine Creation Services (MCS)
 - Resources:
P6-1Q (Zone: Primary)
 - Citrix 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.

At the bottom right, there are three buttons: 'Back', 'Next' (highlighted in blue), and 'Cancel'.

5. For Desktop Experience, select a static, dedicated virtual machine. Then click Next (Figure 65).

Figure 65. Selecting the desktop experience



The screenshot shows the 'Machine Catalog Setup' wizard in the 'Desktop Experience' step. On the left is a 'Studio' sidebar with a list of steps: Introduction, Operating System, Machine Management, Desktop Experience (highlighted), Master Image, Virtual Machines, Computer Accounts, and Summary. The main area contains two questions with radio button options:

Desktop Experience

Which desktop experience do you want users to have?

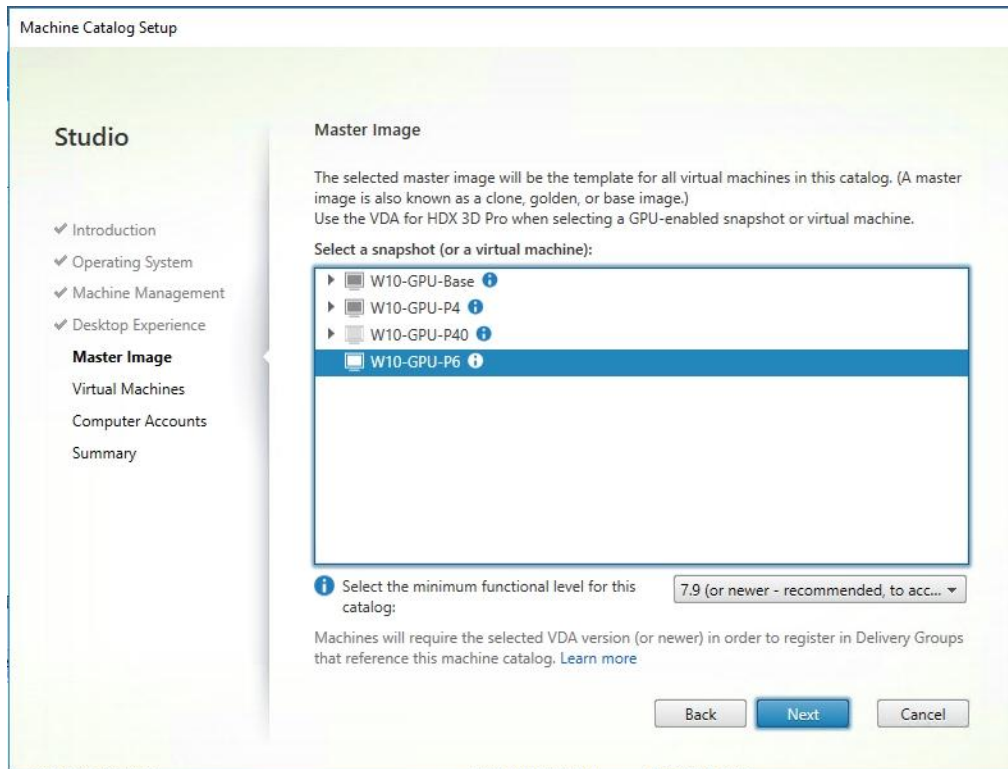
- I want users to connect to a new (random) desktop each time they log on.
- I want users to connect to the same (static) desktop each time they log on.

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

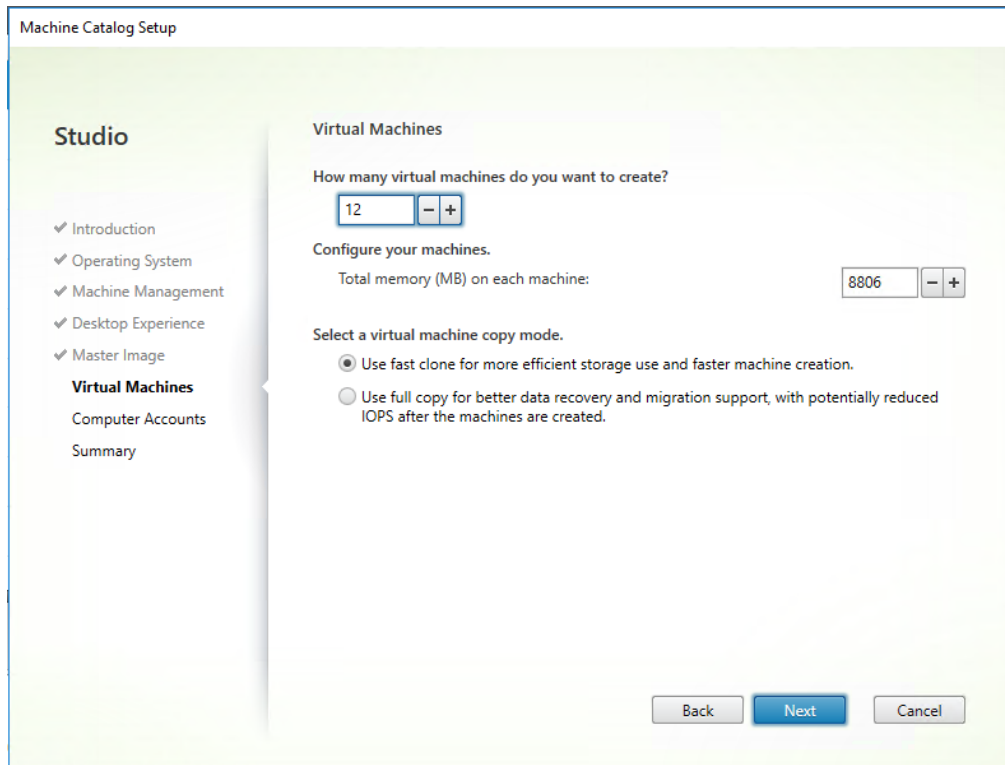
- [Not recommended: Citrix Personal vDisk technology is now deprecated.] Yes, save changes on a separate Personal vDisk.
- Yes, create a dedicated virtual machine and save changes on the local disk.
- No, discard all changes and clear virtual desktops when the user logs off.

At the bottom right are three buttons: 'Back', 'Next' (highlighted in blue), and 'Cancel'.

6. Select a virtual machine to be used for the catalog master Image. Then click Next (Figure 66).

Figure 66. Selecting a virtual machine for the master image

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. For the machine copy mode, select Fast Clone.
10. Click Next (Figure 67).

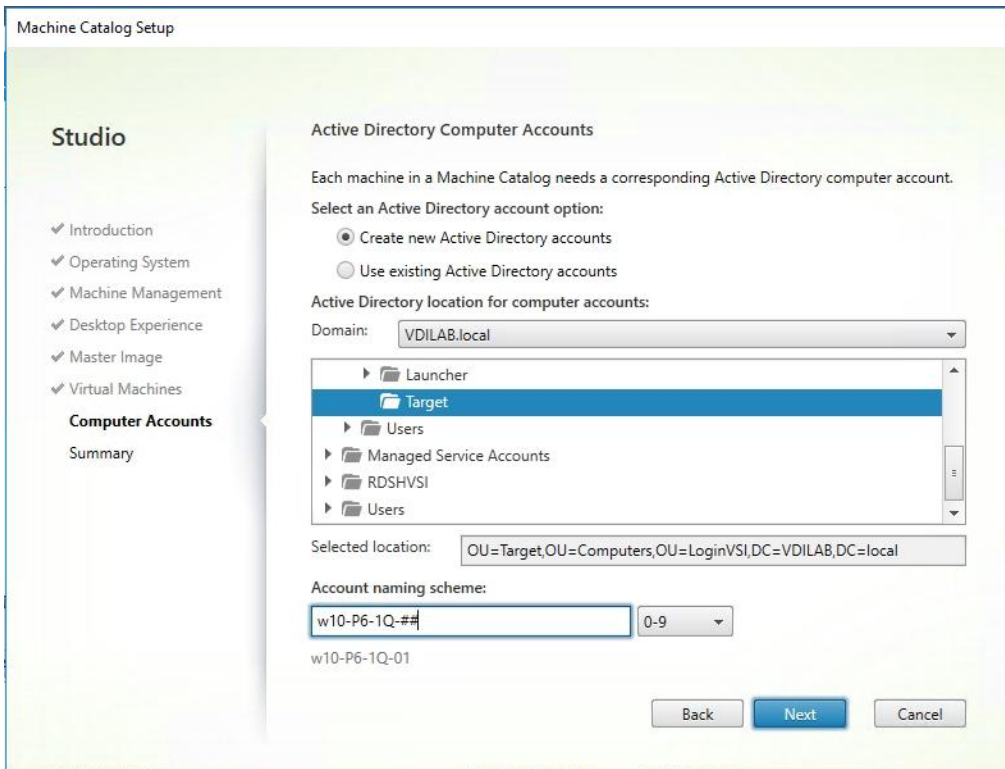
Figure 67. Configuring the virtual machines

The screenshot shows the 'Machine Catalog Setup' wizard in the 'Virtual Machines' configuration step. On the left, a 'Studio' sidebar lists navigation options: Introduction, Operating System, Machine Management, Desktop Experience, Master Image, **Virtual Machines**, Computer Accounts, and Summary. The main area is titled 'Virtual Machines' and contains the following configuration options:

- How many virtual machines do you want to create?** A numeric input field with the value '12' and minus/plus buttons.
- Configure your machines.** A sub-section with the label 'Total memory (MB) on each machine:' and a numeric input field with the value '8806' and minus/plus buttons.
- Select a virtual machine copy mode.** Two radio button options:
 - 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.

At the bottom of the wizard, there are three buttons: 'Back', 'Next' (highlighted in blue), and 'Cancel'.

11. Specify the Active Directory account naming scheme and organization unit (OU) where accounts will be created. Then click Next (Figure 68).

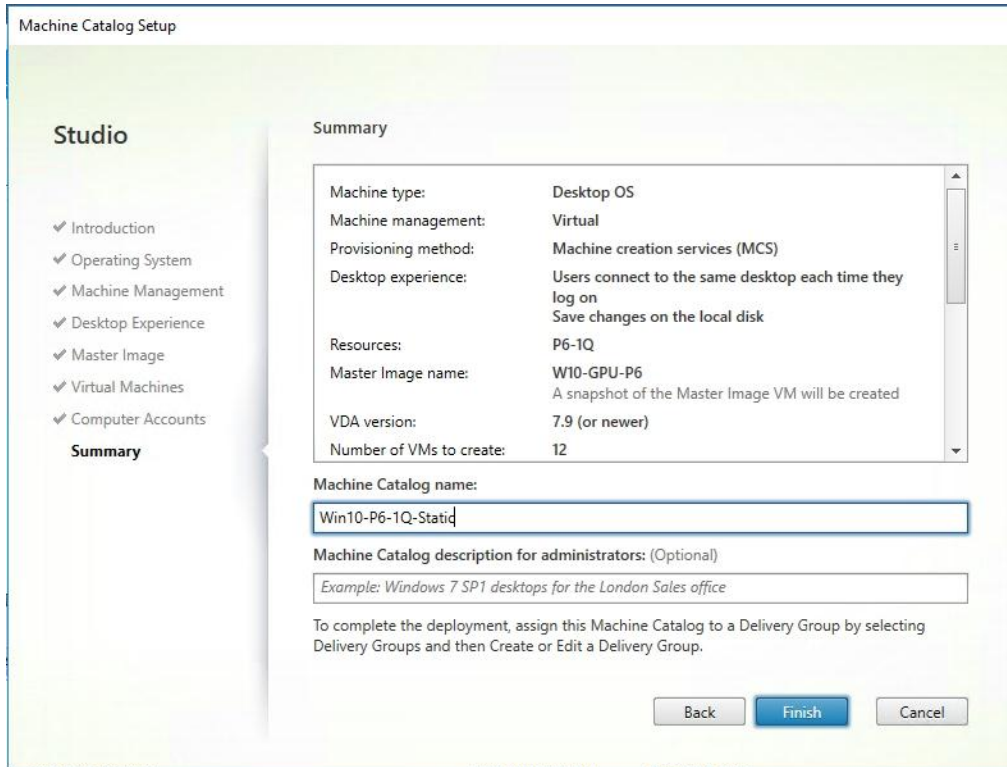
Figure 68. Creating Active Directory accounts

The screenshot shows the 'Machine Catalog Setup' wizard, specifically the 'Active Directory Computer Accounts' step. On the left, a 'Studio' sidebar lists navigation options: Introduction, Operating System, Machine Management, Desktop Experience, Master Image, Virtual Machines, Computer Accounts (selected), and Summary. The main content area is titled 'Active Directory Computer Accounts' and contains the following elements:

- A heading: 'Active Directory Computer Accounts'
- Text: 'Each machine in a Machine Catalog needs a corresponding Active Directory computer account.'
- Text: 'Select an Active Directory account option:'
- Two radio buttons: 'Create new Active Directory accounts' (selected) and 'Use existing Active Directory accounts'.
- Text: 'Active Directory location for computer accounts:'
- A 'Domain:' dropdown menu showing 'VDILAB.local'.
- A file explorer window showing a tree structure with folders: 'Launcher', 'Target' (selected), 'Users', 'Managed Service Accounts', 'RDSHVS1', and 'Users'.
- Text: 'Selected location: OU=Target,OU=Computers,OU=LoginVSI,DC=VDILAB,DC=local'
- Text: 'Account naming scheme:'
- A text input field containing 'w10-P6-1Q-##' and a dropdown menu set to '0-9'.
- Text: 'w10-P6-1Q-01' (displayed below the input field).
- Three buttons at the bottom: 'Back', 'Next' (highlighted in blue), and 'Cancel'.

12. On the Summary page, specify a catalog name and click Finish to start the deployment (Figure 69).

Figure 69. Machine Catalog Setup Summary page



Machine Catalog Setup

Studio

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

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 Save changes on the local disk
Resources:	P6-1Q
Master Image name:	W10-GPU-P6 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:

Machine Catalog description for administrators: (Optional)

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

Back Finish Cancel

Verify vGPU deployment

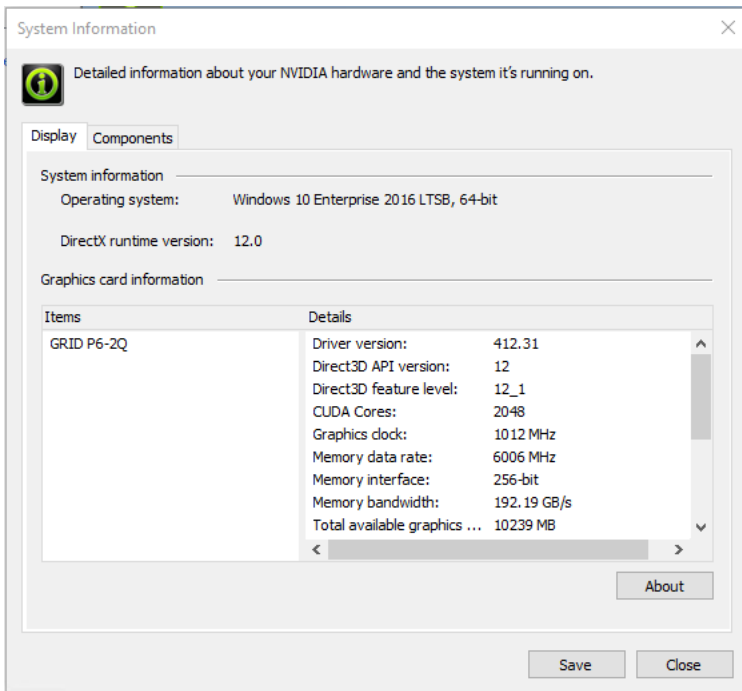
After the desktops are provisioned, use the following steps to verify vGPU deployment in the Citrix XenDesktop environment.

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. From 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 70).

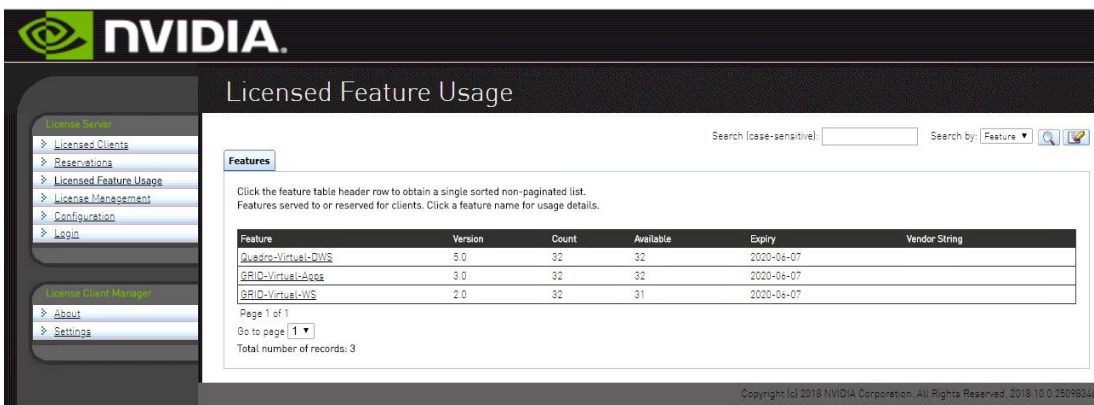
Figure 70. NVIDIA Control Panel: System Information



Verify NVIDIA license acquisition by desktops

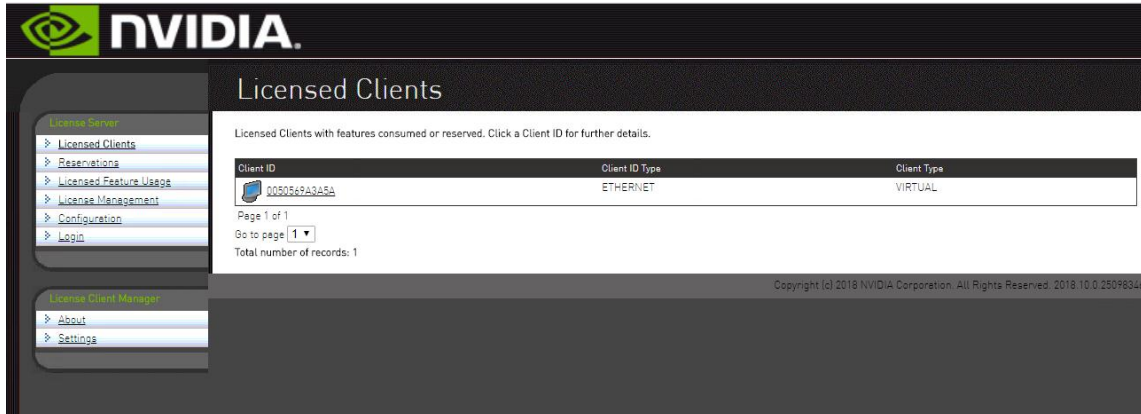
A license is obtained before the user logs on to the virtual machine after the virtual machine is fully booted (Figure 71).

Figure 71. NVIDIA license server: Licensed Feature Usage



To view the details, select Licensed Clients in the left pane (Figure 72).


Figure 72. NVIDIA license server: Licensed Clients



nvidia

Licensed Clients

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

Client ID	Client ID Type	Client Type
 0050569A3A5A	ETHERNET	VIRTUAL

Page 1 of 1
Go to page:
Total number of records: 1

Copyright (c) 2018 NVIDIA Corporation. All Rights Reserved. 2018.10.0.25098348

Create Citrix XenDesktop policies

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

XenDesktop policies control user access and session environments and provide the most efficient method 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 are typically defined through Citrix Studio. (The Windows Group Policy Management Console can also be used if the network environment includes Microsoft Active Directory and permissions are set for managing Group Policy Objects). Figure 73 shows the policies for GPU testing discussed in this document.

Figure 73. GPU testing policy: Very high-definition user experience

GPU Testing

Overview	Settings	Assigned to
▶	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 Direct X APIs. The benchmark's workloads, called viewsets, represent graphics content and behavior from actual applications.

SPECviewperf 13 uses these 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>.

SPECviewperf 13 results

Figures 74, 75, 76, and 77 show SPECviewperf results for various profiles.

Figure 74. SPECviewperf results: Single virtual machine testing on the host

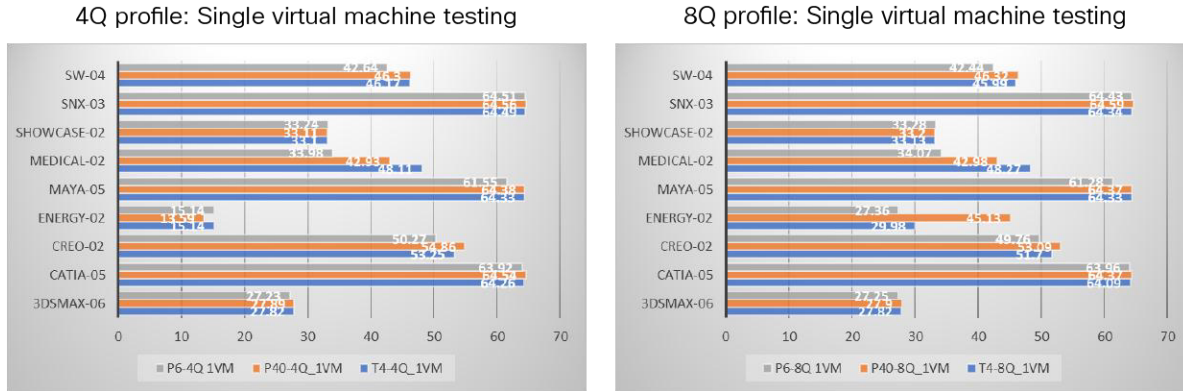


Figure 75. SPECviewperf results for T-4 4Q and 8Q profile testing: Single virtual machine versus maximum density on the host with a single card

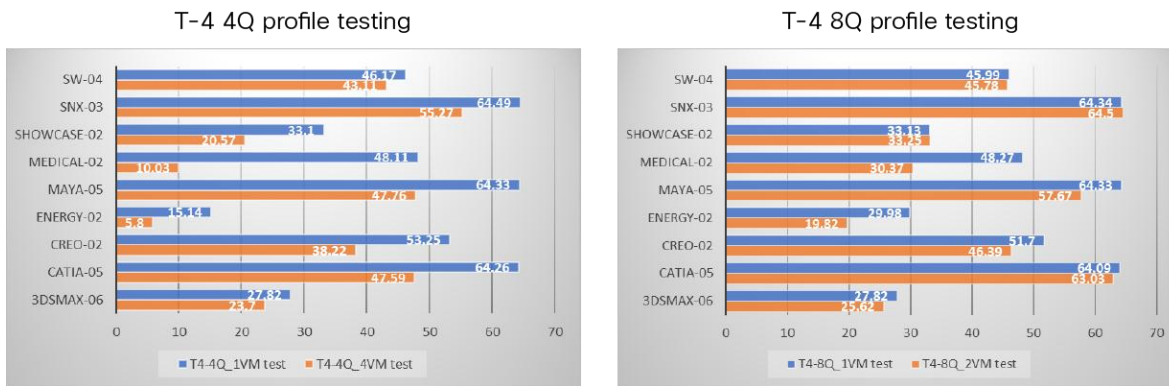


Figure 76. SPECviewperf results for P-40 4Q and 8Q profile testing: Single virtual machine versus maximum density on the host with a single card

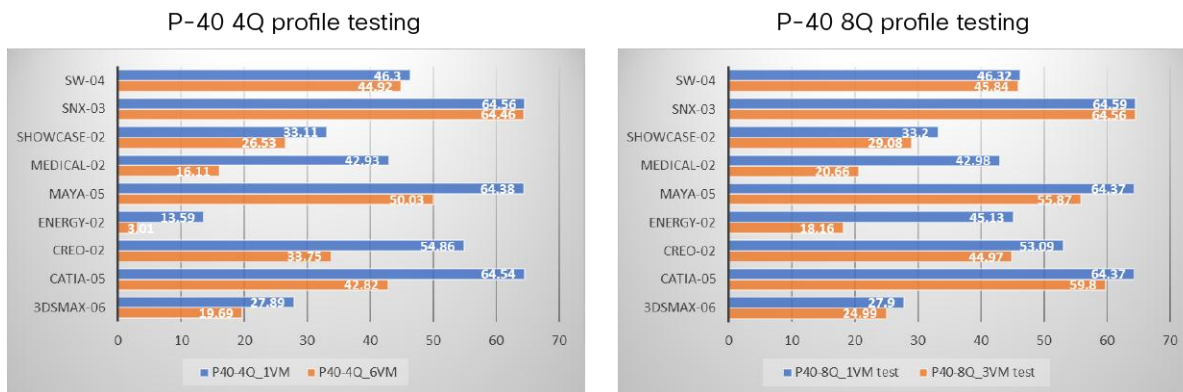
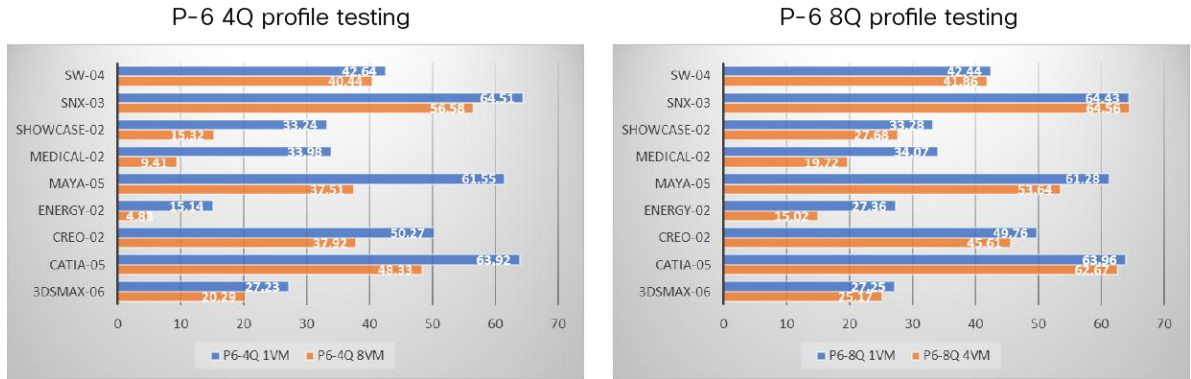


Figure 77. SPECviewperf results for P-6 4Q and 8Q profile testing: Single virtual machine versus maximum density on the host



Host CPU utilization

Figures 78, 79, and 80 show host CPU utilization results.

Figure 78. Cisco UCS C240 CPU utilization with SPECviewperf T-4 4Q and 8Q profile testing: Single virtual machine versus card maximum density.

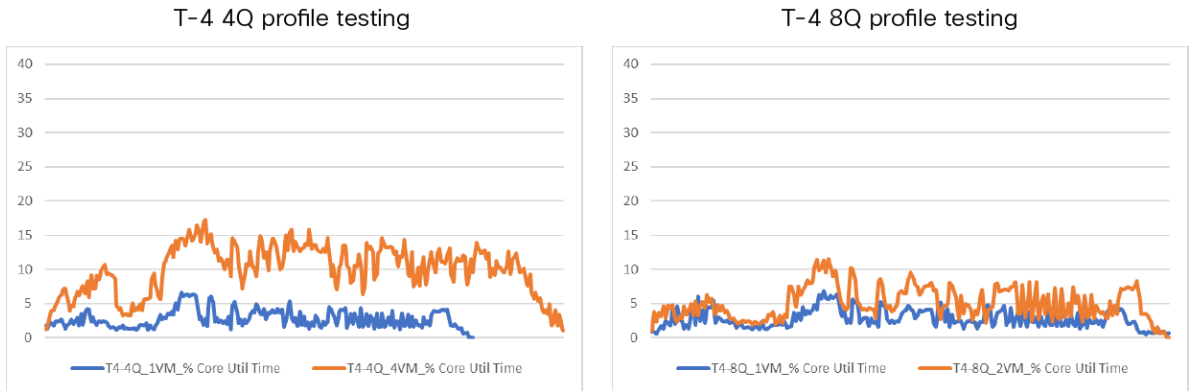


Figure 79. Cisco UCS C240 CPU utilization with SPECviewperf P-40 4Q and 8Q profile testing: Single virtual machine versus card maximum density.

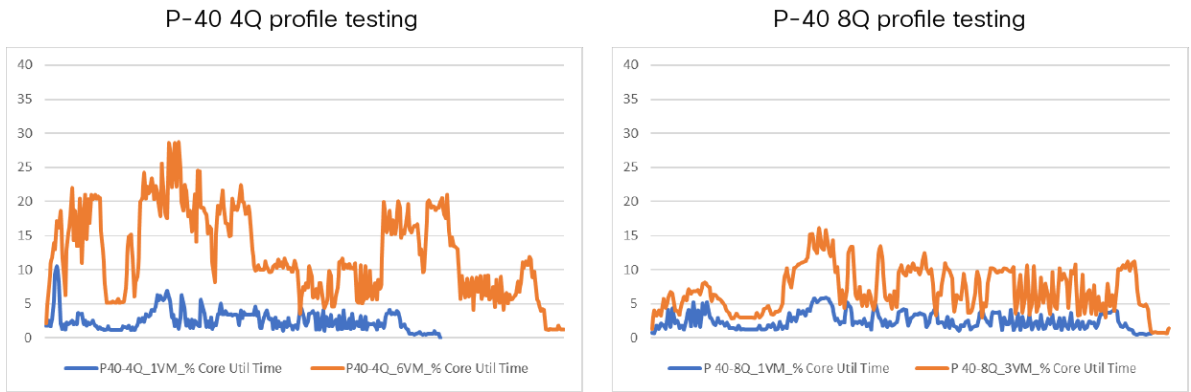
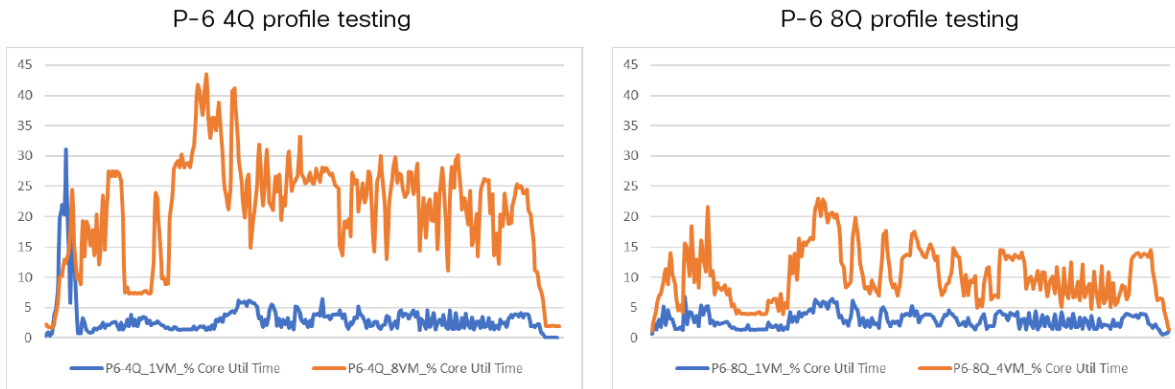


Figure 80. Cisco UCS B200 CPU utilization with SPECviewperf P-6 4Q and 8Q profile testing: Single virtual machine versus maximum density



Host GPU utilization

Figures 82, 83, and 84 show host GPU utilization results.

Figure 81. NVIDIA T-4 with SPECviewperf T-4 4Q and 8Q profile testing: Single virtual machine versus card maximum density

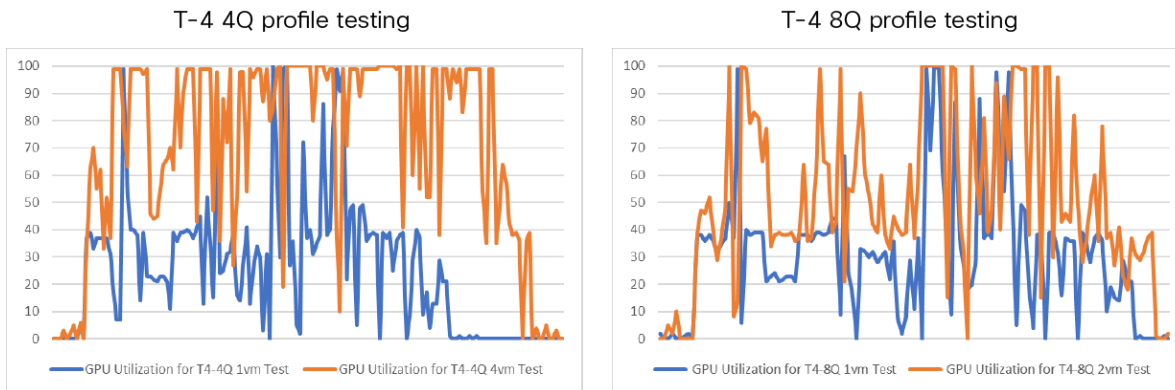


Figure 82. NVIDIA P-40 with SPECviewperf P-40 4Q and 8Q profile testing: Single virtual machine versus card maximum density

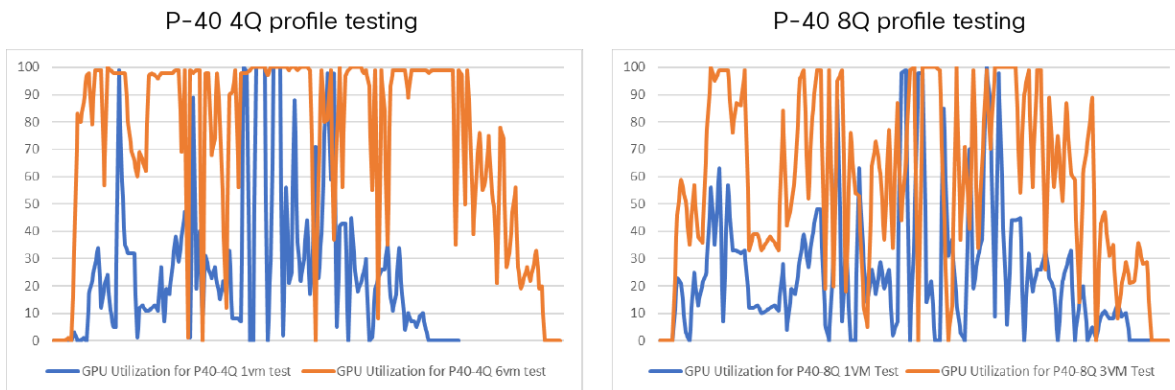
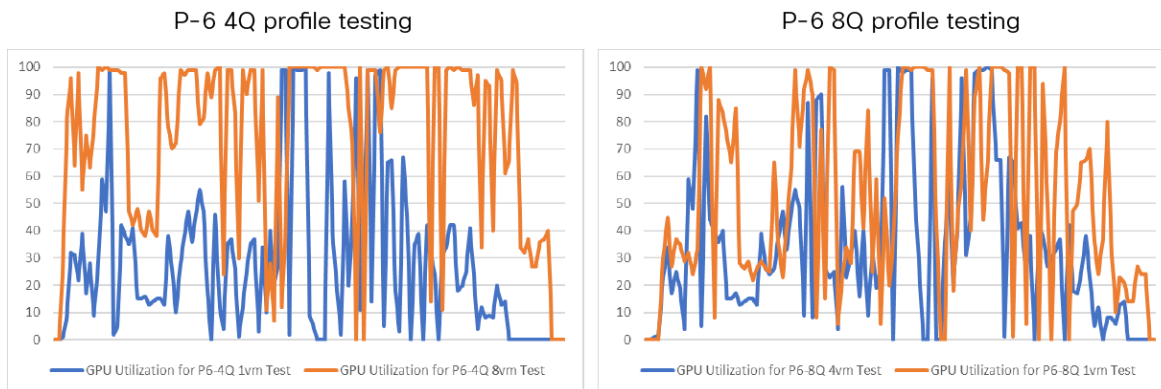


Figure 83. NVIDIA P-6 with SPECviewperf P-6 4Q and 8Q profile testing: Single virtual machine versus maximum density

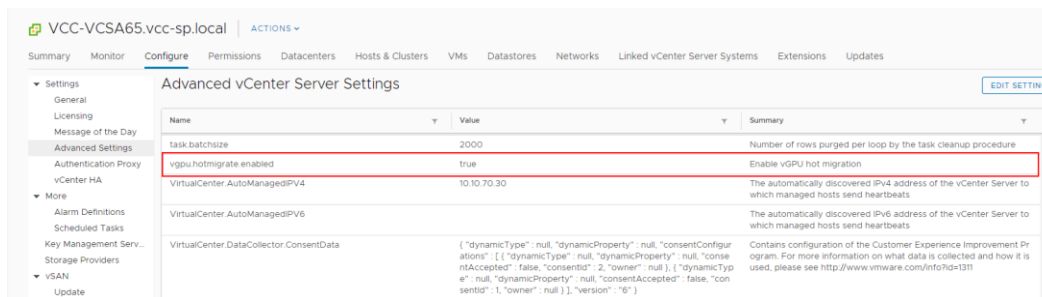


Live vGPU-enabled virtual machine with VMware vMotion

You can use the VMware vMotion Migration wizard to migrate a powered-on virtual machine from one computing resource to another by using vMotion. For more information about support and restrictions, refer to the VMware documentation.

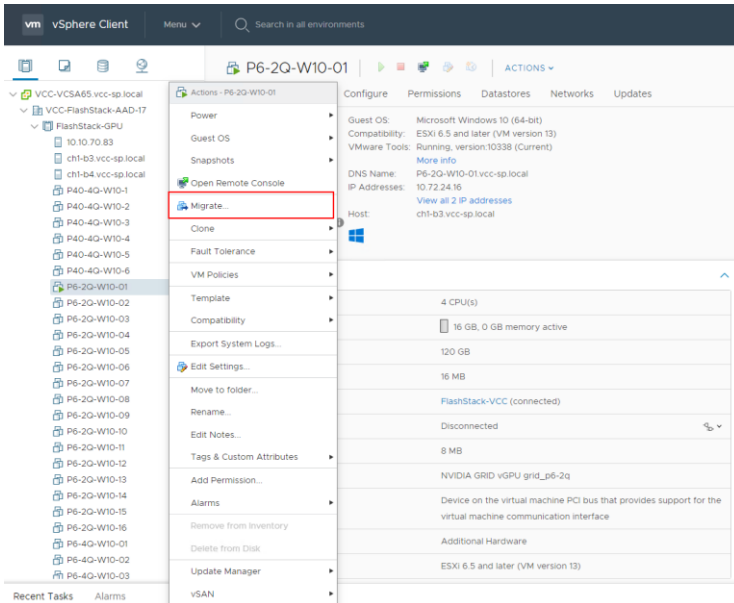
1. Verify that the Advanced vCenter Server setting `vgpu.hotmigrate.enabled` is set to true (Figure 84).

Figure 84. Advanced VMware vCenter Server Settings



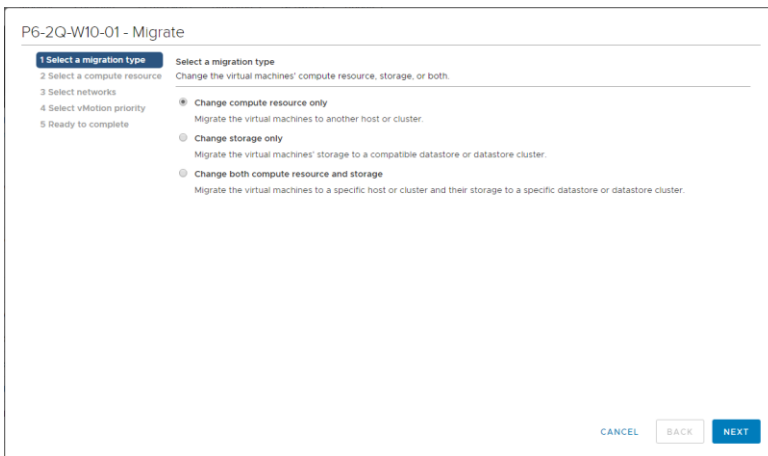
2. Right-click the virtual machine and choose Migrate (Figure 85).

Figure 85. Choosing Migrate



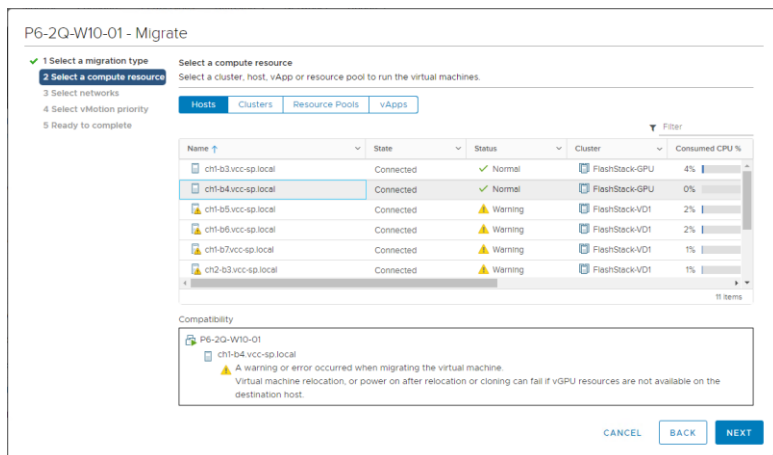
3. Select “Change compute resource only” and click Next (Figure 86).

Figure 86. Selecting a migration type



4. Select a new host to run the virtual machine and click Next (Figure 87).

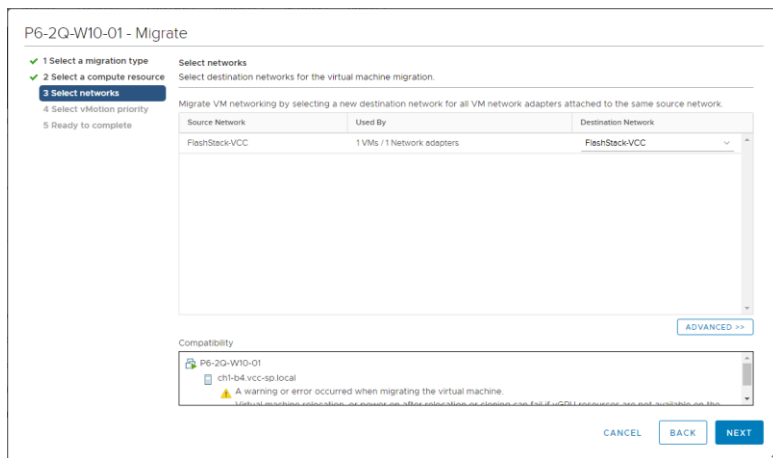
Figure 87. Selecting the host to run the virtual machine



Note: If a compatibility problem arises, it is listed in the Compatibility panel. Fix the problem or select another host or cluster.

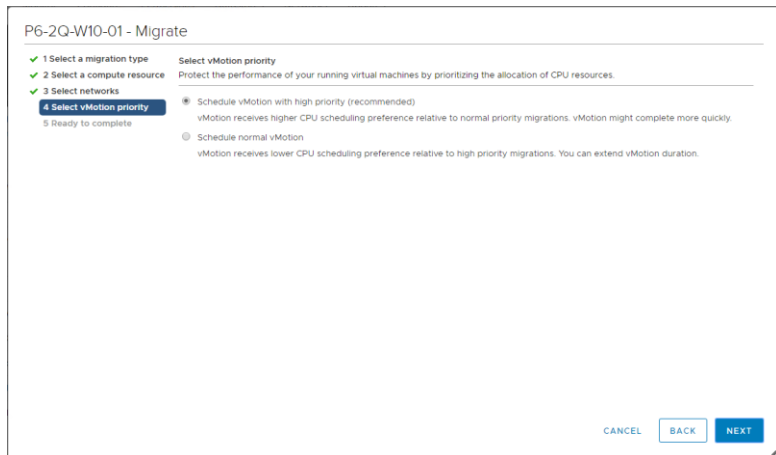
5. Select a destination network for all connected virtual machine network adapters and click Next (Figure 89).

Figure 88. Selecting a destination network



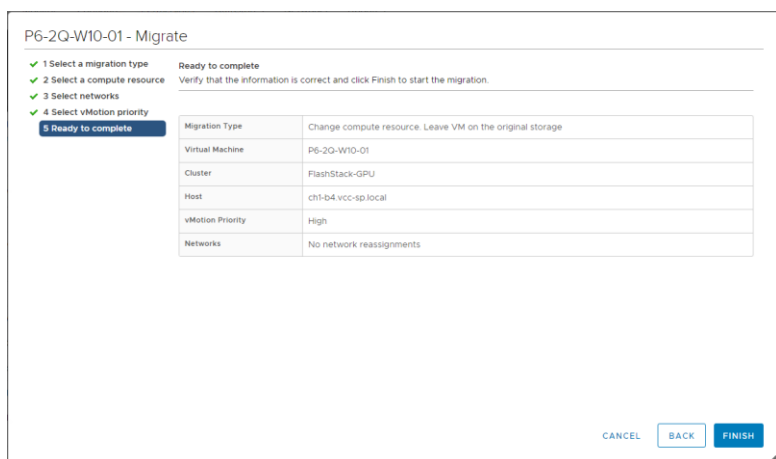
6. Schedule vMotion with high priority and click Next (Figure 90).

Figure 89. Scheduling VMware vMotion



7. Review the settings page and click Finish (Figure 91).

Figure 90. Clicking Finish



8. Verify the migration results (Figure 92 and Figure 93).

Figure 91. VMware vCenter Recent Tasks: vGPU-enabled vMotion progress

Task Name	Target	Initiator	Status	Queued For	Start Time	Completion Time	Server
Relocate virtual machine	P6-2Q-W10-01	VSPHERE.LOCAL\Administrator	45%	4 ms	06/21/2019, 4:05:17 PM		VCC-VCSA65.vcc-e...

Figure 92. VMware vCenter Recent Tasks: vGPU-enabled vMotion completion

Task Name	Target	Initiator	Status	Queued For	Start Time	Completion Time	Server
Relocate virtual machine	P6-2Q-W10-01	VSPHERE.LOCAL\Administrator	Completed	3 ms	06/21/2019, 4:19:55 PM	06/21/2019, 4:20:20 PM	VCC-VCSA65.vcc-e...

Additional configurations

This section presents additional configuration options.

Install and upgrade NVIDIA drivers

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

Use Citrix HDX Monitor

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

Optimize the Citrix HDX 3D Pro user experience

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

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

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

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

For example, if 20 users share a 60-Mbps connection, the bandwidth available to each user can vary between 3 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 Streaming Single Instruction, Multiple Data [SIMD] Extensions (SSE) 4.1 and Advanced Vector Extensions (AVX). The OpenGL accelerator also supports applications using OpenGL versions up to Version 2.1.
- For applications running on a workstation, first try the default version of OpenGL support provided by the workstation's graphics adapter. If the graphics card is the latest version, in most cases it will deliver the best performance. If the graphics card is an earlier version or does not deliver satisfactory performance, then try the OpenGL Software Accelerator.
- 3D OpenGL applications that are not adequately delivered using CPU-based software rasterization may benefit from OpenGL GPU hardware acceleration. This feature can be used on bare-metal devices and virtual machines.

Conclusion

The combination of Cisco UCS Manager, Cisco UCS C240 M5 Rack Servers and B200 M5 Blade Servers, Pure Storage FlashArray //x70 R2, and NVIDIA cards running on VMware vSphere 6.7 Update 1 and Citrix XenDesktop 7.15 LTSR provides a high-performance platform for virtualizing graphics-intensive applications.

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

For more information

For additional information about topics discussed in this document, see the following resources:

- Cisco UCS C-Series Rack Servers and B-Series Blade Servers:
 - <http://www.cisco.com/en/US/products/ps10265/>
- NVIDIA:
 - <http://www.nvidia.com/object/grid-technology.html>
- Citrix XenApp and XenDesktop:
 - <https://docs.citrix.com/en-us/xenapp-and-xendesktop/7-15-ltsr.html>
 - <http://blogs.citrix.com/2014/08/13/citrix-hdx-the-big-list-of-graphical-benchmarks-tools-and-demos/>
- 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>

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.

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